# Cognitively-motivated geometric methods of pattern discovery and models of similarity in music

James C. Forth

Submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy
of the University of London.

Department of Computing
Goldsmiths, University of London

March 2012

I certify that this dissertation, and the research to which it refers, are the result of my own work.

Dedicated to Douglas Tingay (1927—2012)

# Abstract

This thesis is concerned with cognitively-motivated representations of musical structure. Three problems are addressed, each related in terms of their focus on music as an object of perception, and in the application of geometrical methods of knowledge representation.

The problem of pattern discovery in discrete representations of polyphonic music is first considered, and a heuristic proposed which seeks to assist musicological analysis by identifying patterns that may be salient in perception, from a large number of potential patterns. This work is based on geometric principles that are far removed from plausible psychological models of pattern induction, but the method is motivated by psychological evidence for the importance of invariance and repetition in perception.

The second and third problems explicitly adopt a cognitive theory of representation, namely the conceptual space framework developed by Gärdenfors (2000). Within this framework, concepts can be represented geometrically within perceptually grounded quality dimensions, and where distance in the space corresponds to similarity. The second problem concerns the prediction of melodic similarity, and the theory of conceptual spaces is investigated in the novel context of point set representations of melodic structure, employing the Earth Mover's Distance metric (Rubner et al. 2000). This work builds on the work of Typke (2007) concerning the application of Earth Mover's Distance to melodic similarity. Evaluation is performed with respect to published psychological data (Müllensiefen and Frieler 2004), and the MIREX 2005 symbolic melodic similarity evaluation.

The third problem concerns the conceptual representation of metrical structure, informed by the psychological theory of metre developed by London (2004). A symbolic formalisation of this theory is developed, alongside two geometrical models of metrical-rhythmic structure, which are evaluated within a genre classification task.

# Acknowledgements

First and foremost, I would like to thank my supervisor Geraint Wiggins for his support, generosity and unwavering optimism. I am also indebted to the Intelligent Sound and Music Systems group, with which I am honoured to have been associated. Thanks in particular to: Marcus Pearce for his supervisory input; my collaborator Alex McLean; Daniel Müllensiefen for answering my many psychology questions; and David Lewis and Christophe Rhodes for their work on the AMuSE and Gsharp software that was used extensively throughout this research. Thanks to Tim Crawford, Richard Lewis, Ben Fields, Ray Whorley, Polina Proutskova, Alastair Craft, Bruno Gingras, and Dan Jones for the stimulating conversations and for sharing with me their inspirational work. From the wider Department of Computing, I would like to thank in particular Janis Jefferies, Sarah Rauchas, and Mark D'Inverno for their support and encouragement. This dissertation also greatly benefited from the careful reading of my examiners, Simon Dixon and Justin London.

Data used in some of the experiments reported in this dissertation (as well as others that are not), was kindly provided by Rainer Typke, Klaus Keil, Stephan Hirsch, Stephen Downie, Anja Volk, Anna Wolf and Daniel Cameron.

David Burnand, Michael Oliva and Ingrid Pearson at the Royal College of Music were instrumental at the beginning stages of the research reported in this dissertation. I thank them for their feedback on my earlier work and composition, in helping me to secure an AHRC doctoral award (number 118566), and for enabling a smooth transition from South Kensington to New Cross.

I would like to thank my family and friends for their love and support. Last but not least, thanks to Rose.

# Related Publications

Some of the work contained within this dissertation has appeared in the following publications.

An earlier version of the research reported in chapter 3 was also reported in:

J. Forth and G. A. Wiggins (2009). "An approache for identifying salient repetition in multidimensional representations of polyphonic music". In: *London Algorithmics 2008: Theory and Practice*. Ed. by J. Chan et al. Texts in Algorithmics. London, UK: College Publications, pp. 44–58

Early work into conceptual space representations of musical structure was published in:

J. Forth et al. (2008). "Musical creativity on the conceptual level". In: *Proceedings of the 5th International Joint Workshop on Computational Creativity*. Ed. by P. Gervás et al., pp. 21–30

An earlier version of the conceptual space of metre presented in chapter 5 was published in:

J. Forth et al. (2010). "Unifying conceptual spaces: Concept formation in musical creative systems". In: *Minds and Machines*, pp. 11–30. DOI: 10.1007/s11023-010-9207-x

Related work concerning the design and implementation of the AMusE system, which was used as the implementation framework for the experiments reported in this dissertation, was published in:

D. Lewis et al. (2011). "Tools for music scholarship and their interactions: a case study". In: *Proceedings of the Supporting Digital Humanities Conference (SDH 2011)*. Ed. by B. Maegaard. URL: http://sldr.org/SLDRdata/doc/show/copenhagen/SDH-2011/proceedings.html

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

The research reported in this dissertation demonstrates a general approach for the representation and modelling of music informed by psychological and cognitive principles. Music is fundamentally a psychological phenomenon, which is the perspective taken in the development of computational methods for investigating aspects of music and musical behaviour.

Three problems are addressed, each related in terms of their focus on music as an object of perception, and in the application of geometrical methods of knowledge representation. The problem of pattern discovery in discrete representations of polyphonic music is first considered, building on the SIATEC structure induction algorithm. SIATEC is an algorithm for discovering patterns in multidimensional point sets (Meredith et al. 2002). This algorithm has been shown to be particularly useful for analysing musical works. However, in raw form, the results generated by SIATEC are large and difficult to interpret. We propose an approach, based on the generation of set-covers, which aims to identify particularly salient patterns that may be of musicological interest. Our method is capable of identifying principal musical themes in Bach Two-Part *Inventions*, and is able to offer a human analyst interesting insight into the structure of a musical work. This work is based on geometric principles that are far removed from plausible psychological models of pattern induction, but the method is motivated by psychological evidence for the importance of invariance and repetition in perception.

The second and third problems that are addressed explicitly adopt a cognitive theory of representation, namely the conceptual space framework developed by Gärdenfors (2000). Within this framework, concepts can be represented geometrically within perceptually-grounded quality dimensions, and where distance in the space corresponds to similarity. We follow an existing abstract vector space formalisation of the conceptual space theory, but the application in the domain of

music represents a novel contribution.

The second problem concerns the prediction of melodic similarity, and the theory of conceptual space is investigated in the novel context of point set representations of melodic structure. Melodies are represented as point sets in low dimensional spaces. Quality dimensions representing the basic event attributes of onset, pitch and duration are defined, as well as dimensions representing relations between basic attributes. The usefulness of the individual dimensions in affording prediction of melodic similarity is evaluated. The Earth Mover's Distance metric (Rubner et al. 2000) is employed as the measure of distance between point set representations of melodies. Different weighting schemes are defined, giving different EMD measures based on partial or complete matching between point sets. This represents an additional factor in our evaluation. Evaluation is performed with respect to published psychological data (Müllensiefen and Frieler 2004), and the MIREX 2005 symbolic melodic similarity evaluation.

The third problem concerns the conceptual representation of metrical structure, informed by the psychological theory of metre developed by London (2004). London's theory is based on a psychological, cognitive, neuroscientific, and musicological understanding of metre as a process of entrainment. We develop a symbolic formalisation of this theory as metrical tree structures, which forms the basis of two conceptual space representations. The first represents metrical concepts as hierarchical structures of periodic components. The second extends this representation to include the internal sequential structure of periodic cycles. The geometry is defined in terms of the symbolic formulation, and the mappings between the levels of representation associates symbolic metrical tree structures with points in geometrical space. Expressively varied metres are naturally represented in the space as regions surrounding prototypical metrical points. The developed models of metrical-rhythmic structure are evaluated within a standard genre classification task involving stratified 10x10-fold cross-validation over a labelled dataset using $k$-nearest-neighbour clustering.

Following this introduction, chapter 2 considers general issues within music representation, before introducing Gärdenfors' theory of conceptual space. Chapter 3 addresses the problem of pattern discovery. Chapter 4 concerns the problem of melodic similarity. Chapter 5, the most substantial portion of the thesis, considers the conceptual representation of metre. Conclusions arising from the individual chapters are brought together and considered in chapter 6.

# Chapter 2

# Music representation

A representation is a formal language used to express information. Honing (1993) identifies four categories of approaches to the subject of representation within music research. The first pair of categories concerns the motivations of those interested in music representation. The first are those for whom the representation of musical information supports their music research, for example, musicologists or composers. The second are those for whom the subject of representational languages itself is the area of research. The overall aim of this thesis is concerned with the issue of musical representation itself.

The second pair of categories identified by Honing concern the types of the representations proposed for music. Representations are either predominantly technical, that is, are designed in response to a particular technical problem, or else aim to represent conceptual or mental musical structures. The former category emphasises observable and measurable musical attributes, such as the key press of a piano keyboard, or the position of a note head on a musical score. The latter seeks to capture aspects of the listening experience of music, predominantly as part of a computational theory aiming to predict aspects of musical behaviour. The representations pursued here naturally fall into this latter category, and are motivated by the aim of capturing some of the perceptual qualities salient to musical perception and conceptualisation.

## 2.1 Musical information

In an early paper discussing the application of computer technology to music research, Babbitt (1965, p. 76) employs the terms *graphemic*, *acoustic*, and *auditory* to distinguish three related domains of musical information. The acoustic domain encompasses the physical manifestations of music, such as the propagation of sound

waves, and representations of such properties, such as the analogue representation of music stored on electromagnetic tape, or the stream of bits resulting from analogue-to-digital conversion. The representation of acoustic information most naturally falls within Honing's category of technical representations (Honing 1993, p. 222), since there exists concrete referents in the world to which the representational language refers.

The graphemic domain pertains to the graphical notation of music, such as conventional musical scores and tablature. Graphical notations are themselves representations of music, serving primarily as musical aide-mémoires and for the communication of musical ideas. From the computational perspective, there is scope here for both technical and cognitive representational approaches. Where the aim is simply to represent the exact layout of notation symbols on a score, a purely technical representation is adequate. However, if the aim is to also represent associated music-theoretical meaning, or possible performance interpretations, then the representation language must necessarily express, at least in part, the musical knowledge assumed by each notation system. Such information could also be described as declarative knowledge or procedural knowledge. For example, the representation could describe a trill declaratively as an object of ornamentation, or alternatively, as a form of procedural knowledge describing how the trill is created (Honing 1993, p. 229).

The auditory domain covers information about music as perceived by the listener, aligning with Honing's category of conceptual and mental representations. The characterisation of musical information into the domains of the acoustic, graphemic, and auditory is not exhaustive; for example, gestures made by performers would be another potentially relevant domain of information (Selfridge-Field 1997, p. 7). However, the distinctions are nonetheless important categories of musical information, and help to clarify the issues of what is entailed in representing "music". The phenomenon of music itself cannot be said to exist in any one domain exclusively, but instead can be understood as something that exists between the domains, with each one offering a particular perspective from which to study music (Wiggins 2008).

Most representations of music are designed for a particular task, which often remains within one of the domains characterised by Babbitt. The well known problem in music information retrieval of music synchronisation aims to bridge the gap between at least two domains, the graphemic and acoustic. Systems such as SyncPlayer (Fremerey et al. 2007) aim to allow users to navigate through a library of audio recordings using automatically aligned digital images of musical scores. Bab-

bitt notes the difficulty in automatically translating knowledge between domains, and the problem is still very much present today.

## 2.2 Generality in representation

To consider the issue of music representation at the most general level is to question the nature of music itself, and is arguably 'chimerical and counterproductive' (Huron 1992, p. 34). Huron advocates a practical goal-oriented approach, in which the question of what aspects of music to represent is largely determined by the intended application. Huron (1992, p. 20) states that 'the problem of music representation hinges on elucidating finite sets of properties which are suited to the achievement of certain goals.' This statement is equally applicable to the problem of knowledge representation in other fields of discourse. However, the multifaceted nature of musical information may require, for certain tasks, quite different forms of knowledge to be expressed and interrelated. For example, the requirements of a representation suitable for manipulating audio recordings are very different to those necessary to represent medieval music notation. However, one could argue that for an application whose intended purpose is to allow the user to relate information across different domains of musical information, such as between graphemic and acoustic information, then a certain degree of generality is beneficial, if not necessary, to facilitate such mappings.

Further problems can also be identified in taking a strictly goal-oriented approach to music representation. Defining a representation for one specific task potentially limits the sharing of data and tools, as inevitably information considered irrelevant for one task will be necessary for another. This is perhaps not a serious issue for individual researchers, but is potentially inhibiting of progress across the research field as a whole. Furthermore, it is impossible to anticipate all possible situations in which users might want to use a representation, and indeed Huron states extensibility as one of the qualities of a good representation. However, there are arguably better and worse ways to facilitate user extensibility. In the worse case, users may be able to implement ad hoc extensions to meet their immediate needs, at the expense of portability. A better approach to allow customisation is to employ a suitably general framework within which extensions can be defined within the semantics of the representation itself.

One possible way to resolve these tensions is to clarify where exactly generality should be sought. Huron (1992) seems to be referring to all possible manifestations of music when warning against generality. When one considers the potential scope

of music information—from the range of world musics as well as the numerous ways of thinking about and analysing musical phenomena—this position seems ultimately to be the only tenable approach. However, even within a goal-oriented approach, generality can still be sought within the representation itself, which overcomes many of the potential problems associated with a plethora of bespoke representational schemes.

In order to assist answering such questions, Wiggins et al. (1993) propose a framework for the evaluation of music representation systems based on the two dimensions of *expressive completeness* and *structural generality*. Expressive completeness refers to the 'range of raw musical data that can be represented', and structural generality refers to the 'range of high-level structures that can be represented and manipulated' (Wiggins et al. 1993, p. 31). According to these criteria, audio recordings of musical performances are high in expressive completeness because they capture in great detail the raw physical manifestation of music. However, audio recordings are also very low in structural generality because there is no explicit representation of the structural components of the music, such as notes, instrumentation, or metrical grouping. Representations designed for graphemic or auditory musical information necessarily require a higher level of structural generality since the information directly concerns identifiable musical objects and concepts. For example, a representation designed to express the perceptually salient features of a melody will require at least the means of representing discrete note events, together with their pitch and duration. However, such representations are low in expressive completeness since they lack the means of recreating the continuous waveform corresponding to a musical performance.

### 2.2.1 Charm

A simple, yet powerful approach to a general representation of music is proposed by Wiggins et al. (1989), Harris et al. (1991), and Smaill, Wiggins, and Harris (1993). The Common Hierarchical Abstract Representation for Music (Charm) aims to support a high degree of generality within the representation itself, which is a distinct notion to the quest of representing music in the most general sense, which would ultimately require a representation capable of encompassing all possible mental, physical, and cultural manifestations of music and musical behaviour. The need for stated objectives for any representational formalism is maintained, and indeed is considered necessary for evaluating differences between representations (Wiggins et al. 1993). Charm is defined initially as a representation of music at the symbolic level, in which identifiable aspects of music are represented by discrete symbols.

As such, Charm is appropriate for representing a wide range of graphemic information, but may appear less appropriate for acoustic or other continuously-valued musical information. However, as a general framework for musical representation, developing Charm compliant representation of acoustic and auditory information is perfectly feasible. Symbolic, or discrete, representations are particularly appropriate for the high-level description of a range of perceptual attributes and concepts, such as for representing discrete musical events, groupings of events, and for expressing the formal properties of relationships between such structures. The representations developed in chapters 4 and 5 seek to extend the purely symbolic approach of Charm with perceptually-grounded geometrical forms of representation, in order to capture more fluid and continuous music-psychological phenomena.

Charm is based on the computer science concept of abstract data typing. The authors note that despite the direct incompatibility of many music representation schemes, that a considerable degree of commonality exists at an abstract level. For example, most schemes define some way of representing pitch, whether in terms of MIDI note numbers, scale degree, microtonal divisions of the octave, or frequency. However, at an abstract level, common patterns of operations can be observed, which are irrespective of the underlying implementation. Therefore, the authors propose an *abstract* representation, in which musically meaningful operations can be defined in terms of abstract data types. Harris et al. (1991) define basic data types for pitch (and pitch interval), time (and duration), amplitude (and relative amplitude) and timbre. Therefore, the abstract event representation is the Cartesian product:

$$\mathit{Pitch} \times \mathit{Time} \times \mathit{Duration} \times \mathit{Amplitude} \times \mathit{Timbre}$$

In the case of time, the following functions can be defined where the arguments in $\{t, d\}$ denote *Time* or *Duration* data types respectively.

$$\begin{aligned} \mathit{add}_{dd} &: \quad \mathit{Duration} \times \mathit{Duration} \rightarrow \mathit{Duration} \\ \mathit{add}_{td} &: \quad \mathit{Time} \times \mathit{Duration} \rightarrow \mathit{Time} \\ \mathit{sub}_{tt} &: \quad \mathit{Time} \times \mathit{Time} \rightarrow \mathit{Duration} \\ \mathit{sub}_{dd} &: \quad \mathit{Duration} \times \mathit{Duration} \rightarrow \mathit{Duration} \end{aligned} \tag{2.1}$$

Typed equivalents of arithmetic relational operators (e.g., $\leq, \geq, =, \neq$) are also defined, permitting ordering and equality relations to be determined. With the exception of timbre, the internal structure of each basic data type is the same, allowing comparable functions to be defined modulo renaming (Harris et al. 1991).

Given a specification of abstract musical data types, a user can implement specific functionality required for their data and application according to the abstract definitions. Implementing specific functionality effectively means supplying a concrete implementation for each operation defined on each abstract type. Or better still, given a suitable high-level language capable of expressing the type information, one could declaratively specify concrete types, with the additional benefit of being able to infer mappings between types.

The abstract data type approach to representing music extends beyond the representation of surface level events. Charm formally defines the concept of the *constituent*, which allows arbitrary hierarchical structures to be specified (Harris et al. 1991). At the abstract level, a constituent is defined as the tuple:

$$\langle \textit{Properties/Definition}, \textit{Particles} \rangle$$

*Particles* is a set whose elements, called particles, are either events or other constituents. No constituent can be a particle of itself, defining a structure of constituents as a directed acyclic graph. *Properties/Definition* is the 'logical specification of the relationship between the particles of this constituent in terms of the membership of some class' (Harris et al. 1991, p. 8). The distinction between *Properties* and *Definition* is made explicit in a concrete implementation. However, at the abstract level, they both logically describe the structure of the constituent. Properties refer to 'propositions which are *derivably* true of a constituent' (Harris et al. 1991, p. 10); for example, that no particle starts between the beginning and end of any other particle, defined by Harris et al. (1991) as a *stream*:

$$\begin{aligned} \textit{stream} \quad \Leftrightarrow \quad & \forall p_1 \in \textit{particles}, \neg\exists p_2 \in \textit{particles}, \\ & p_i \neq p_2 \wedge \\ & \textit{GetTime}(p_1) \leq \textit{GetTime}(p_2) \wedge \\ & \textit{GetTime}(p_2) < \textit{add}_{td}(\textit{GetTime}(p_1), \textit{GetDuration}(p_1)) \end{aligned} \tag{2.2}$$

where *GetTime* and *GetDuration* are selector functions returning the timepoint and duration respectively of a given particle. Definitions are propositions that are true by definition; for example, that a set of particles contains all the events notated in a score of a particular piece of music.

An implementation of a Charm-compliant representation requires some additional properties, both for computational efficiency and user convenience. The following is an example of a simple 'motif' constituent (Smaill, Wiggins, and Harris 1993).

```
constituent(c0, stream(0, t1), motif, [e1, e2, e3, e4])
```

Every event and constituent defined within the system must be associated with a unique identifier, shown as `c0`, `e0`, `e1` and so forth in the above example. The constituent is a stream, with a start time and a duration, denoted by the property `stream(0, t1)`, which is derivably true from the events it contains. In contrast, the constituent is defined as a motif, and a user is free to provide such definitions for their own purposes.

Smaill, Wiggins, and Miranda (1993) present an application utilising the Charm representation to support human creativity. The system is designed to aid exploration of a timbral space, and utilises machine learning techniques to simulate the human process of concept formation. Although concerning timbre, the motivation behind the work is very close to that pursued in this thesis.

A wider benefit of adopting an abstract data type approach to music representation is that it provides the basis for developing a common platform for the sharing of data, as well as software tools. This is demonstrated in Smaill, Wiggins, and Harris (1993) in which both the implementation language and the concrete representations of the data are shown to be immaterial given that the correct behaviour of the abstract data types is observed. From a formal perspective, many issues of representation discussed in the field can be seen as concerning merely arbitrary matters of encoding or data serialisation. Although encoding schemes may well be designed to meet particular needs, such as to facilitate efficient human data entry or to be space efficient, the ontological commitments implicit in the encoding can be left unstated, and therefore potentially ambiguous, or even unquestioned, ultimately limiting potential usefulness. The Advanced MUSical Encoding (AMusE) system (Lewis et al. 2011), a Charm-compliant software framework for music computation, is used for the implementation of all experiments reported in this dissertation.

## 2.3 Theory of conceptual space

Honing (1993, p. 221) notes the necessity of incorporating cognitive aspects of music into its representation: '[s]ince a representation of the real world (*represented* world) has to do with cognition, the image (*representing* world) will have most of cognition's characteristics.' Constructing cognitively-informed representations is no trivial problem, even within very confined domains. Computational systems operate over representations of the real world, and have no access to meaning beyond what is formally defined within the representation language. In other words,

a 'representation is only syntax and should have all knowledge embodied in this syntax' (Honing 1993, pp. 224–225). We conjecture that perceptual groundedness is one of the most important aspects for any cognitively-motivated representation of music. The importance of similarity in mental processing is long established (Shepard 1987), and is the guiding principle underlying the computational theories proposed in this thesis. In short, representations that afford the efficient and flexible manipulation of conceptual similarity may prove generally useful in modelling music. In order to pursue this task, we employ the representational framework of conceptual space.

Peter Gärdenfors (2000) proposes the theory of conceptual space as a geometric form of representation, which can be viewed as being situated between the levels of sub-symbolic representation and symbolic representation. The theory states that concepts, which are entirely mental entities, can be represented within sets of dimensions with defined geometrical, topological or ordinal properties. This formalism places *betweenness* at its core, upon which the notion of conceptual similarity is derived. Similarity between concepts is implicitly represented in terms of the distance between points or regions in a multidimensional space, in a manner comparable to the spatial view of similarity proposed by Shepard (1962a, 1962b).

### 2.3.1 Definitions

Gärdenfors' theory of conceptual space begins with an atomic but general notion of *betweenness*, in terms of which he defines *similarity*, represented as (not necessarily Euclidean) distance. This allows models of cognitive behaviours (such as creative ones) to apply geometrical reasoning to represent, manipulate and reason about concepts. Similarity is measured along *quality dimensions*, which 'correspond to the different ways stimuli are judged to be similar or different' (Gärdenfors 2000, p. 6). An archetypal example is a colour space with the dimensions hue, saturation (or chromaticism), and brightness. Each quality dimension has a particular geometrical structure. For example, hue is circular, whereas brightness and saturation correspond with measured points along finite linear scales. Identifying the characteristics of a dimension allow meaningful relationships between points to be derived, and it is important to note that the values on a dimension need not be numbers—though how an appropriate algebra is then defined is not discussed.

Quality dimensions may be grouped into *domains*. A domain is a set of *integral* (as opposed to *separable*) dimensions, meaning that no dimension can take a value without every other dimension in the domain also taking a value. Therefore, hue, saturation, and brightness in the above colour model form a single domain. A

domain is also equipped with a distance measure, which may be a true metric, or non-metric, such as a measure based on an ordinal relationship or the length of a path between vertices in a graph. It follows that Gärdenfors' definition of a conceptual space is simply 'a collection of one or more domains' (2000, p. 26). For example, a conceptual space of elementary coloured shapes could be defined as a space comprising the above domain of colour and a domain representing the perceptually salient features of a given set of shapes.

Since the quality dimensions originate in betweenness, similarity is directly related to proximity, though not necessarily Euclidean proximity. Such spatial representations naturally afford reasoning in terms of spatial regions. For example, in the domain of colour, one can identify a region that corresponds with the concept RED. Boundaries between regions are fluid, an aspect of the representation that may be usefully exploited by creative systems searching for new interpretations of familiar concepts.

Gärdenfors identifies various types of regions with differing topological characteristics. *Convex* regions allow us to define *natural* properties:

**CRITERION P** A *natural property* is a convex region of a domain in a conceptual space. (Gärdenfors 2000, p. 71)

Again taking the example of RED in the domain of colour: given any two shades of RED, any shade between would also be RED. Therefore, the region corresponding to RED must be convex. These convex regions in conceptual domains can be closely related to basic human perceptual experience.

For relatively straightforward domains such as the above three-dimensional domain of colour, we can think of concepts as natural properties. However, more complex concepts, such as coloured shapes or metrical structure, may exist over multiple domains. To admit these more complex structures, Gärdenfors defines a *natural concept* as follows:

**CRITERION C** A *natural concept* is represented as a set of regions in a number of domains together with an assignment of salience weights to the domains and information about how the regions in different domains are correlated. (Gärdenfors 2000, p. 105)

Our interpretation of Criterion C is that a natural concept is a set of one or more natural properties and salience weights for the dimensions and domains, and information about how they are correlated. Semantic distance between concepts is determined by calculating the distance between points in the space. For purely numerical dimensions, as a rule of thumb Gärdenfors suggests Euclidean distance is appropriate for integral dimensions, while the city-block metric is appropriate

for separable dimensions (Gärdenfors 2000, pp. 24–26).

### 2.3.2 Levels of representation

Gärdenfors (2000) presents the theory of conceptual space as a representational tool for approaching problems concerning the modelling and understanding of cognitive processes. As a representation, it is situated at a particular level of abstraction. He argues that conceptual structures should be represented using geometry on what he terms the *conceptual level*. This level of representation is situated between the *symbolic level*, which includes, for example, formal grammar, and the *sub-conceptual level* of high-dimensional representations such as neural networks.

Symbolic representations are used within cognitive science to model cognitive processes at a high level of abstraction. Discrete symbols, representing objects, properties of objects, relationships, concepts and so forth, are precisely defined, constituting the semantics of the representation. Sets of rules can also be defined, operating over and manipulating the symbolic language. Within a symbolic representation, meaning is internal to the representation itself; symbols have meaning only in terms of other symbols, and not in terms of any real world objects or phenomena they may represent.

Symbolic representations are often associated with Good Old Fashioned AI (GOFAI), yet symbolic representation in itself does not entail classic GOFAI methodology, and plays a key role in contemporary cognitive science. An underlying assumption of GOFAI research is that human thinking can be understood in terms of symbolic computation, in particular, computation based on formal principles of logic. Expert systems are one particularly successful outcome based on symbolic representations. A typical expert system will have access to a large body of symbolically encoded knowledge, over which it is able to operate in order to solve specific problems or to derive new facts. However, symbolic systems have proved less successful in modelling aspects of human cognition beyond those closely related to logical thinking. For example, cognitive processes closely related to perception tend to require extremely large numbers of rules in order to account for the vast range of perceptual input that a system situated in a real environment may encounter. As well as being an implausible model for brain processing, such systems can become brittle, having limited ability to adapt their behaviour to new input.

Conventionally termed sub-symbolic representations include artificial neural networks, or more generally, connectionist representations. Connectionist approaches seek to model cognitive processes by exploiting the emergent properties

of densely connected networks of primitive units. A particular strength of connectionist networks is their ability to adapt their behaviour according to observed data. However, since the learned behaviour is represented as weightings between units in the network, they offer limited explanatory insight into the process being modelled.

Gärdenfors acknowledges the strengths and weaknesses of different forms of representation, but makes the more general point that different representational formalisms should be seen as complementary, rather than competing, methodologies. As such, choices of representation should be made in accordance with scientific aims and in response to the challenges of the particular problem at hand. Furthermore, Gärdenfors takes the view that the conceptual level can unify traditional symbolic and sub-symbolic representations, providing a means to develop hybrid representations that combine the strengths of the various approaches, as investigated by Aisbett and Gibbon (2001). A particular strength of a conceptual spaces representation is its ability to offer a parsimonious account of concept combination and acquisition (Gärdenfors 2004, pp. 114–126), both of which are closely related to conceptual similarity. By defining dimensions in terms of perceptual qualities, conceptual space representations are grounded in our experience of the physical world, providing a semantics closely aligned with a human sense of meaning. Therefore, hybrid representations comprising mappings between sub-symbolic (or more appropriately sub-conceptual), conceptual, and symbolic forms of representation creates the possibility for grounding traditional representational formalisms within a cognitive semantics framework.

### 2.3.3 Music and conceptual space

The theory of conceptual space is closely aligned with cognitive semantics, which proposes that 'meanings are mapping to conceptual structures, which themselves refer to real-world entities' (Raubal 2004, p. 154). Gärdenfors' initial theory of conceptual space concentrates primarily on tangible properties and concepts, where quality dimensions typically relate to attributes directly available to our sensory system, for example, the colour of apples. Although musical phenomena are closely linked with physical events in the world, which are experienced via the senses, musical understanding cannot be equated with the physical stimuli themselves. One consequence of this for any representation of musical understanding is the necessity for relatively abstract quality dimensions—relative at least to the dimensions required to represent tangible physical properties or concepts.

Music, and art in general, are interesting application domains in which to in-

vestigate Gärdenfors' theory of conceptual space, not least because of the primacy of subjective experience within them. Music is notoriously difficult to describe with language, although humans have very little difficulty distinguishing between music and non-music when heard. Music exists over time, and there is a delicate interplay between what has gone before, and what might come next (Pearce and Wiggins 2006). Furthermore, all musical experiences are shaped by past musical experiences. Despite a long tradition within musicology of concentrating primarily on notated musical structure, particularly within the analysis of Western music, Gabrielsson (1993) points out that there is often considerable overlap between the perspectives of music theory and music psychology. Therefore, much insight into music conceptualisation can be gained from music theory, and usefully for the theory of conceptual space, music theory offers a vocabulary for distinguishing between what may be relatively abstract concepts, and provides clues as to possible structures of quality dimensions within which they may be represented—which may then be tested empirically.

In principle, an approach to the representation of music based on conceptual space should not need to be confined to any one specific conceptualisation of music. In fact, the conceptual space theory itself supports an elegant model of learning, which accords with evolutionary views of musical development (Bown and Wiggins 2009), in which the process of developing understanding of unfamiliar concepts is modelled by extending a conceptual space with additional quality dimensions, affording greater discrimination between novel stimuli. Furthermore, the notion of dimensional salience, modelled by weightings associated with quality dimensions and domains, allows for the possibility of adapting a conceptual space to take into account individual musical backgrounds and experience. Therefore, it is assumed that differing conceptualisations of music, such as those evident across Western classical or pop music, Balkan folk music, or Ghanian drumming (Patel 2008, pp. 97–99), can be represented consistently within the conceptual spaces theory.

To briefly consider music in relation to natural language, a cognitive semantics view of natural language implies that the meaning of words and sentences cannot be understood as a direct mapping between symbols and real-world objects, but is something that is inherently mediated by human minds. Within linguistics, a distinguished set of *concrete words* is often defined, where word meaning is understood in terms of the 'perceived physical attributes or properties associated with the referents of words' (Andrews et al. 2009, p. 463). Alternatively, the meaning of more abstract concepts, such as *truth* or *virtue*, cannot be defined in terms of

real world referents. In these cases, meaning is much more related to the function of the words within the language itself. In this sense, meaning is intralinguistic—context and the relationships between words is the primary determinant of meaning. Such meaning can be derived from an analysis of the statistical distribution of words in a suitable corpus. Andrews et al. (2009) argue that both the relationship between words and real-world objects, and those between words themselves, play necessary roles in the learning of the meaning of words in natural languages, with abstract words affording generalisation, and concrete words serving to connect, or ground, more abstract concepts in direct sensory experience.

In the case of music—that is, music as perceived—the issue of representational grounding is mediated by subjective experience, since the phenomenon itself, or any 'objects' which one might consider meaningful, are primarily psychological or cultural in nature. Therefore, the semantics of any representation of musical responses to patterns of sound events in the world has more in common with the representation of abstract concepts in language, rather than concrete words. This is further evidenced by the success of statistical methods in modelling and predicting musical behaviour (Pearce and Wiggins 2004).

To be explicit, the purpose of the representational theory pursued here should not be confused with the representation of musical scores or of physical musical sound. In both these cases, which are sometimes ambiguously referred to as 'music', there are clear concrete referents, whose correspondence with the cognitive constructs may not be straightforward. The aim of the conceptual space representations in chapters 4 and 5 is precisely to capture the cognitive constructs associated with aspects of musical experience, and in principle the theory of conceptual space offers a viable approach for representing such fluid, yet richly structured phenomena.

### 2.3.4 Formalisation

Two approaches to the mathematical formalisation of Gärdenfors' theory of conceptual space appear in the literature, both building on an initial formalisation by Aisbett and Gibbon (2001). One strand of research, based on fuzzy set theory, is presented in detail by Rickard et al. (2007b), drawing on previous work by Rickard (2006) and Rickard et al. (2007a). Another strand of research, employing vector spaces, is presented by Raubal (2004), with subsequent related work by Schwering and Raubal (2005a, 2005b), and Raubal (2008a, 2008b).

The vector space formalisation is followed in this thesis. Firstly, we are interested in developing the theory of conceptual space in a direction suitable for the

representation of melodic sequences. Towards this end, we build on the work of Typke (2007) employing the Earth Mover's Distance measure (Rubner et al. 2000), which is a measure of distances between sets of points in a vector space model. Secondly, a vector space formalism provides us with a single general formalism with which to define both the symbolic point set representation underlying the pattern discovery algorithm developed in chapter 3, as well as the conceptual-space representations of melodic structure and metrical-rhythmic structure developed in chapters 4 and 5 respectively.

# Chapter 3

# Point set methods of pattern discovery

This chapter concerns the problem of identifying perceptually salient instances of repetition in symbolically represented polyphonic music. A geometrical approach is adopted in which pieces of music are represented as multidimensional datasets. Following the work of Meredith et al. (2002), Meredith et al. (2003), and Meredith (2006), we have implemented SIATEC, a pattern induction algorithm, and investigate the properties of the generated results in terms of musicological value and perceptual salience. SIATEC is known to discover many more patterns than are typically of interest to any musical analysis. In fact SIATEC guarantees the enumeration of *all* instances of maximal-length patterns that occur more than once within a given dataset. This formally defined pattern type has interesting characteristics that may be exploited for the development of computational methods to assist in the analysis of musical structure.

A known problem with SIATEC is the volume of the discovered patterns, which can be difficult to interpret (Meredith et al. 2002, p. 340). We propose a post-processing step, similar in character to the NP-hard minimum-weighted set-cover problem (Karp 1972), in which various heuristics can be employed in order to optimise the results in terms of specific music-analytic objectives. The generic set-cover problem informally involves two finite sets: a set of basic elements; and a set of candidate subsets of basic elements. A solution to the problem involves finding the smallest number of candidate subsets that together contain, or *cover*, all basic elements. In the present musical context, musical events are the basic elements, and SIATEC patterns are the candidate subsets of elements. The minimum-weighted variant of the generic set-cover problem introduces a weight or cost associated with each candidate subset, and the objective is to minimise the total weight

of a cover solution. In our case, weight is equated with structural salience, and the objective is to *maximise* the total weight of a cover.

Set-covering problems are NP-hard, meaning that an exact solution cannot be computed in polynomial time. A considerable amount of research, both theoretical and applied in focus, has been conducted in trying to establish methods for deriving solutions to set-covering problems within acceptable bounds of approximation. The standpoint of computer science informs understanding of the nature of this problem, and provides examples of rigorously tested methods that may be applicable to our case. The necessity for approximation within this approach provides the opportunity for involving domain-specific musically-informed heuristics, which themselves may be parameterised to achieve a range of analytic objectives. Furthermore, the inherent ambiguity in set-covering problems accords with the common situation in musical analysis whereby different interpretations of a work may be considered equally valid and correct. In order for an analyst to reach any firm conclusion, compromises must be made, which are often informed by conventions (heuristics) of music theory.

An applied aim of this research is to develop tools suitable for various music-analytic tasks. Within the field of musicology such tools may assist conventional score analysis, and may prove particularly useful for larger-scale corpus analysis. The latter overlaps with interests of music information retrieval, where such techniques may be applied in order to extract commonly occurring patterns as the basis for classification. A composer may also be able to gain inspiration by analysing a work in progress, arriving at a fresh perspective. Novel applications may also be found in music psychology or artificial intelligence, where large collections of music could be analysed in order to derive data for training or testing models of musical behaviour.

## 3.1 Algorithmic approaches to structure induction

The concept of a musical pattern entails repetition. The definition of SIATEC ensures the enumeration of all maximal repeated patterns (Meredith et al. 2002, pp. 331–333). A large number of these discovered patterns will usually prove to be of little interest from a musical or perceptual perspective, and this is one problem our heuristics must address. Yet a more complicated issue concerns the many *types* of salient repetitive patterns that may exist in a musical work. In other words, the kinds of patterns that are likely to be of interest, and the ways in which they are interesting, may vary considerably.

There is agreement amongst both musicologists and music psychologists as to the importance of repetition in music (for example Jones 1981; Lerdahl and Jackendoff 1983; Nattiez 1990; Krumhansl 1997). One cross-cultural study based on fifty musical works found that 94% of all musical passages longer than a few seconds in duration were repeated at some point in the work (Huron 2006, pp. 228–229). However, this result does not account for the role of repetition in music in its entirety, because repetition may exist in many forms beyond the exact repetition of musical events in sequence. For example, melodies may still be perceived as instances of the same basic melodic motif despite being transposed in pitch or scaled in time. Indeed, perceptual similarity may pertain for any individual listener under an arbitrary number of processes of elaboration and transformation. In the context of computational analysis, therefore, careful consideration must be given to the notion of pattern equality.

Much previous work in this area has concentrated on techniques for string matching, with considerable successes in certain specialised tasks, notably concerning monophonic melodies. However, in the wider context various limitations of string methods become apparent, particularly in the case of polyphonic music as considered here (Lemström and Pienimäki 2007).

An alternative approach to string matching exists in the form of geometrically-based algorithms. Within a geometrical framework, the individual note events of a piece of music correspond to single points in an ordered vector space.[1] A family of *Structure Induction Algorithms* have been developed for pattern discovery and matching in multidimensional datasets by Meredith et al. (2002), Wiggins et al. (2002), Meredith et al. (2003), and Meredith (2006). The initial development of these techniques was motivated to a large extent for application to music, but are equally applicable in other domains where objects may be adequately represented in a multidimensional space.

Following Meredith et al. (2002, p. 328), we define a datapoint as a $k$-tuple of real numbers, and a pattern $P$ or dataset $D$ as a finite set of $k$-dimensional datapoints. We reserve the term *dataset* in this chapter to refer to a complete set of datapoints we wish to process, for example, a piece of music, while *pattern* refers to a subset of a dataset. A typical datapoint representation in a musical context will include dimensions representing time and pitch attributes of musical events. A translator is a vector that maps from one instance of a pattern to another within a dataset. More precisely, a vector $\mathbf{t}$ is a *translator* for $P$ in $D$ if and only if the translation of

[1] A variation on this approach, based on sets of line segments in space, is discussed by Ukkonen et al. (2003).

$P$ by $\mathbf{t}$ is also a subset of $D$.

The basic SIA algorithm computes all maximal repeated patterns in a dataset (Meredith et al. 2002, pp. 334–335). The algorithm finds the largest non-empty set of translatable datapoints for every positive translation possible within the dataset. Hence, each pattern discovered by SIA is called a *maximal translatable pattern* (MTP). For $n$ datapoints, the worst case running time of SIA is $O(kn^2 \log_2 n)$ and its worst-case space complexity is $O(kn^2)$.

An important extension to SIA is SIATEC (Meredith et al. 2002, pp. 335–338). SIATEC underlies both the approach to pattern discovery adopted in the present chapter, as well as the closely related COSIATEC algorithm, which will be discussed below. Like SIA, the SIATEC algorithm enumerates all the maximal translatable patterns in a dataset, but also groups them into equivalence classes. A *translational equivalence class* (TEC) is represented compactly as an ordered pair $\langle P, T(P, D)\rangle$, where $P$ is a maximal translatable pattern, and $T(P, D)$ is the set of translators for $P$ in dataset $D$. The worst-case running time of SIATEC is $O(kn^3)$, and its worst-case space complexity is $O(kn^2)$.

Even for small datasets, the raw output of SIATEC can quickly become unmanageably large. Table 3.1 shown the number of TECs discovered by SIATEC within J. S. Bach's Two-Part *Inventions*, where the notes of each piece are represented within a two-dimensional space of onset and morphetic pitch.[2] Furthermore, the patterns are diverse in size and structure, and on the whole are not readily intuitive. It would be straightforward to rank the discovered patterns based on a set of criteria; for example, to sort by pattern size $|P|$, or the number of pattern repetitions $|T(P, D)|$. However, such a simplistic approach presents two particular difficulties. Firstly, the method does not lend itself to any principled means of deciding how many of the most highly ranked patterns should be selected as being representative of the repetition in the dataset. Secondly, this method would preclude the ability to make inter-pattern judgments, that is, for the value of one pattern to influence the value of another, due to combinatorial explosion.

COSIATEC is one method for automatically identifying a subset of 'interesting' patterns from amongst the many patterns discovered by SIATEC (Meredith et al. 2003; Meredith 2006). COSIATEC is designed to generate compressed representations of datasets by representing them in terms of highly repetitious subsets. The algorithm first runs SIATEC, generating a list of $\langle P, T(P, D)\rangle$ pairs, and then selects the best pattern based on heuristics. The algorithm then removes from the orig-

[2]Morphetic pitch represents the position of a note head on a staff. Both onset and morphetic pitch are defined in section 3.2.1.

Table 3.1: Number of notes (`onset` × `mpitch` datapoints) and the number of discovered TECs in J. S. Bach's Two-Part *Inventions*.

| Composition | Number of datapoints | Number of TECs |
|---|---|---|
| BWV 772 | 458 | 9035 |
| BWV 773 | 634 | 11724 |
| BWV 774 | 494 | 9882 |
| BWV 775 | 443 | 9304 |
| BWV 776 | 733 | 15978 |
| BWV 777 | 547 | 17209 |
| BWV 778 | 473 | 11103 |
| BWV 779 | 598 | 11731 |
| BWV 780 | 558 | 11995 |
| BWV 781 | 439 | 9038 |
| BWV 782 | 568 | 11306 |
| BWV 783 | 685 | 15969 |
| BWV 784 | 564 | 12250 |
| BWV 785 | 592 | 16782 |
| BWV 786 | 477 | 10407 |

inal dataset all the datapoints that are members of the occurrences of the chosen pattern $P$. The process continues until all the datapoints have been removed from the dataset. The resulting set of patterns are collectively termed a *cover* (Wiggins et al. 2002). In this case, each datapoint is represented in a cover exactly once.

For each iteration of `COSIATEC`, the remaining patterns are evaluated according to three heuristic measures: *coverage*; *compression ratio*; and *compactness*. The most highly valued pattern according to a factor combining these measures is selected to become part of the resulting cover. Compression ratio is also employed in the present work, along with a music-specific variant of compactness, which incorporates voicing information. Both heuristic measures are defined in section 3.2.3. A variant of coverage is also employed here, specific to our formulation of pattern selection as a weighted set-cover problem, and is defined in section 3.2.2.

Although motivated by compression, `COSIATEC` has been shown to identify principal musical themes in pieces of music (Meredith et al. 2003; Meredith 2006). This is explicable given the very nature of a musical theme, which will typically appear numerous times during a work, making it an ideal pattern for use in the encoding of a compressed representation. Therefore, `COSIATEC` offers a tidy solution to the difficulties of sifting through the output of `SIATEC`. The problem becomes one of generating optimal covers given particular heuristics. Furthermore, being a greedy algorithm, the generation of covers entails a degree of pat-

tern co-dependency, since previously selected patterns will affect the outcome of later iterations.

## 3.2 A novel method for the identification of salient patterns

The approach to pattern discovery in the present work follows a similar strategy to that of COSIATEC. We again formulate the problem as one of cover generation, but explore possibilities created by shifting the emphasis away from purely optimising compression. The foremost difference in this approach is that we only apply SIATEC once—to initially process the entire dataset. Furthermore, we evaluate the structural salience of each pattern discovered by SIATEC only once, prior to cover generation. Therefore, the value of each pattern, with respect to a particular music-analytic focus, is determined within the same initial context prior to selection. The rationale being that, in contrast to COSIATEC, our selection process more closely relates to the process of musical listening, since listeners perceive patterns in a musical work in the context of all the notes. The heuristics used for determining structural salience are based on measures of compression and compactness. These values, for each candidate pattern, are scaled each iteration of a greedy selection algorithm by a single varying factor equal to the number of currently uncovered datapoints that are covered by the pattern.

In further contrast to COSIATEC, our cover generation method relaxes the constraint requiring that each datapoint be included in a cover exactly once, enabling us to consider datapoints as potentially belonging to multiple patterns within a single cover. This creates the opportunity to make connections between patterns based on intersecting elements, with the intention of revealing structural relationships between sets of repeated elements. A similar strategy could be pursued within COSIATEC, but only within the context of each iteration as covered datapoints are removed at each stage. We first introduce additional notational elements, then state the problem formally as an instance of the generic weighted set-covering problem, before finally defining the pattern evaluation heuristics used to determine structural salience.

### 3.2.1 Definitions

To distinguish between the compact representation of TECs generated by SIATEC, $\langle P, T(P, D)\rangle$, and their set-theoretic definition, we introduce the symbol $\mathcal{T}$ to de-

note a set of translationally equivalent pattern occurrences. Set $\mathcal{T}$ can be thought of as the enumeration of the compressed representation of TECs generated by SIATEC, i.e., each set $P \in \mathcal{T}$ is an explicit representation of a maximal translatable pattern belonging to a particular TEC.

**Definition 3.1.** The set $\mathcal{T}$ is a set of maximal, translationally equivalent pattern occurrences, containing all instances of a pattern discovered by SIATEC belonging to a single translational equivalence class.

Following the method of cover generation introduced in COSIATEC, covers are constructed at the level of TECs, meaning that when a TEC is selected to become part of the cover, the datapoints belonging to each occurrence of the pattern $P \subset D$ are considered covered. Therefore, we introduce the symbol $P'$ to denote the subset of $D$ that is covered by the union of all occurrences of pattern $P \in \mathcal{T}$.

**Definition 3.2.** The set $P' \subset D$ is the union of a set of translationally equivalent patterns $P \in \mathcal{T}$.

$$P' = \bigcup_{P \in \mathcal{T}} P$$

We also assume a symbolic musical surface containing the following event attributes:

- onset, the score-time (quantised) representation of event onset measured in crotchets, where a crotchet is equal to one time unit;
- mpitch, the morphetic pitch representing the position of a note head on a staff, where $\text{A}\natural_0$ is defined to be 0. Therefore, the morphetic pitch of middle-C ($\text{C}\natural_4$) is 23;
- voice, a binary value indicating whether an event is notated in the left-hand (= 0) or the right-hand (= 1) part.

### 3.2.2 Set-cover generation

The approach taken here for the generation of covers from musical patterns discovered by SIATEC can be described in terms of the widely known NP-hard set-covering problem. Cormen et al. (2001, pp. 1033–1034) state this problem as follows.

> An instance $(X, \mathcal{F})$ of the ***set-covering problem*** consists of a finite set $X$ and a family $\mathcal{F}$ of subsets of $X$, such that every element of $X$ belongs to at least one subset in $\mathcal{F}$:

$$X = \bigcup_{S \in \mathcal{F}} S. \tag{3.1}$$

We say that a subset $S \in \mathcal{F}$ ***covers*** its elements. The problem is to find a minimum-size subset $\mathcal{C} \subseteq \mathcal{F}$ whose members cover all of $X$.

$$X = \bigcup_{S \in \mathcal{C}} S. \tag{3.2}$$

In other words, the desired outcome of this optimisation problem is to find the smallest number of subsets in $\mathcal{F}$ that account for (cover) each element in $X$ at least once.

Considering the generic set-cover problem in the context of SIATEC, set $X$ is equivalent to a dataset $D$. As in the selection process of COSIATEC, we consider all datapoints that are members of pattern occurrences in $\mathcal{T}$, as constituting a single subset of $D$, defined above as $P'$ (3.2). Therefore, $\mathcal{F}$ is equivalent to the entire set of $P'$ subsets of $D$ discovered by SIATEC: $\mathcal{F} = \left\{P'_0, P'_1, \ldots, P'_{n-1}\right\}$. A set-cover solution $\mathcal{C}$ is equal to the set of $P'$ subsets of $D$ that optimally cover $D$.

The standard approach to set-cover generation utilises a greedy algorithm based on the heuristic of selecting within each iteration the set $S$ that covers the largest number of currently uncovered elements in $X$. If this heuristic results in a tie, the algorithm randomly selects a single subset from amongst the best-rated candidate subsets to become part of the cover solution. This algorithm has an approximation ratio of $\ln n$ (Johnson 1973), meaning that the ratio between the size of the discovered set-cover and the optimum cover is bounded by the natural logarithm of the size, $n$, of the input set. Feige (1998) shows that the approximation ratio of $\ln n$ guaranteed by the greedy algorithm, among others, is the lower-bound for polynomial time algorithms.

In the context of SIATEC, we define the above greedy selection heuristic in terms of $\text{coverage}_\text{i}$ (3.4), based on the definition of coverage introduced for COSIATEC. Meredith et al. (2003, p. 7) define coverage as 'the number of datapoints in the dataset that are members of occurrences of the pattern', which we can state equivalently here as the cardinality of $P'$ (3.3).

$$\text{coverage} = |P'| \tag{3.3}$$

Since set $\mathcal{C}$ is constructed incrementally within a greedy strategy, we must update the coverage of each candidate pattern each iteration of the algorithm. We denote

a partial set-cover solution $\mathcal{C}_i$, where $i \in \mathbb{Z}^+$ indexes the iterations of the selection algorithm. We then define coverage$_\text{i}$, where $i \in \mathbb{Z}^+$ again indexes algorithm iterations, as an extension of coverage taking into account the set of previously selected patterns $P' \in \mathcal{C}_i$.

$$\text{coverage}_\text{i} = |P' \setminus \cup_{P' \in \mathcal{C}_i} P'| \tag{3.4}$$

In the case $\mathcal{C}_1 = \emptyset$, where $\emptyset$ denotes the empty set, our definition of coverage$_\text{i}$ is equivalent to coverage as defined by Meredith et al. (2003). To generate a set-cover, the greedy algorithm therefore selects, for each iteration $i$, the pattern with the greatest coverage$_\text{i}$.

We adopt a greedy strategy here, but for our purpose, simply finding a minimum-sized subset $\mathcal{C} \subseteq \mathcal{F}$ by maximising coverage$_\text{i}$ does not adequately characterise the problem: we require an additional means of specifying which patterns should be considered better or worse by the selection algorithm in terms of their musical characteristics. Therefore, a more appropriate model for the problem is the equally well-known generalisation of the minimal set-cover problem: the minimum-weighted set-cover problem (Chvatal 1979). Typically, a greedy algorithm is also adopted, except that covering subsets are selected in the order that minimises the ratio of cover weight to number of elements covered.

To place the SIATEC cover problem in this context, it is necessary to attach weighting values to each of the discovered patterns in $\mathcal{F}$. This step is similar to the use of heuristics in COSIATEC, except that in this case the values are calculated only once, prior to the actual selection process. The higher a pattern scores according to a heuristic, the more relevant it is considered to be to the analysis. The heuristics used to calculate these values are discussed in the following section. Contrary to the more typical formation of weighted set-cover problems, the selection process in this case seeks to maximise weight thus,

$$\text{coverage}_\text{i} \cdot \text{weight} \tag{3.5}$$

where weight is equal to the structural salience of a pattern with respect to the evaluation heuristics defined in section 3.2.3.

A minimum coverage$_\text{i}$ threshold, which must be exceeded for a pattern to become a member of the cover, has proved a useful parameter in the generation of set-covers. In practice, a minimum coverage$_\text{i}$ threshold of between 10 and 30 datapoints is the typical useful range. Higher values in this range are particularly useful in order to generate covers consisting of only a small number of patterns.

High coverage$_i$ values may lead to not every datapoint in $D$ being represented in the set-cover solution. However, this is not necessarily unsatisfactory, since not every note in a piece of music is necessarily part of a repeated pattern.

Once it has been determined that a pattern should become a member of the set-cover, a final step is taken to determine whether a pattern should be considered a *primary* or *secondary* pattern. This step is simply intended to make the generated results more comprehensible for the human analyst, by attempting to group together similar patterns. If a pattern is the first pattern to be selected, it is simply defined as primary. Each subsequently selected pattern is compared to each existing primary pattern in terms of the number of datapoints they commonly cover. This is in order to identify the primary pattern that is 'most similar' to the newly selected pattern, quantified in terms of overlapping coverage. If the proportion of commonly covered datapoints is greater than an arbitrarily defined threshold—50% in this case—then the newly selected pattern is declared a secondary pattern, and grouped together with the most similar primary pattern. If the newly selected pattern is not considered similar to any of the other primary patterns it is declared a primary pattern. Whether a pattern is defined as primary or secondary has no bearing on the actual selection process, it is purely a means of organising the selected patterns, as well as offering an estimation of the number of distinct musical ideas present in the work.

### 3.2.3 Pattern evaluation heuristics

As noted above, there may be many different forms of repetition in a piece of music. It is therefore necessary to establish evaluation criteria in order to automate the extraction of the kinds of repetitions that are considered relevant to an analytical objective. Here we describe two heuristics that are used to provide static, or absolute, measures of the structural salience of patterns prior to cover generation.

*Compression ratio* is defined as 'the compression ratio that can be achieved by representing the set of points covered by all occurrences of a pattern by specifying just one occurrence of the pattern together with all the non-zero vectors by which the pattern is translatable within the dataset' (Meredith 2006, p. 13). Compression ratio can, therefore, be stated as follows.

$$\text{compression ratio} = \frac{|P'|}{|P| + |T(P, D)| - 1} \tag{3.6}$$

Compression ratio is particularly useful for identifying large, non-overlapping patterns that have many occurrences in a dataset.

The second heuristic used, denoted compactness-v (3.13), measures the compactness of a pattern, taking into account given voicing information. Meredith (2006, p. 13) defines a generic notion of compactness as 'the ratio of the number of points in the pattern to the number of points in the region spanned by the pattern'. This measure applies to each *occurrence* of a pattern $P \in \mathcal{T}$. Therefore, unlike compression ratio, which generates a single value for each TEC, there are $|\mathcal{T}|$ compactness values for each TEC.[3] In order to arrive at a single value for a TEC, since the selection algorithm generates covers by selecting $P'$ subsets of $D$, the obvious approach is to use either the mean or maximum compactness value as the TEC weighting value. From a musical perspective, selecting the maximum pattern compactness value to determine the weighting of a TEC can be justified on the principle that a significant musical theme will typically have at least one relatively prominent (compact) occurrence in a work.

As discussed by Meredith et al. (2003) and Meredith (2006), the definition of region, for example, as a segment, bounding box, or convex hull, has implications for computing the value of compactness for any given pattern occurrence. For our purpose, we define a region as a segment of time bounded by the onset timepoints associated with the first and last datapoints of a pattern $P$, as defined in (3.7) and (3.8), where $\mathit{onset}(\mathbf{p})$ gives the timepoint of the onset of datapoint $\mathbf{p}$.

$$\mathit{start}(P) = \min_{\mathbf{p} \in P} \mathit{onset}(\mathbf{p}) \tag{3.7}$$

$$\mathit{end}(P) = \max_{\mathbf{p} \in P} \mathit{onset}(\mathbf{p}) \tag{3.8}$$

Therefore, all datapoints $\mathbf{d} \in D$ occurring inclusively between the start and end timepoints of a segment are said to be members of the region spanned by a given pattern $P$ (3.9).

$$\mathit{segment}(P, D) = \{\mathbf{d} \in D \mid \mathit{start}(P) \leq \mathit{onset}(\mathbf{d}) \leq \mathit{end}(P)\} \tag{3.9}$$

We can now state the generic form of compactness thus.

$$\text{compactness} = \max_{P \in \mathcal{T}} \frac{|P|}{|\mathit{segment}(P, D)|} \tag{3.10}$$

However, unlike previous work, we wish to calculate the ratio using only those notes in the region that are also members of the voices present in the pattern. This decision is based on the assumption that notes belonging to the musical voices

[3] Or equivalently, in terms of the compressed representation of TECs: $|T(P, D)|$.

present in a pattern are more likely to influence its perceptual salience, compared with notes belonging to other musical voices. This assumption is consistent with empirical findings related to melodic streaming (Bregman 1990, pp. 61–64). Such a definition of compactness, relying to a certain degree on specific musical knowledge, is less generic than the original geometrical definition. However, it has proved to be the most satisfactorily performing variant in our exploratory study. Furthermore, our testing dataset consists of the fifteen J. S. Bach *Inventions*. Notwithstanding the caveat that the articulation of distinct voicing in timbrally homogeneous textures, particularly in the case of keyboard instruments, should not be general assumed, the strict two-part texture of the *Inventions* gives some support for utilising voicing information during the selection process.

Extending the above definition of *segment* (3.9), we define *segment-v* (3.11), restricting segment membership to datapoints that belong to a voice also present in the pattern,

$$\begin{aligned} \textit{segment-v}(P, D) = \{\mathbf{d} \in D \mid \quad & \textit{start}(P) \leq \textit{onset}(\mathbf{d}) \leq \textit{end}(P) \wedge \\ & \textit{voice}(\mathbf{d}) \in \textit{voices}(P)\} \end{aligned} \tag{3.11}$$

where $\textit{voice}(\mathbf{d})$ gives the notated voice of datapoint $\mathbf{d}$, and $\textit{voices}(P)$ gives the set of voices in $P$ (3.12).

$$\textit{voices}(P) = \bigcup_{\mathbf{p} \in P} \textit{voice}(\mathbf{p}) \tag{3.12}$$

Our preferred measure of pattern compactness can now be defined thus.

$$\text{compactness-v} = \max_{P \in \mathcal{T}} \frac{|P|}{|\ \textit{segment-v}(P, D)|} \tag{3.13}$$

The values of compression ratio and compactness-v are unit-normalised and independently subject to minimum and maximum thresholds. Simple min-max normalisation (equation (3.14); Jain et al. 2005, p. 2276) is used, which linearly scales the weightings produced by each heuristic, $W_h$, to values between zero and one, $W_h^{'}$.

$$W_h^{'} = \frac{W_h - \min}{\max - \min} \tag{3.14}$$

Minimum and maximum thresholds, defined over the interval $[0, 1]$, and where minimum $<$ maximum, can be applied to the values of each heuristic independently. If the value produced by a heuristic falls beyond the given range, the value defaults to zero. Setting a minimum threshold for a heuristic is particularly useful as a means of excluding a subset of obviously uninteresting patterns prior to

generating a set-cover. The removal of redundant subsets is common in the literature (Caprara et al. 1998, p. 2). In the musical context, excluding very poorly rated patterns prior to set-cover generation may lead covers that are overall more musically interesting, even if a smaller cover may be achieved by including a number of more uninteresting patterns.

Normalised heuristic measures are combined into a final estimation of structural salience by multiplication.

$$\text{weight} = W'_{\text{compression ratio}} \cdot W'_{\text{compactness-v}} \tag{3.15}$$

Therefore, if a pattern is rated maximally by both heuristics, it will have an overall weighting of one. If a pattern is given a rating of zero by either heuristic, the overall weight will also be zero.

## 3.3 Case study: Motivic analysis in Bach two-part *Inventions*

We have applied the cover generation method and heuristics discussed above to J. S. Bach's Two-part *Inventions* (BWV 772–786). Each piece is represented as a set of `onset` × `mpitch` × `voice` tuples. `SIATEC` was applied to the two-dimensional projection of `onset` × `mpitch`, and voicing information was used only during set-cover generation. Each piece was analysed using a range of heuristic thresholds. Findings from the analysis of BWV 772 are reported below. This piece was also subject to analysis by Meredith et al. (2003) and Meredith (2006), and comparisons between the findings are noted below.

There are a total of 9035 TECs discovered by `SIATEC` in BWV 772. 540 of these are patterns of less than three notes, which we exclude from the study on the grounds that at least three notes (a pair of note intervals) is a reasonable minimum requirement for a pattern to be considered as a potentially salient constituent of musical structure. We then calculate the structural salience weight of each of the remaining 8495 TECs according to equation (3.15), applying the following thresholds:

- compression ratio (min: 0.2, max: 1.0);
- compactness-v (min: 0.2, max: 1.0).

Within these thresholds, only 268 TECs are assigned a non-zero weight. This considerable reduction indicates that a very large proportion of the patterns discovered by `SIATEC` are not relevant to this particular analytical focus. Figures 3.1—3.4

show the distribution of structural salience values according to each heuristic measure. Both heuristics rate the majority of patterns as very low in salience, and as can be seen in the sorted plots, the experimentally-determined minimum threshold of 0.2 intersects at a point above which salience dramatically increases.

We then generate a set-cover applying the following threshold:

- $\text{coverage}_i$ (min: 15 datapoints).

Generating a cover with the relatively high $\text{coverage}_i$ threshold of 15 datapoints produces a cover consisting of only six patterns—three primary and three secondary. Setting a lower threshold in this case tends to increase the number of secondary patterns selected, as considerable coverage is achieved by the first two selected primary patterns.

Figure 3.5 shows the first occurrence of each selected pattern in score notation. Patterns 1 and 2, the first and second patterns discovered, are the inversion of the subject, and subject itself respectively. These patterns are the same as those discovered by `COSIATEC`, and are labelled as the subject of the work as analysed by Dreyfus (1996, p. 10).

The two secondary patterns, 1.1 and 2.1, are both clearly subsets of their parent patterns. From an analytical perspective, the most interesting aspect of these patterns is how their individual pattern of occurrence *differs* from that of their parent, as can be seen in the schematic representation of pattern occurrence in figure 3.6. This is particularly apparent for pattern 2.1 in bars 16–20, figure 3.7, where many instances of the pattern overlap. As well as highlighting the high density of this simple descending three-note quaver pattern at the end of the piece, the change in the translations of pattern 2.1 in relation to pattern 2 suggest some sort of developmental change to the primary pattern. In fact, this change corresponds to the note that immediately follows an occurrence of pattern 2, which in these closing bars forms an interval of a 2nd. All previous occurrences, except for the occurrence preceding bar 9, are followed by a larger interval, most commonly a 5th. Hence the overlapping occurrences of pattern 2.1 are not present at these locations.

It cannot really be argued that pattern 1.2 is a perceptually salient pattern when considered as a single occurrence. However, when taking into account the larger pattern that is formed by the overlapping occurrences, an important pattern emerges. Pattern 1.2 is indicative of the gap-fill quaver pattern that accompanies pattern 1 in bars 3–4 in the bass, and in bars 11–12 in the treble, shown in figure 3.8. Pattern 1.2 is also embedded in the structure of pattern 1 itself, shown by the arrows below the semi-quaver passages, each indicating an embedded occurrences of pattern 1.2 in pattern 1.

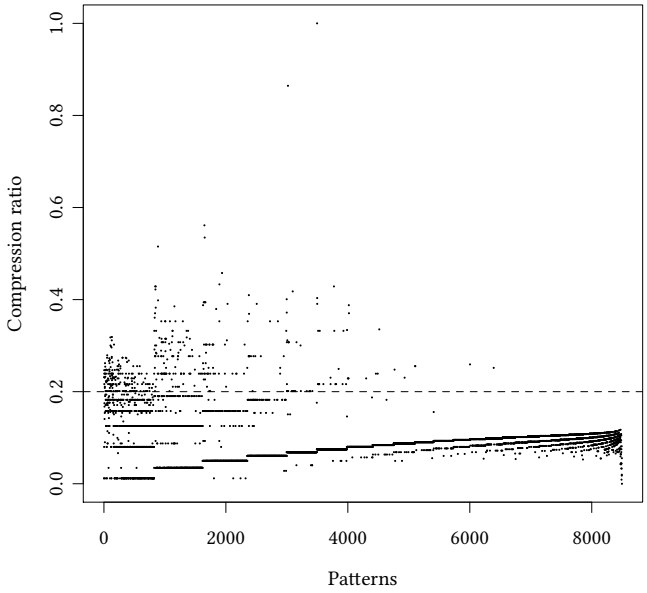



Figure 3.1: Compression ratio values for all three-note and above TEC patterns discovered in BWV 772, presented in the order that patterns are discovered by SIATEC. The dotted line marks the experimentally-determined threshold below which TECs are excluded from the set-cover generation phase.

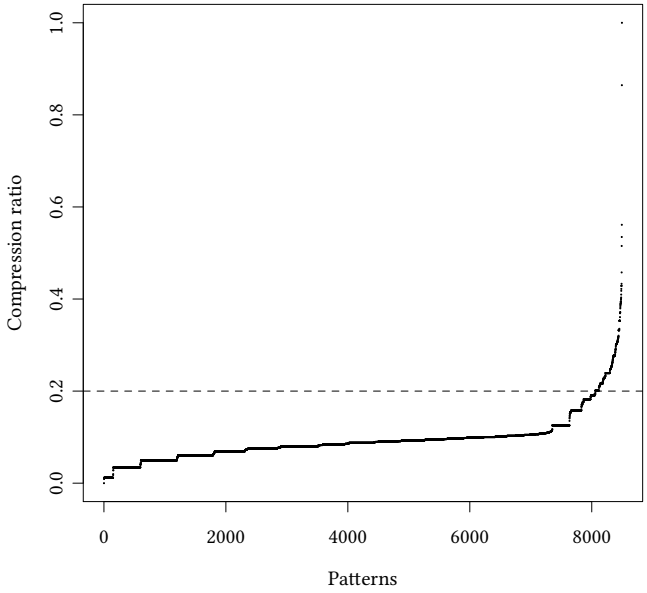



Figure 3.2: Compression ratio values for all three-note and above TEC patterns discovered by SIATEC in BWV 772, sorted by value. The dotted line marks the experimentally-determined threshold below which TECs are excluded from the set-cover generation phase.

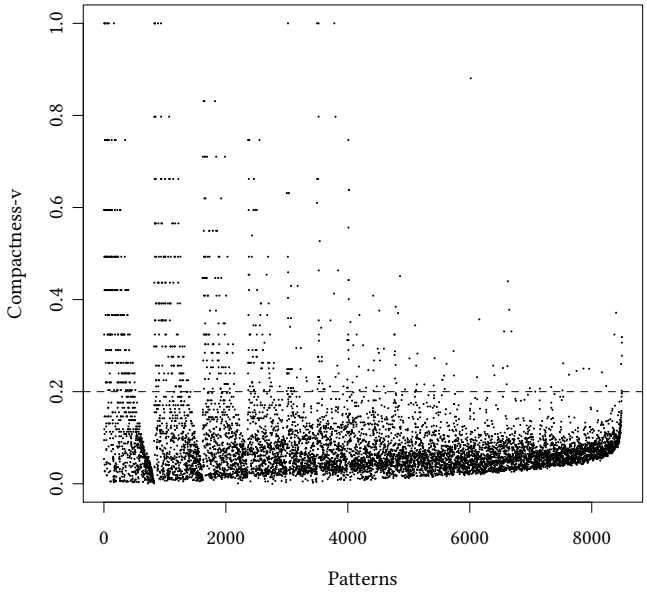



Figure 3.3: Compactness-v values for all three-note and above TEC patterns discovered in BWV 772, presented in the order that patterns are discovered by SIATEC. The dotted line marks the experimentally-determined threshold below which TECs are excluded from the set-cover generation phase.

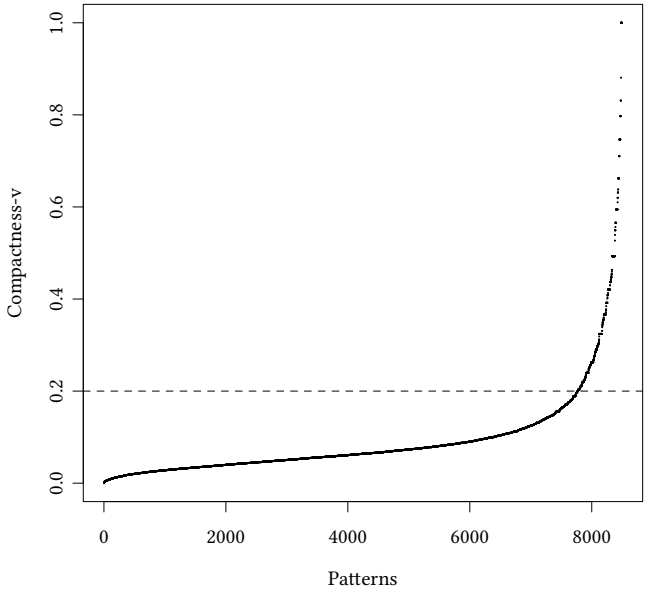
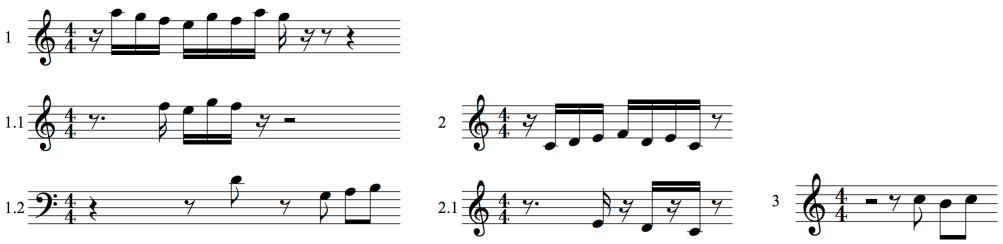



Figure 3.4: Compactness-v values for all three-note and above TEC patterns discovered by SIATEC in BWV 772, sorted by value. The dotted line marks the experimentally-determined threshold below which TECs are excluded from the set-cover generation phase.




a: Pattern 1: bar 3 b: Pattern 2: bar 1 c: Pattern 3: bar 1

Figure 3.5: The primary and secondary patterns selected from the SIATEC analysis of BWV 772. Each extract is shown as a complete bar, and rests have been inserted to indicate the metrical position of each initial occurrence.

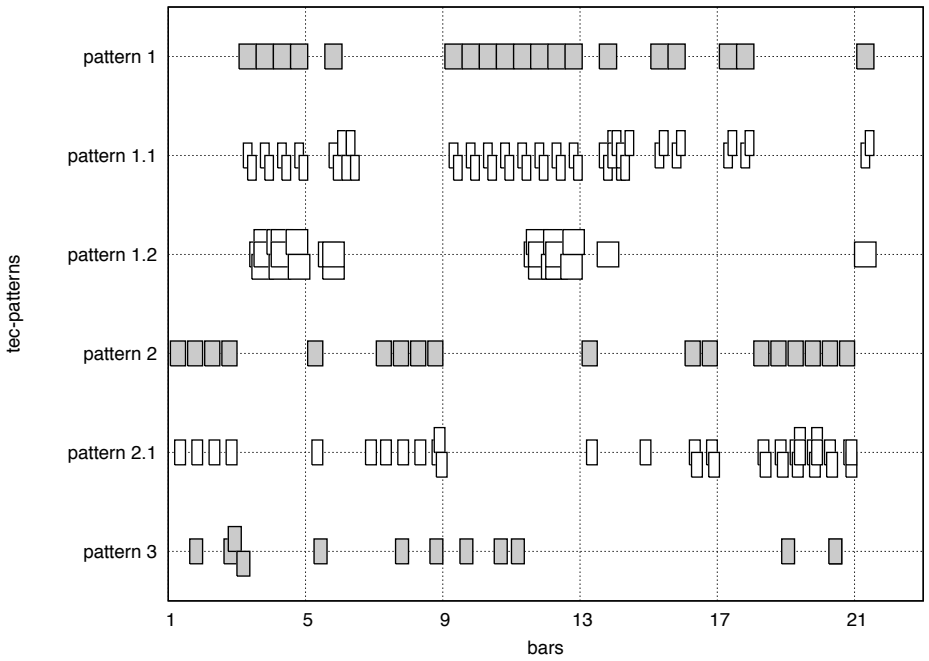



Figure 3.6: A schematic representation of the primary and secondary patterns selected from the SIATEC analysis of BWV 772. The filled boxes are primary patterns, the empty boxes are secondary patterns. Each box represents a pattern occurrence. To aid clarity, patterns that overlap are draw alternately above and below the line.

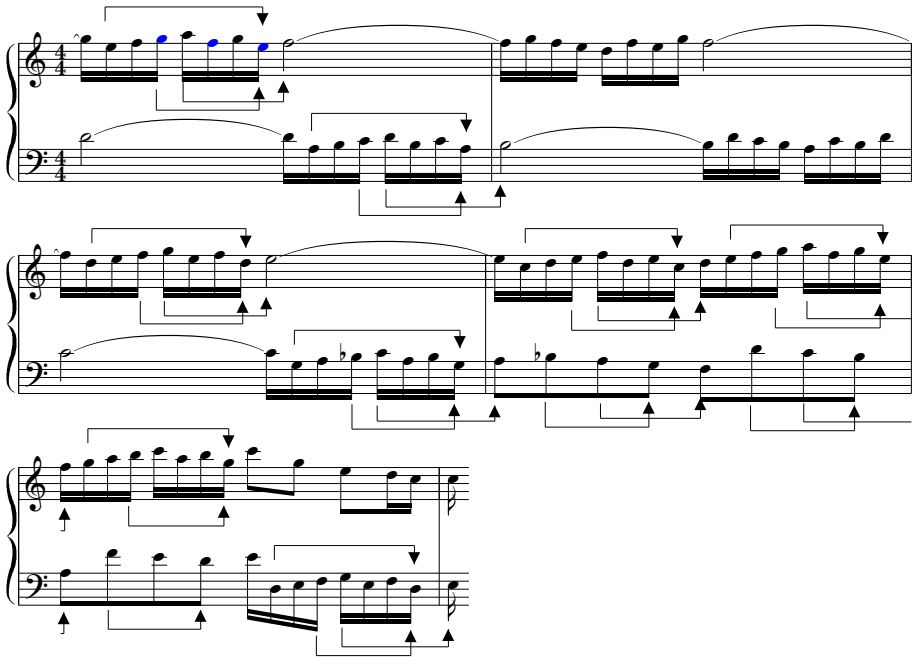


Figure 3.7: Patterns 2 and 2.1 in bars 16–20 of BWV 772. Arrows above each staff indicate occurrences of pattern 2, and arrows below indicate occurrences of pattern 2.1. Blue note heads are also used to highlight the first occurrence of pattern 2.1.

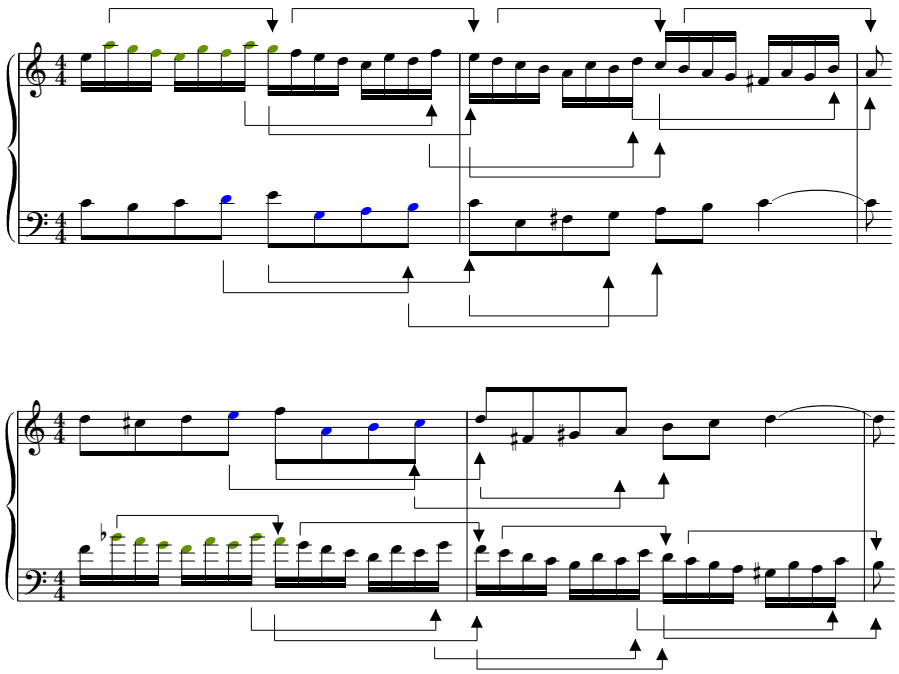


Figure 3.8: Patterns 1 and 1.2 in bars 3–4 (above) and 11–12 (below) of BWV 772. Arrows above each staff indicate occurrences of pattern 1, and arrows below indicate occurrences of pattern 1.2. Green note heads are used to highlight the first occurrence of pattern 1, and blue note heads for the first occurrence of pattern 1.2.

## 3.4 Future work

Applied examples from the literature present several variations on the greedy algorithm that have proved useful in particular domains, which may similarly be beneficial in our case. Marchiori and Steenbeek (1998) describe the `Enhanced Greedy` algorithm, which has a more sophisticated heuristic for breaking ties when adding new covering sets of equal size to the solution. At each iteration the algorithm also checks for (and possibly removes) sets that become 'nearly' redundant in the solution due to the addition of new sets. The `Iterated Enhanced Greedy` algorithm is also described, in which a subset of the currently best (smallest) cover is used as an initial partial solution for a further iteration of the algorithm. Another approach that would also warrant empirical investigation in this context is the multiple weighted set cover problem, which is a further generalisation of the basic set-cover problem where events must be covered a specified minimum number of times (Yang and Leung 2005). Alternative approaches to the basic greedy algorithm, including approximate linear programming and exact branch and bound method, are discussed in Caprara et al. (1998).

## 3.5 Conclusion

A computational method for assisting in the structural analysis of music has been presented. The method is essentially a selection algorithm based on a set of heuristics that attempt to determine the quality of discovered patterns in terms of musical salience. The `SIATEC` algorithm is integral to the process, since it provides the initial set of discovered patterns from which the selection is made. Our method also owes much to the `COSIATEC` algorithm, which is also a means of selecting important patterns from the patterns discovered by `SIATEC`. The primary difference between our approach and `COSIATEC` is that our method is not based solely on the principle of optimising compression, but instead allows musicological principles to influence the outcome alongside information theoretic measures. As a result, we are able to select patterns that are deemed to be of musicological interest, but which may not lead to the generation of a complete or optimally compressed representation of the dataset, as is generated by `COSIATEC`. For example, we are able to select multiple patterns that share notes in common, but which have different patterns of occurrence within a piece. The ability to analyse the occurrences of closely related patterns within a work can provide interesting insight into the compositional treatment of thematic ideas.

There is still a great deal of work to be done in order to improve the qual-

ity and reliability of automated music analysis. However, even as it stands, our system opens up some very interesting possibilities for future work. Automated systems cannot currently hope to match the quality of analysis performed by professional musicologists, but do have an advantage of being able to process very large amounts of data. The ability to reliably isolate significant musical patterns and infer basic structural relationships between patterns from across a database of many thousands of pieces of music could form the basis of a rich source of information—both in the sense of musical knowledge, with potential applications in domains such as artificial intelligence and music information retrieval, but also in the sense of data to be analysed by musicologists, potentially extending the scope of traditional musicological enquiry.

# Chapter 4

# Conceptual space point set models of melodic similarity

This chapter employs a representational strategy for musical information based on point sets in low-dimensional spaces, as in chapter 3, but one in which the space is also equipped with a norm, plus a transformational distance measure capable of measuring distance between *sets* of points. The objective of this approach is the modelling of melodic similarity. The Earth Mover's Distance (EMD) metric (Rubner et al. 2000) is employed as the transformational similarity measure between point set representations of melodies (Typke 2007). Evaluation is firstly conducted against a dataset of human similarity rating between a set of existing popular music melodies and a set of carefully constructed melodic variations. This dataset originates from a psychological experiment reported by Müllensiefen and Frieler (2004), and models developed therein provide the basis for comparison. Further evaluation is performed over the dataset used in the MIREX 2005 symbolic melodic similarity evaluation.[4] The objective of the MIREX evaluation was to compare the retrieval accuracy of a number of melodic similarity algorithms given a set of query melodies. For each melodic query, the corresponding set of most similar melodies was established in advance by musical experts (Typke et al. 2005). A comparison between the performance of the models developed in this chapter against the results of this MIREX evaluation is particularly appropriate as the evaluation method and data originate from Typke's work on EMD-based melodic similarity (Typke 2007), on which the present research is based. The EMD-based algorithm developed by Typke was also submitted to the MIREX evaluation, thus affording direct comparison. The relationship between the two experiments reported in the chapter, one psychologically oriented, the other towards music information

[4] http://www.music-ir.org/evaluation/mirex-results/sym-melody/index.html

retrieval, raises some interesting issues concerning the respective approaches of each discipline towards music. In particular, we show that models motivated by psychological concerns, and trained on psychological data, can exhibit comparable performance to models trained directly over exemplar training sets, as is the typical approach in machine learning and information retrieval communities. This is not to argue that the approach adopted by one community is necessarily better than the other, since the motivations of each can be very different, but to highlight the potential mutual benefits of multi-disciplinary perspectives in furthering the scientific understanding of music.

In the context of the narrative of this thesis, the representation and consequent models developed in this chapter can be understood as a hybridisation of the ordered vector space representation studied in chapter 3 and the normed vector space representations explored in chapter 5.

## 4.1 Earth Mover's Distance

### 4.1.1 Background

The Earth Mover's Distance, along with an associated compressed representation of feature distributions called *signatures*, was first described as a distance metric in the context of image retrieval by Rubner et al. (2000). EMD offers a means to compare the difference between multidimensional distributions, based on the minimum cost required to transform one distribution into the other, an idea initially proposed by Peleg et al. (1989).

Images are often summarised by multidimensional distributions over some feature space, or more typically fixed-bin histograms of such distributions. A common issue faced when using histograms is determining the optimal partitioning of the feature space, resulting in a trade-off between expressiveness (too few bins to accurately characterise the distribution) and time and space complexity (too many bins). Signatures are instead proposed as a space efficient, adaptable representation. A signature $\left\{\mathbf{s}_j = \left\langle \mathbf{m}_j, w_{\mathbf{m}_j} \right\rangle\right\}$ represents a set of feature clusters, where $\mathbf{m}_j$ is the mean (or mode) of cluster $j$, and $w_{\mathbf{m}_j}$ the fraction of pixels belonging to that cluster, corresponding to its prominence or weight in characterising the image (Rubner et al. 2000, pp. 104–105). Clustering must first be performed for each image to determine the representative clusters. The key point is that the length of each signature depends on the complexity of the image, and EMD is a metric capable of measuring distance between signatures of varying length.

Rubner et al. (2000) provide a proof that EMD is a true metric when distributions are of equal mass, and present empirical evidence that EMD can offer a better account of perceptual similarity in the context of image retrieval, compared to a number of preexisting histogram-based dissimilarity measures.[5]

In addition to improved retrieval accuracy, EMD naturally affords partial matching. When two distributions differ in mass, EMD is the minimum cost of transforming the smaller distribution into a subset of the larger. In this case EMD is not a true metric, although metric variants have since been developed (Pele and Werman 2008; Pele and Werman 2009).

The original description of EMD focuses on distances between distributions because of the domain specific requirement of image processing. However, the authors note that EMD can be viewed more generally as a method that 'extends the notion of a distance between single elements to that of a distance between sets, or distributions, of elements' (Rubner et al. 2000, p. 105). In the context of discrete representations of music, viewing EMD as a measure of distance between sets of elements, such as a set of points in a two-dimensional space of time and pitch, is most appropriate for a number of reasons. First, as music exists over time it is essential that sequentiality not be thrown away casually, as would be the case if one were to analyse distributions of duration or IOI values. Second, in comparison to images, symbolic music data is typically smaller, and can often be processed directly, without the need for summarising features. However, more importantly, discrete representations of a musical surface are already situated at a sufficiently high level to be considered as representations of objects of cognition, unlike individual pixel information.

### 4.1.2 A weighted point set representation of melody

Before turning to the definition of EMD itself, it is necessary to first define a basic representation appropriate to music over which EMD can be measured. This representation will be developed further in section 4.2, developing Gärdenfors' notion of conceptual space in the context of point set representations of complex stimuli. Based on Typke (2007), weighted point sets are defined as follows.

**Definition 4.1.** Let $A = \{\langle \mathbf{a}_1, w_1 \rangle, \ldots, \langle \mathbf{a}_m, w_m \rangle\}$ be a weighted point set of $m$ point-weight pairs, $i = 1, \ldots, m$, where $\mathbf{a}_i \in \mathbb{R}^k$ with $w_i \in \mathbb{R}^+ \cup \{0\}$. Let $\mathrm{W} = \sum_{i=1}^{m} w_i$ be the total weight of set $A$.

[5] The histogram dissimilarity measures tested were: $L_1$ distance, Jeffrey divergence, $\chi^2$ statistic, and Quadratic-form distance.

Typke represents each note of a melody as a point in a two-dimensional Euclidean space of notated onset-time $\times$ pitch. Rests are not represented explicitly. The onset time of each event is normalised with respect to the encoding timebase, to ensure a common time scale. Pitch is represented in Hewlett's Base-40 notation (Hewlett 1992).[6] Duration is then used as the weight associated with each point. The point of departure for Typke's representation is conventional score notation, which broadly speaking represents pitch and time in two dimensions, and uses different symbols to represent note duration. The rationale for the use of duration as EMD weight is that longer notes are typically more salient in a melody, and therefore deserve greater weight in determining overall melodic similarity.

### 4.1.3 EMD definition

Computing EMD relies on a solution to the well known *transportation problem* (Hitchcock 1941). This problem involves finding the most cost-effective means of transporting a quantity of goods from a given a number of suppliers, to a given number of consumers. Each supplier has a stated amount of goods, while each consumer has a limited amount of capacity. The demand for supply must be met for each consumer, assuming sufficient quantity of goods are available. The cost of transportation between each supplier-consumer pair is given. A solution to the transportation problem describes a particular *flow* of goods between each pair of suppliers and consumers. The amount of *work* involved is characterised by the amount of flow, together with the cost of transportation, between all pairs. The amount of work involved is used as a term in the definition of EMD.

Casting weighted point set representations of melodies into this scenario, one melody point set is arbitrarily designated as the set of suppliers, and the other as the set of consumers. The weight associated with each point becomes the respective supply or demand of goods to be transported, and the cost of transportation between each supplier-consumer pair is the distance between the points in the space. In the context of EMD, distance in this space is called the *ground distance*, referring to the analogy characterising EMD as the amount of work required to move piles of earth into holes in the ground. Rubner et al. (2000, p. 104) stress the importance of defining a perceptually-motivated ground distance. This question is addressed in terms of conceptual space representations in section 4.2. Intuitively, this process can be thought of as the cost of transforming one point set (melody) into another. When two identical point sets are considered, zero cost will be incurred because all pairs of supply and consumer points are co-located in the space,

[6] MIDI pitch number is used in the MIREX 2005 evaluation corpus.

requiring no movement of weight.

Based on Rubner et al. (2000), the transportation problem can be formulated as the following linear programming problem: Let $P = \left\{\left\langle \mathbf{p}_1, w_{\mathbf{p}_1}\right\rangle, \ldots, \left\langle \mathbf{p}_m, w_{\mathbf{p}_m}\right\rangle\right\}$ be the first weighted point set with $m$ points, where $\mathbf{p}_i$ is a point representing an event, and $w_{\mathbf{p}_i}$ is the associated weight; $Q = \left\{\left\langle \mathbf{q}_1, w_{\mathbf{q}_1}\right\rangle, \ldots, \left\langle \mathbf{q}_n, w_{\mathbf{q}_n}\right\rangle\right\}$ the second point set with $n$ points; and $\mathbf{D} = [d_{ij}]$ the ground distance matrix where $d_{ij}$ is the ground distance between points $\mathbf{p}_i$ and $\mathbf{q}_j$.

The optimal solution involves finding a flow $\mathbf{F} = [f_{ij}]$, where $f_{ij}$ is the flow between $\mathbf{p}_i$ and $\mathbf{q}_j$, which minimises the overall cost

$$\text{WORK}\left(P, Q, \mathbf{F}\right) = \sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} f_{ij}, \tag{4.1}$$

subject to the following constraints:

$$f_{ij} \geq 0 \qquad 1 \leq i \leq m, 1 \leq j \leq n \tag{4.2}$$

$$\sum_{j=1}^{n} f_{ij} \leq w_{\mathbf{p}_i} \qquad 1 \leq i \leq m \tag{4.3}$$

$$\sum_{i=1}^{m} f_{ij} \leq w_{\mathbf{q}_j} \qquad 1 \leq j \leq m \tag{4.4}$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} = \min\left(\sum_{i=1}^{m} w_{\mathbf{p}_i}, \sum_{j=1}^{n} w_{\mathbf{q}_j}\right) \tag{4.5}$$

Constraint 4.2 ensures that all weight only flows in one direction, from the suppliers to the consumers. Constraint 4.3 ensures that the flow from any supplier cannot exceed its weight, and similarly constraint 4.4 ensures that no consumer can receive more than its weight. Constraint 4.5 ensures that the maximum weight possible is moved, which is equal to the total weight of all suppliers, or all consumers, whichever is the smaller. This amount is called the *total flow*. Solving the transportation problem results in the optimal flow $\mathbf{F}$. EMD is then defined as the resulting work normalised by the total flow:

$$\text{EMD}\left(P, Q\right) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} f_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}}. \tag{4.6}$$

Normalising by total flow ensures that the resulting work remains proportional to the weights of the point sets under consideration, thus avoiding favouring smaller weighted point sets.

A number of approaches exist for efficiently solving EMD. Rubner et al. (2000)

implement the streamlined transportation simplex (Hillier and Lieberman 1990), taking advantage of the particular structure of the underlying transportation problem. We apply a standard simplex solver[7], which provides more than adequate performance for our purpose.

## 4.2 EMD model definitions

Building on the general weighted point set representation defined in section 4.1.2, we now specify the models to be tested. Each model consists of three variable components: a space, a norm, and a weighting scheme. A space is a set of dimensions within which the events of a melody are represented as points. A norm defines the length of vectors in the space, which determines the distance between points in the space—the ground distance. Finally, a weighting scheme defines how weight is assigned to each point, which here is used primarily to determine whether EMD measures are based on partial or complete matching between point sets.

### 4.2.1 Ground distance dimensions

#### Basic attributes

We begin with a symbolic definition of the musical surface from which we subsequently derive a number of conceptual space representations. Let a melody $\mathbf{m}_i^j = \langle e_1, e_2, \ldots, e_j \rangle$ be an ordered sequence of events of length $j \in \mathbb{Z}^+$, indexed by $i \leq j$. Each event is a 4-tuple $\langle \texttt{onset}, \texttt{cpitch}, \texttt{cpitch}_c, \texttt{dur} \rangle$ consisting of four basic attributes: onset time, chromatic pitch, centred chromatic pitch, and duration. A fixed-width font is used to distinguish the symbolic definition of attributes from their derived conceptual-level geometric definition below.

Onset is defined as the score-time representation of an event onset relative to the notional beginning of the melody, defined as timepoint zero, and measured in crotchets, where a crotchet equals one time unit. Unlike Typke (2007) we do not insist that the onset time of the first event of all melodies must be equal to zero. For melodies that do not begin on the first beat of the bar, but are notated as beginning equal to or before the mid-point of the bar, leading rests are respected. Therefore, in this situation, the onset of the first event will be greater than zero, but less than or equal to the number of beats in the bar divided by two. For melodies that are notated as beginning in the second half of the bar, we assume this represents

[7] `glp_simplex` from the GNU Linear Programming Kit (GLPK), using a custom foreign function interface bindings from Common Lisp to the GLPK Graph API.

an anacrusis figure, and shift the melody earlier in time by one bar, to the effect that anacrusis events have negative onset values, and the first downbeat of the melody is equal to zero. Where melodies contain multiple bars rest before the first event, as is common in the RISM dataset, all leading empty bars are removed. It is also common in the RISM dataset for leading rests not to be explicitly notated. Therefore, when the number of beats of the first notated bar is detected as being less than the expected number of beats given the notated time-signature, the metrical position of events in the opening short bar is computed relative to the downbeat of the next bar, ensuring the correct metrical interpretation is derived. Duration is defined as the interval between an event onset and offset as given in the score, also measured in units of crotchets.

Chromatic pitch is simply represented by MIDI note number, providing a representation of absolute pitch height. Centred chromatic pitch is the same as chromatic pitch, except transposed so that the duration-weighted mean pitch height of each melody is equal to middle-C, represented by MIDI note number 60. Duration-weighted mean pitch height is defined in equation (4.7), where $\mathbf{m}_i^j$ is the melodic sequence under consideration, and the functions $cpitch(.)$ and $dur(.)$ here return the chromatic pitch and duration values respectively associated with event $\mathbf{m}_i$. The motivation for considering centred pitch height is to minimise the impact of different tessitura, and to some extent key, when comparing pitch height. Only one absolute pitch representation is projected into a conceptual space in any model. More musically sophisticated representations of pitch, such as pitch-class or scale-degree, could also be used as the basis for alternative transposition-invariant pitch height representations, and will be evaluated in future work. However, being an absolute representation of pitch height, centred chromatic pitch has the additional benefit of preserving melodic contour, without requiring additional dimensions.

$$DWMPH(\mathbf{m}_i^j) = \frac{\sum_{i=1}^{j} cpitch(\mathbf{m}_i) \cdot dur(\mathbf{m}_i)}{\sum_{i=1}^{j} dur(\mathbf{m}_i)} \tag{4.7}$$

It is perfectly possible to treat the above event tuples as points in a vector space, analogous to the point set representation used by `SIA` in chapter 3. However, as our objective here is the construction of a conceptual space in which distance between points corresponds to a notion of psychological similarity, treating the basic attributes directly as, for example, values in a three-dimensional space of `onset` $\times$ `cpitch` $\times$ `dur` is unlikely to yield meaningful distances due to the different ranges and scales of each attribute. Furthermore, the perception of event attributes need not necessarily manifest in the same linear form as their sym-

bolic representation. For example, several authors have investigated the cognitive structures associated with tonal pitch perception, and have proposed various multidimensional spiral-like representations (Jones 1981; Shepard 1982; Chew 2004; Krumhansl 2005). More complex cognitive representations such as these, and those developed in chapter 5 concerning metre, are not considered here. However, the perceptually-motivated scaling of attribute values is investigated.

In order to project basic event attributes into a *conceptual* space, as opposed to merely a geometrical space, it is necessary to scale the values of the attributes appropriately so that all values are represented in the same relative unit of measurement. Standardisation of variables is intended to serve two primary purposes. First, it affords a degree of compatibility between dimensions. Intuitively, it allows variance in one attribute to be meaningfully compared and combined with variance in another. Second, we predict that using standardised dimensions will afford a degree of model generality, meaning that models trained on one dataset should be applicable to another broadly similar dataset, even if the underlying attribute representation or overall corpus statistics differ. Standardising dimensions is closely related to dimension salience weights introduced below. In fact, both are identical processes applying a linear scaling to a dimension. However, each serves a distinct purpose, and it is useful to consider them independently, as will be discussed below.

Following the standardisation strategy recommended by Raubal (2004), we simply map attributes onto quality dimensions using a $z$-transformation (4.8).

$$z_i = \frac{x_i - \bar{X}}{s_X} \tag{4.8}$$

Where $x_i$ is the attribute value to be transformed, $z_i$ is the standardised quality dimension value, and $\bar{X}$ and $s_X$ are respectively the mean and standard deviation of attribute $X$. In the case of `onset`, the mean and standard deviation of melody length over the corpus is used for standardisation. Each of the other basic attributes are simply treated as distributions directly, and standardised according to the mean and standard deviation computed over the corpus. This method assumes that attribute values follow a normal distribution, which is a reasonable assumption in our case. A alternative method would be to use distributional distance (Müllensiefen 2009), whereby normalised attribute distance is defined in terms of an empirically derived cumulative distribution function. The method would be particularly useful for non-normal distributions, but it will not be considered further here.

We can now define four quality dimensions, corresponding to the above four $z$-transformed basic attributes thus:

- $\textsc{onset} = \mathbb{R}$
- $\textsc{cpitch} = \mathbb{R}$
- $\textsc{cpitch}_c = \mathbb{R}$
- $\textsc{dur} = \mathbb{R}$

The ground distance space of a model is defined as the Cartesian product of individual dimensions, allowing different models to be specified combining different dimensions. Two further dimensions representing relative attributes are defined below, and these can be combined analogously. The space $\textsc{onset} \times \textsc{cpitch}_c$ is taken as our baseline ground distance space, as onset and centred chromatic pitch are the attributes investigated by Typke (2007).

Every quality dimension is also equipped with a salience weight. Salience weights are not shown in the basic representation for clarity, but when it is necessary to refer to them directly they will be notated as a preceding multiplying term, for example, $0.5{\cdot}\textsc{onset}{\times}1.0{\cdot}\textsc{cpitch}$. As mentioned above, salience weights apply a linear scaling to a dimension, in much the same way as dimension standardisation. However, standardisation happens at the beginning of a modelling process, when surface-level attributes are mapped to quality dimensions, and salience weights are adjusted subsequently during model fitting. If standardisation were omitted, the fitting process should be able to still find the optimum scaling parameters. However, this collapses issues of dimensional scale and genuine perceptual salience into a single factor, sacrificing both interpretation and potential model generality. For example, one approach used by Typke addresses this issue from the perspective of attribute scaling, whereby time values are multiplied by three in order to bias the solution of the transportation problem by not making it 'too cheap to move weight in the time dimension in comparison to the pitch dimension' (Typke 2007, p. 33). In this case, the value three does not lend itself to further interpretation: it is meaningless to consider it in relation to the relative importance of the time and pitch dimensions in the model of similarity.[8] This issue becomes more serious when further dimensions are added to the space, and standardisation offers some assurance that a high salience weight as a result of model fitting is a genuine indication of perceptual importance.[9] Furthermore, when dimensional scale

[8] Another more sophisticated approach based on the optimal alignment of segmented queries is use in Typke's EMD-based retrieval algorithm submitted to the MIREX 2005 symbolic melodic similarity evaluation. This algorithm is discussed further in section 4.4.1.

[9] A post-hoc investigation into the information theoretic characteristics of individual dimensions could prove useful here. If any such link was found to exist, it could lead to the ability to *predict* salience weights, rather than relying on search.

and salience are conflated, it becomes less likely that a model that works on one dataset will exhibit comparable performance over different data, especially in cases where different attribute representations are used. For example, representing onset values in milliseconds will require a very different scaling factor than onsets represented in units of crotchets. However, using standardised values removes this complication. Furthermore, it allows one to probe the question of whether empirically discovered dimensional saliences generalise across corpora.

#### Relative attributes

All melodies are assumed to be Charm stream constituents, meaning that no events overlap in time, and events are ordered by time (equation (2.2); Harris et al. 1991). This allows us to construct two further *relative attributes*: inter-onset interval (`ioi`) and pitch interval (`cpint`).[10] Relative attributes represent relations between basic attributes. However, in order to simplify the representation, they are treated as properties of events in the same way as basic attributes. This process is formalised in equations (4.9) and (4.10).

$$\texttt{ioi}_i = \begin{cases} \texttt{onset}_i - \texttt{onset}_{i-1} & \text{if } 1 < i \leq j \\ \top & \text{otherwise.} \end{cases} \tag{4.9}$$

$$\texttt{cpint}_i = \begin{cases} \texttt{cpitch}_i - \texttt{cpitch}_{i-1} & \text{if } 1 < i \leq j \\ \top & \text{otherwise.} \end{cases} \tag{4.10}$$

The $\top$ symbol denotes an undefined value. Therefore, the above definitions specify that relative attributes are undefined for the first event in a sequence. Moving from symbolic attributes to a vector space, the presence of undefined values presents a serious problem. A mathematically correct approach might be to keep basic and relative attributes separate, in distinct spaces. However, this raises a question regarding EMD, which assumes all points are represented in a single space. One approach might be to compute EMD separately in basic and relative spaces, and somehow combining the resulting distances into a single measure. However, this is beyond the scope of the present work.

In order to investigate the inclusion of relative attributes into EMD models, we adopt the following workaround, acknowledging its semantic inconsistency and potential source of additional noise. When mapping undefined relative attributes

[10] The notion of relative attributes here is related to derived viewpoints in a multiple viewpoint system (Conklin and Witten 1995) developed in the context of statistical modelling of symbol sequences.

in the first event of each melody, the undefined value is replaced with the mean value of that attribute in the melody under consideration. In this way, the ground distance from the first event to all subsequent events in this dimension is minimised and evenly distributed. This gives us two further dimensions to investigate when computing ground distance:

- CPINT $= \mathbb{R}$
- IOI $= \mathbb{R}$.

Following Typke (2007), all spaces investigated here contain an onset dimension, and either a centred or non-centred chromatic pitch dimension, providing a basic representation of melodies as an ordered sequences of events in time, where each event contains a definite pitch height. All other defined dimensions are combined in turn with the basic representation to test their impact independently on the task of predicting melodic similarity. Two further spaces are defined, containing all additional dimensions along with a centred or non-centred chromatic pitch dimension, in order to test the combined impact. Therefore, the complete list of spaces to be tested is as follows.

- ONSET $\times$ CPITCH$_c$
- ONSET $\times$ CPITCH
- ONSET $\times$ CPITCH$_c$ $\times$ DUR
- ONSET $\times$ CPITCH $\times$ DUR
- ONSET $\times$ CPITCH$_c$ $\times$ CPINT
- ONSET $\times$ CPITCH $\times$ CPINT
- ONSET $\times$ CPITCH$_c$ $\times$ IOI
- ONSET $\times$ CPITCH $\times$ IOI
- ONSET $\times$ CPITCH$_c$ $\times$ DUR $\times$ CPINT $\times$ IOI
- ONSET $\times$ CPITCH $\times$ DUR $\times$ CPINT $\times$ IOI

A consequence of including ONSET within each ground distance space is that the representation does not provide shift-invariance in time. As described in detail in section 4.4.1, Typke (2007) addresses this issue algorithmically by segmenting melodies, and then translating and scaling each segment in time to find an optimal alignment. Therefore, in cases where melodies under comparison contain identical segments, but which appear in a different order, they may still be matched. An alternative approach to be pursed in future work will be to extend the representation to afford greater degrees of shift invariance by including metrical information. In the most simple case, the inclusion of a quality dimension representing the position of events in the bar, either instead of or alongside ONSET, would allow for more flexible matching that is less dominated by strict sequential ordering.

### 4.2.2 Norms

We now have a number of different conceptual vector spaces within which to represent melodies. Recall that the ground distance matrix $\mathbf{D}[d_{ij}]$ required in the computation of EMD represents the distances from all the points representing the notes of one melody, to all the points representing another. In order to measure these distances, a norm must be specified defining the length of the vectors between points.

We use the common $L^1$ norm, corresponding to city-block distance, and the $L^2$ norm, corresponding to Euclidean distance. Each space defined above will be tested with both $L^1$ and $L^2$ norms for comparison. We predict that $L^1$ norm will provide the most accurate measure of perceptual distance based on Gärdenfors' rule of thumb that city-block distance is more appropriate for modelling separable dimensions, and that Euclidean distance is more appropriate for integral dimensions (see section 2.3.1). While in one sense pitch and time dimensions are integral because a sequence of perceived pitches exists in time, our intuition is that pitch and time dimensions will be better modelled as an additive combination because the effect of change in either can be readily apparent as a distinct qualitative difference in perception.

### 4.2.3 Weighting schemes

The weights associated with each point are integral to the computation of EMD, since it is a measure of the amount of work required to move weight from the points in one point set, to the points in another. One way to conceptualise weight is as an extra dimension, external to the ground distance space, but exerting an influence on the overall measure of distance between point sets from 'outside' the vector space formalism. In applications of EMD, weight is typically described as reflecting the relative importance of individual points in a point set. Typke (2007) follows this logic by using note duration as weight, under the reasonable assumption that longer notes are more important in perception. In conceptual space terms, in adopting this view weight becomes analogous to quality dimensions, because it is a representation of a perceptually important variable quality.

A potential problem with endowing weight with perceptual significance is determining the appropriate level of influence weights have in the computation of distance relative to the ground distance dimensions. This is an issue closely related to the previous discussion of dimension standardisation and salience weights. For quality dimensions, we have proposed a uniform mechanism for addressing the

scaling and relative importance of dimensions, which is possible because quality dimensions are defined within a general vector space formalism. However, weights exist outside this framework, and their influence on distance can not be interpreted in terms of simple geometrical transformation. This is not to argue that EMD weights should not represent perceptually relevant qualities at all, only that to do so requires careful consideration of how ground distance and weighs interact, a thorough investigation of which is out of the scope of the present study.

Our preferred method of assigning weight to points is to consider it in the abstract, and not as a representation of a variable perceptual quality at all. In doing so, we define two weighting schemes, each serving a well-defined purpose entirely independent of variable melodic qualities.

Weighting scheme $P$ assigns all points an equal weight. The value of the weight is defined relative to the largest point set in the corpus (4.11).

$$w_{a_i} = \frac{1}{\max_{X \in Y} |X|}, \quad \forall a_i \in A, \forall A \in Y \tag{4.11}$$

$w_{a_i}$ is the weight according to weighting scheme $P$ for point $a_i$ in set $A$. $X$ is a point set from a corpus $Y$, and $|X|$ is the cardinality of the point set $X$. The total weight of the longest melody will therefore sum to one. All shorter melodies will have a total weight of less than one.

Applying weighting scheme $P$ in a comparison between melodies comprised of different numbers of events will lead to all events from the shorted melody being matched to a subset of events from the longer melody. Within this scheme, non-matched events from the longer melody do not contribute to the measure of similarity. We refer to this scheme as the *partial matching weighting scheme*, and the resulting measure of EMD employing this weighting scheme as *partial matching EMD*, because for melodies of unequal length, longer melodies are only partially matched in order to calculate EMD. All events of shorter melodies do contribute to the resulting measure of similarity, and in this sense are completely matched. However, the terminology used here reflects the fact that for unequal length melodies, EMD is calculated based on partial matching with respect to the longer melody. Where melodies under comparison are comprised of an equal number of events, this scheme does lead to complete matching between both melodies. However, we reserve the label *complete matching weighting scheme* for the following weighting scheme, which explicitly enforces complete matching between both melodies irrespective of length.

Weighting scheme $C$ assigns all points in a set an equal weight proportional to

the cardinality of the set, as defined by equation (4.12). The total weight for each point set will sum to one, giving an EMD measure based on complete matching because all the weight of the source point set is able to flow to the sink point set. An EMD measure incorporating this weighting scheme will be referred to in the text as *complete matching EMD*. Although all points within a point set will have equal weight, the specific weight value will vary between point sets depending on their size, with the points representing the events of longer melodies receiving less weight than those representing the events of shorter melodies.

$$w_{a_i} = \frac{1}{|A|}, \quad \forall a_i \in A \tag{4.12}$$

In partial matching EMD, all points have a weight in units of a common standard, which in the simplest case all events have the same unit weight, as adopted here. This leads to a measure of distance based on the differences between the best matching events across a pair of melodies. In psychological terms, this method embeds an assumption that common or similar elements between stimuli are more important in determining psychological distance than very unrelated events. Furthermore, the solution to the underlying partial matching transportation problem can be given a more natural higher-level interpretation compared with complete matching. In partial matching, weight is more likely to be transported on an event-by-event basis due to the common unit of weight. Weight is in no way constrained to flow exclusively between disjoint pairs of events, and is free to flow across multiple events in order to achieve the optimum solution. However, this is always the case when applying the complete matching weighting scheme to melodies of unequal length, because the weights associated with the points of each point set are determined by the size of each set respectively, which must *sum* to a specified common amount. This leads to a flow that is more abstractly related to the identifiable concepts captured in the representation.

The partial matching weighing scheme offers a means of ignoring unmatched portions of melodies. For otherwise highly similar melodies only differing in length by a small number of events, this offers a potential advantage over complete matching. The addition of a small number of melodically coherent events is likely not to greatly affect human judgements of similarity, and in the partial matching scenario, the addition of such events would not affect the predicted similarity. However, the presence of even a single additional event when applying the complete matching weighting scheme can substantially alter the flow of weight between point sets, and disproportionately affect the prediction of similarly. How-

ever, as the difference in melody length grows, additional events are likely to exert an increasing influence on perceived similarity, which cannot be accounted for by the partial matching EMD measure. In this situation, the more abstract notion of EMD based on the complete transformation of one point set into another may offer a better account of perceptual similarity. A variant of EMD developed by Pele and Werman (2009) does take into account unmatched weight within a partial-matching weighting scheme. This algorithm may offer a compromise between the two classical partial and complete matching EMD approaches, and should be considered in future work.

Returning to the present study, and in light of the above discussion, we would expect partial matching EMD to perform well given the controlled nature of the Müllensiefen and Frieler (2004) dataset. This dataset consists of original melodies and variants, all of roughly equal length, as can be seen in figure 4.1. However, the variation in melody length is substantially greater in the MIREX 2005 symbolic melodic similarity dataset, as can be seen in figure 4.2. Therefore, complete matching EMD may be better suited to this dataset, due to the increased proportion of potentially salient melodic content that is ignored in the calculation of partial matching EMD.

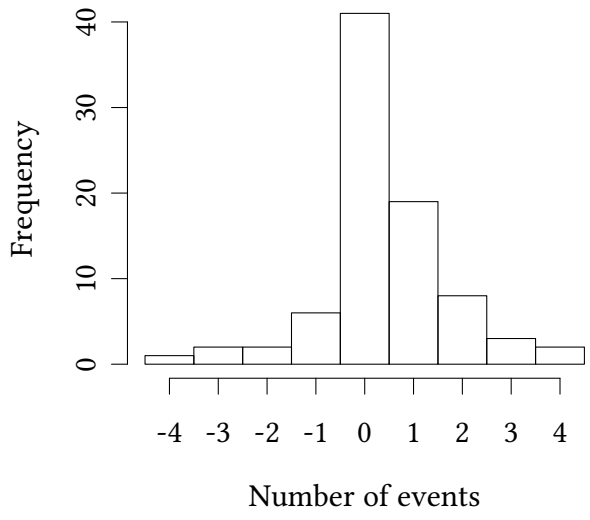



Figure 4.1: Histogram of the difference in length between original and variant melodies in the Müllensiefen and Frieler (2004) dataset.

For means of comparison, two further weighting schemes are defined allowing both partial and complete matching but where weight is also allocated proportion-

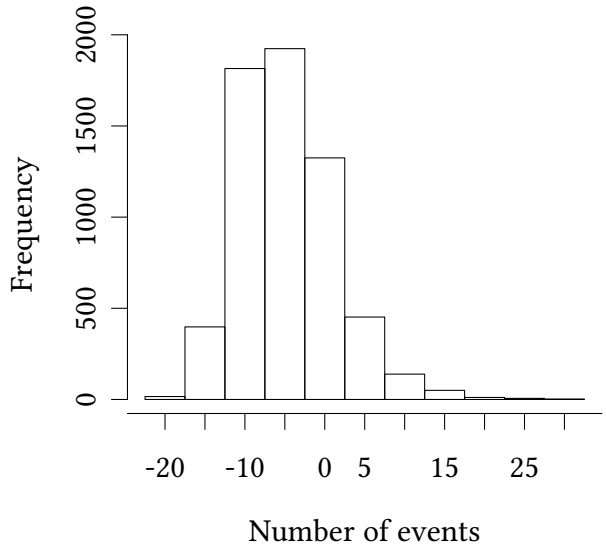



Figure 4.2: Histogram of the difference in length between query and candidate melodies in the MIREX 2005 symbolic melodic similarity dataset.

ally to note duration. In weighting scheme $P_d$ (4.13), all events of equal duration will have equal weight, and the total weight for each set will vary. The total weight for the melody containing the greatest sum of durations will be equal to one. The function *dur*(.) here returns the duration value associated with the event represented by each point.

$$w_{a_i} = \frac{\textit{dur}(a_i)}{\max_{X \in Y} \sum_{j=1}^{m} \textit{dur}(x_j)}, \quad \forall a_i \in A, \forall A \in Y \tag{4.13}$$

In weighting scheme $C_d$ (4.14), all events of equal duration within a melody will have equal weight, but the specific amount will vary across melodies. The total weight for each point set will sum to one, ensuring complete matching.

$$w_{a_i} = \frac{\textit{dur}(a_i)}{\sum_{j=1}^{m} \textit{dur}(a_j)}, \quad \forall a_i \in A \tag{4.14}$$

Defining these additional duration-based weighting schemes will afford direct comparison between models incorporating duration information as quality dimensions, and alternatively as EMD weight. We expect duration information to be more useful as a quality dimension because this method allows for the optimum scaled combination of dimensions to be discovered.

### 4.2.4 Model specification

An EMD model is constructed from three components: a space, a norm and a weighting scheme. Models are notated as a tuple within parentheses, for example:

$$(\textsc{onset} \times \textsc{cpitch}_c, L^1, P)$$

Combining the ten space definitions from section 4.2.1, with the two norms defined in section 4.2.2, and the two preferred weighting schemes from section 4.2.3, produces 40 separate models. A further eight models are constructed combining the two spaces $\textsc{onset} \times \textsc{cpitch}_c$ and $\textsc{onset} \times \textsc{cpitch}$ with the two norms, and the two duration-based weighting schemes from section 4.2.3.

Each model also has a set of parameters that will be optimised during model fitting. The parameters are the salience weights associated with each quality dimension in the space, and are assumed implicit in the above notation, as discussed in section 4.2.1. One further global scaling parameter $\epsilon$, is necessary for each model due to the use of regression methods in section 4.3 to compare model predictions with human similarity ratings. The parameter $\epsilon$ is an exponent used to apply an exponential transformation to EMD distance measures so that linearity between predictions and human similarity ratings is maximised. The value of $\epsilon$ is determined empirically during model fitting, and across all models was found to take on a value of between 0.28 and 0.69 (see appendix C). The value of $\epsilon$ is irrelevant when dealing with only the rank order of predicted similarity, as is the case for the MIREX experiment in section 4.4. Where it is necessary to explicitly refer to $\epsilon$ using the model notation, it will be included as an exponent appended to the model definition, as in the following example.

$$(\textsc{onset} \times \textsc{cpitch}_c, L^1, P)^{0.5}$$

## 4.3 Experiment 1: Pop melody similarity

For the purposes of model fitting and evaluation, experiment 1 uses a corpus of carefully selected pop melodies and associated expert human similarity ratings gathered as part of a psychology experiment into melodic similarity conducted by Müllensiefen and Frieler (2004). The corpus contains fourteen Western pop music melodies, listed in appendix B, all of between seven and ten bars long (15–20 s). For each original melody, six variations were constructed, giving a total of 84 comparison melodies.

Similarity ratings were collected for all pairs of original and variation melodies. In order to make the task more realistic, participants, all musicology students, were asked to imagine a familiar aural training scenario. They were told to imagine that the first melody presented from each comparison pair (an original melody) corresponded to a reference melody played by a teacher on a piano, and that the second melody (a variation melody) corresponded to a student's attempt to reproduce the first from memory, which may contain between none and many errors. Participants were told then to rate, on a scale of 1 to 7, how accurately each variation melody corresponded to the original, with 7 meaning that they judged the variation melody to be identical to the original, and 1 that it contained many errors.

Variation melodies were constructed according to known limitations of melodic memory (see references in the original article). The "errors" introduced, with respect to the hypothetical aural training scenario, are categorised as follows: rhythmic errors; pitch errors not changing pitch contour; pitch errors changing the contour; errors in phrase order; modulation errors (pitch errors that result in a transition into a new tonality). An example original melody, along with a set of variations are presented in figure 4.3. Variation D2 (omitted in the figure) is identical to the original melody, and variations D1 and D3 contain subtle changes of pitch, retaining the basic structure of the original melody. Variations D5 keeps the same pitch structure as the original melody, but breaks up the even flows of quavers by introducing semiquavers, changing the rhythmic character of the melody. The phrase structure of D4 and D6 are altered from the original in the same manner, with D4 introducing further variation through modulation of key. As can be seen, all variations are coherent and stylistically plausible melodies in their own right.

Similarity ratings were collected from 82 musicology students. Since the aim of the experiment was to investigate an *expert* notion of similarity, the subjects were required to demonstrate consistency over multiple trials, and the ability to recognise identical melodies with high accuracy. Out of the 82 subjects that participated, only 23 proved sufficiently expert at the task. The inter-subject reliability of this subset of subjects was analysed, and found to be high (Cronbach's $\alpha$ = 0.962), giving some assurance that the notion of melodic similarity is at least a stable concept for musical experts.

Müllensiefen and Frieler (2004) test a total of 39 models of melodic structure, covering a range of algorithms, assessing their ability to predict the human similarity ratings. Linear regression models between predicted and actual similarity ratings are used to evaluate predictive ability. EMD-models are not considered in

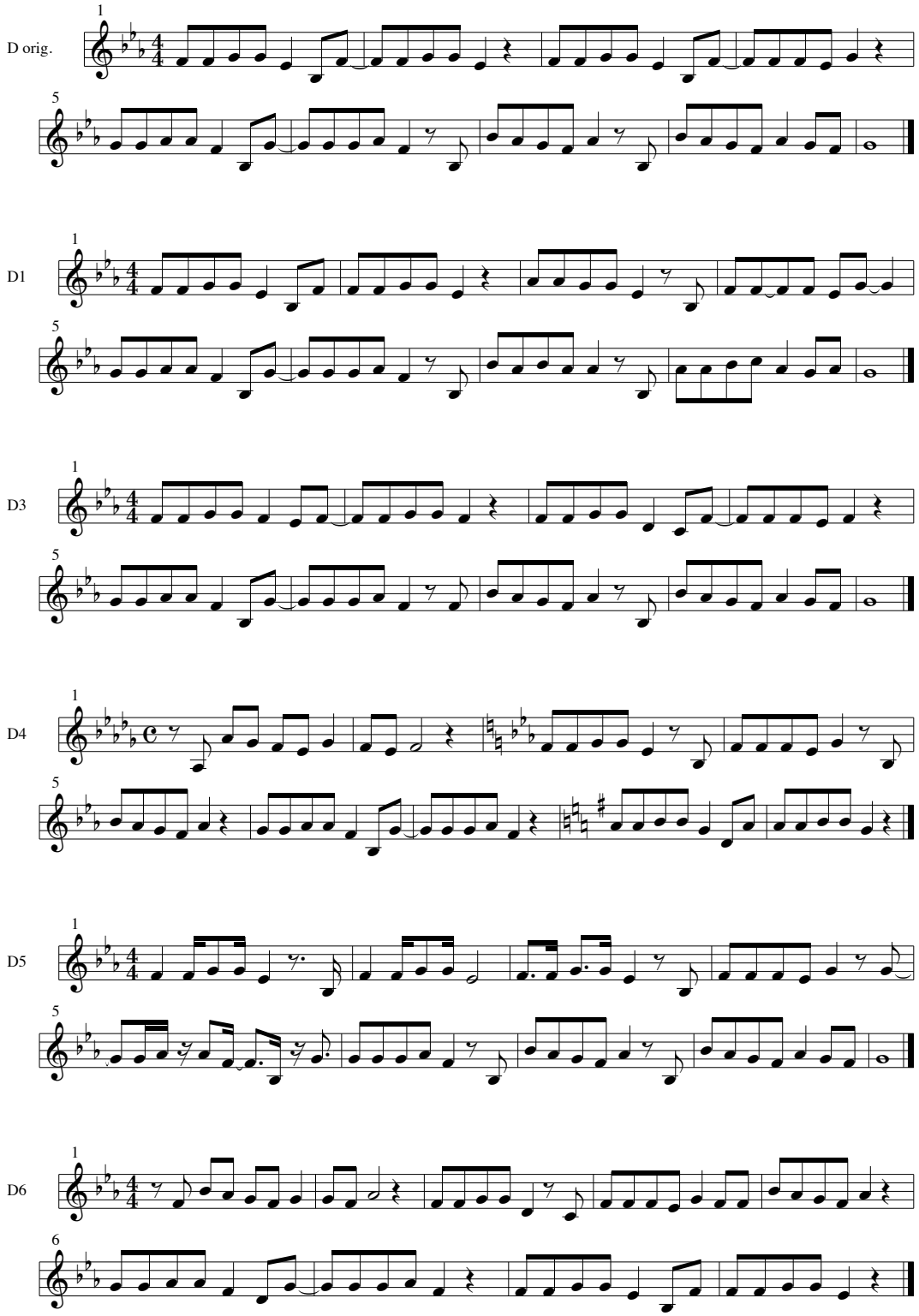



Figure 4.3: Example stimulus melodies from Müllensiefen and Frieler (2004). The original melody (D orig.) in stimulus set D is taken from the song Passion Fruit by Wonderland. Melodies D1–D6 are variation melodies. D2 is omitted here as it is identical to the original.

the original study, but we adopt the same evaluation method so that our results may be directly compared with the models used.

### 4.3.1 Model fitting

To reiterate, the free parameters in each model are the salience weights associated with each quality dimension defined in the ground distance space, plus a global exponent scaling parameter $\epsilon$. At initialisation, all model parameters default to 1.0. We expect that different quality dimensions will be important to different degrees in modelling human similarity judgements. From preliminary testing, it was discovered that model predictions of similarity and human ratings tended to be exponentially related. Therefore, we also determine the optimum value of $\epsilon$ empirically, along with the values of the salience weights of each model, within a random sub-sampling cross-validation scheme to avoid overfitting.

The stochastic search technique Simulated Annealing (SA) (Kirkpatrick et al. 1983) is used to discover the approximate optimum parameter values. SA tries to find a point in the parameter space that minimises a cost function. Each iteration, the algorithm generates a new set of parameter values following a random walk. If the parameters lead to a lower cost value according to the cost function, then the algorithm accepts the new point. If not, the algorithm may still accept the new point based on the probability of a Boltzmann distribution. Over time this probability, or "temperature", gradually decreases. The end solution is an approximation, and is not guaranteed to be equal to the true global minimum, but the strategy is known to perform well in avoiding local minima.

In our case, the parameter space consists of all but one of the salience weights applicable to each model, plus $\epsilon$. It is necessary to hold one salience weight constant so that the search space does not include all trivial transformations of the space. Or to put it another way, what we are interested in is the *relative* values of salience weights, for which the held out parameter serves as the point of reference.[11] In all cases the chromatic pitch salience weight is held constant at 1.0, and all other parameters are free to take arbitrary positive real numbers. The cost function is the standard error of ordinary least squares linear regression between predicted and human similarity ratings. From examining the behaviour of the SA algorithm on our data, 1000 iterations was shown to be more than sufficient for the SA process to explore enough of the parameter space for us to be confident of the validity of the optimisation. Therefore, the SA algorithm was configured to perform 1000 iterations in the generation of all results reported below.

[11] This strategy effectively fixes the scale of the ground distance relative to the EMD weights.

Applying SA over the entire dataset would likely lead to overfitting, and since one of our objectives is to subsequently test the generality of our models over different data, this must be avoided. A 5x2 cross validation (5x2cv) scheme is adopted (Dietterich 1998), which consists of five iterations of 2-fold random sub-sampling cross validation. Each cross validation run randomly partitions the data into two equal halves, or folds. A model is first fitted to one fold, and tested on the other. The same random number generator seed is used for each model, ensuring that the same random subsets are used in all cases.

The 5x2cv technique is useful here for key two reasons. First, given our evaluation method, and consequently the optimisation cost function, both of which rely on the linear relation between model predictions and human ratings, it is necessary that the size of our testing and training subsets be as large as possible for each run. Second, the 2x5cv method has been shown in general to lead to representative samples of model performance. For evaluating model performance we require a method that enables the statistical significance of the *difference* between models to be quantified (not just the significance of how well individual models fit the human ratings). Therefore, instead of only reporting individual model performance, quantified in terms of standard error and $R^2$, as an average across CV folds as is common in the machine learning literature, we perform one further testing run over the entire dataset using the mean parameter values discovered during 5x2cv. This results, for each model, in a final set of predictions that can be compared pairwise using a paired-*r* test, which provides a $t$ statistic for the difference between paired correlations (Revelle 2011).[12]

### 4.3.2 Results

After fitting all models following the above 5x2cv sampling method, we are able to obtain a prediction of the similarity between each pair of melodies according to each model. The performance of each model can be quantified by considering both the standard error between model predictions and human similarity ratings—the term that was minimised during training—as well as the amount of variance in the ratings data that is explained by a model ($R^2$). To quantify the significance of the *difference* between pairs of models, the $t$ statistic for the difference between the dependent correlations $r$ is calculated (Revelle 2011).

[12]Dietterich (1998, pp.10–13) also defines a modified $t$ statistic suitable for hypothesis testing based on the error terms derived from multiple models evaluated using 5x2cv over the same data. This is beyond the scope of planned analysis, but should be considered in future work.

**Preliminary analysis**

Before analysing model performance in detail it is first necessary to consider the data in light of the assumptions of the modelling technique employed. The assumptions of linear regression are linearity, homoscedasticity, normality and independence. Standard diagnostic plots were examined for each model, and standard diagnostic tests performed.

The relatively large number of factors involved, and consequently the high number of models produced, prohibit the discussion of the diagnostic characteristics of each model in turn. Furthermore, the characteristics across all models are broadly similar. Therefore, two models are taken as typical exemplars, and their characteristics can be considered broadly representative of all the EMD-based models here developed. Where any model was found to exhibit significantly different diagnostic characteristics, it will be noted in the text. An arbitrary choice of exemplar models would suffice for the purposes of this preliminary analysis. However, models $(\textsc{onset} \times \textsc{cpitch}, L^1, P)$ and $(\textsc{onset} \times \textsc{cpitch} \times \textsc{cpint}, L^1, P)$ are chosen as these models play a distinguished role in the subsequent analysis.

Non-linearity between human ratings and model predictions was minimised by the inclusion of the exponent parameter in each model, which was optimised during fitting. Scatter plots of predicted versus actual similarity ratings for all models were examined, each demonstrating a reasonable linear correlation between the two variables. Scatter plots for the two exemplar models, including the regression line, are shown in figures 4.4 and 4.5.

It is worthwhile to note the spread of points located at 1 on the predicted similarity axes in figures 4.4 and 4.5, corresponding to model predictions of similarity judgements between identical stimuli. Variance here is typically less than for other ranges along the scale, but the fact there is variance at all makes apparent a false assumption of the EMD models. The psychological evidence shows that subjects were not able to reliably perceive identity between melodies under the experimental conditions. However, all the EMD models naturally predict maximal similarity between identical melodies. Therefore, there is clearly an import factor, or factors, in melodic similarity unaccounted for in the current models. We conjecture that factors influencing the encoding of melodic stimuli in short-term memory may be related here, for example, simply the length of a melody, or more high-level features related to complexity or familiarity. Future work focusing on better understanding the factors influencing the perception of identity in melodic perception could well lead to models better able to predict similarity across the board.

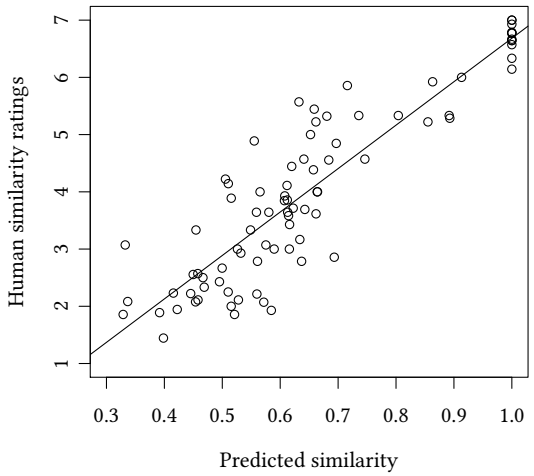



Figure 4.4: Scatter plot of (ONSET × CPITCH, $L^1$, $P$) versus human ratings.

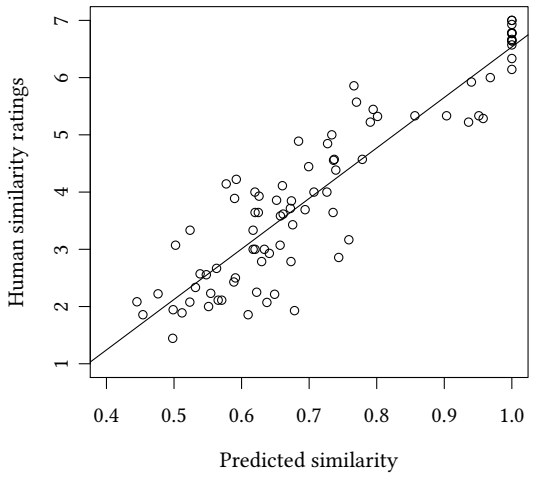



Figure 4.5: Scatter plot of (ONSET × CPITCH × CPINT, $L^1$, $P$) versus human ratings.

Plots of residual versus fitted values for all models reveal that residuals are broadly symmetrically distributed about the horizontal line, providing further support for linearity between predicted and actual similarity ratings. The left plots in figures 4.6 and 4.7 are examples of this. It can also be seen that the range of residuals tends to decrease as a function of fitted values, suggesting a violation of homoscedasticity. However, this was found not to be significant for any model according to the heteroscedasticity directional test statistic ($0.21 < p < 0.98$) (Peña and Slate 2006). In the context of similarity judgements, this trend could perhaps be at least partly attributed to the intuition that similarity between more similar

objects is easier to predict than that between more dissimilar objects.

The quantile-quantile plots in figures 4.6 and 4.7 are normal probability plots showing the standardized empirical distribution of residuals against a normal distribution of the same mean and variance. Ideally, points should lie on the diagonal line, signifying a normal distribution. The distribution of residuals in figure 4.6 shows a degree of kurtosis. However, kurtosis was found not to be significant for any model according to the kurtosis directional test statistic ($0.19 < p < 1.0$) (Peña and Slate 2006).

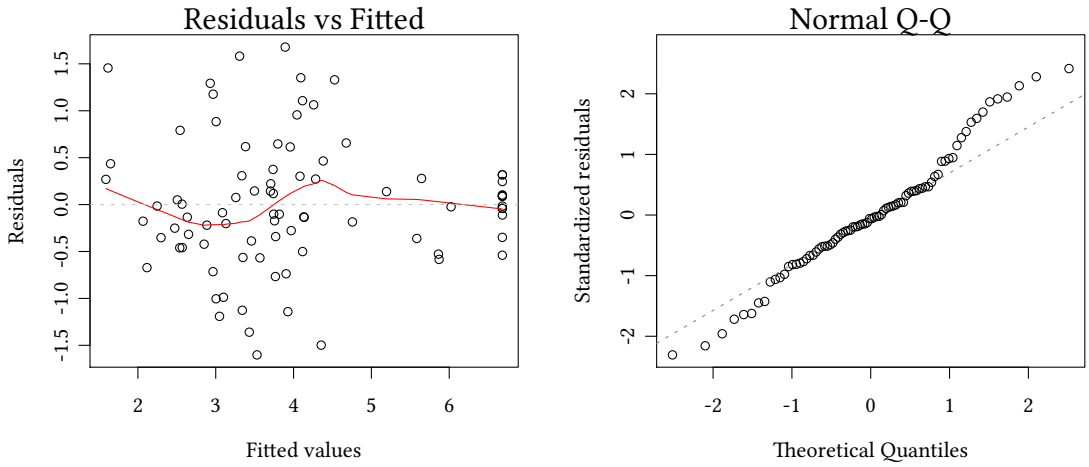



Figure 4.6: Diagnostic plots for $(\textsc{onset} \times \textsc{cpitch}, L^1, P)$.

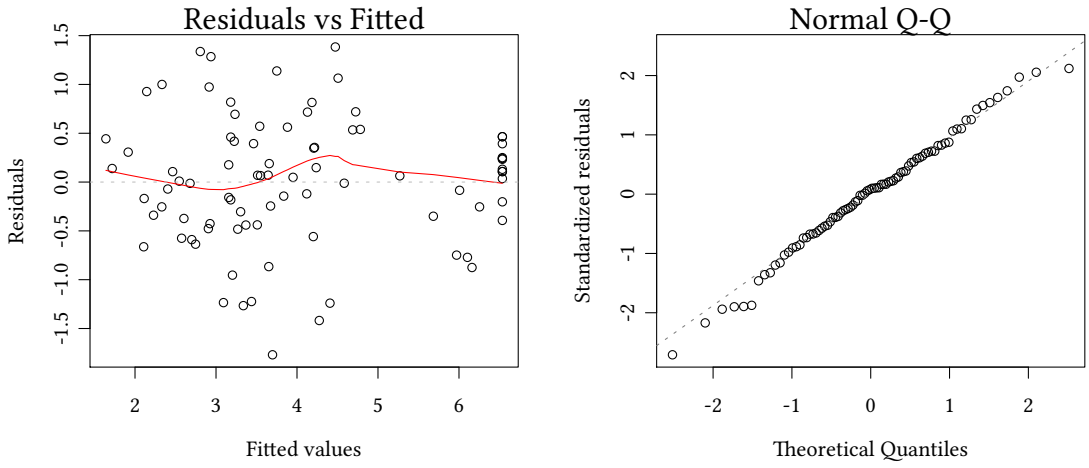



Figure 4.7: Diagnostic plots for $(\textsc{onset} \times \textsc{cpitch} \times \textsc{cpint}, L^1, P)$.

Violation of independence is particularly problematic for time-series data, where there can often be strong colinearity between variables. This was unlikely to be a problem with this data as firstly, it is not a time series, and secondly, the sequence of comparisons between original melodies and their variations was randomly determined. Nonetheless, Durbin-Watson tests were performed on each model to test for independence. The DW-statistic ranges from 0 to 4, and a value of 2 represent no autocorrelation. Very little residual autocorrelation was found across all models ($1.82 < DW < 2.13$).

Applying the Global Validation of Linear Models Assumptions test (Peña and Slate 2006) provided additional support that the assumptions were acceptable for all models according to the global statistic ($p > 0.05$). However, significant skewness ($p < 0.05$), violating the assumption of normality, was identified for the following four models:

- $(\textsc{onset} \times \textsc{cpitch}_c \times \textsc{ioi}, L^1, C)$
- $(\textsc{onset} \times \textsc{cpitch}, L^1, C)$
- $(\textsc{onset} \times \textsc{cpitch} \times \textsc{dur}, L^1, C)$
- $(\textsc{onset} \times \textsc{cpitch} \times \textsc{ioi}, L^1, C)$

Results for these models should be treated with extra caution. These models are kept in the study for completeness, yet, notwithstanding caution, they do not appear to exhibit any distinguishing qualities and thus are not subject to subsequent detailed analysis.

### $L^1$ vs. $L^2$ norm

The choice of ground distance norm is shown not to significantly affect model accuracy in the majority of cases. As shown in figure 4.8, out of the 24 model pairs, only five $L^1$ models performed significantly better than the corresponding $L^2$ variant at $\alpha = 0.05$ level, and only two at $\alpha = 0.01$ level. However, even when the difference was not significant, every $L^1$ norm model slightly outperformed its $L^2$ counterpart. Although the evidence is not strong enough to support our hypothesis that $L^1$ norm is the most appropriate for these models, as a practical measure, subsequent analysis will concentrate on the relative performance of $L^1$ norm models only.

Our prediction that $L^1$ normed spaces would perform better than $L^2$ spaces was based on Gärdenfors' rule of thumb that the $L^1$ measure is more appropriate for separable dimensions, and $L^2$ more appropriate for integral dimensions. However, another possibility, carrying Gärdenfors' logic still further, would be to construct hierarchical domains, each with different norms as appropriate. For example, to

define separate pitch and time subspaces, incorporating all respective absolute and relative quality dimensions, each equipped with an $L^2$ norm on the grounds that qualitative differences within each domain are integral, and then treat the resulting distances as points in an $L^1$ normed meta-level space. However, if one also considers the problem of combining absolute and relative surface-level features as discussed in section 4.2, for which one possible solution could involve treating absolute and relative features as weighted points sets in different spaces, then a simple hierarchical composition of pitch and time subspaces becomes problematic. However, further experimentation with new models in this direction could provide new insight into how various surface-level musical qualities interact in perception.

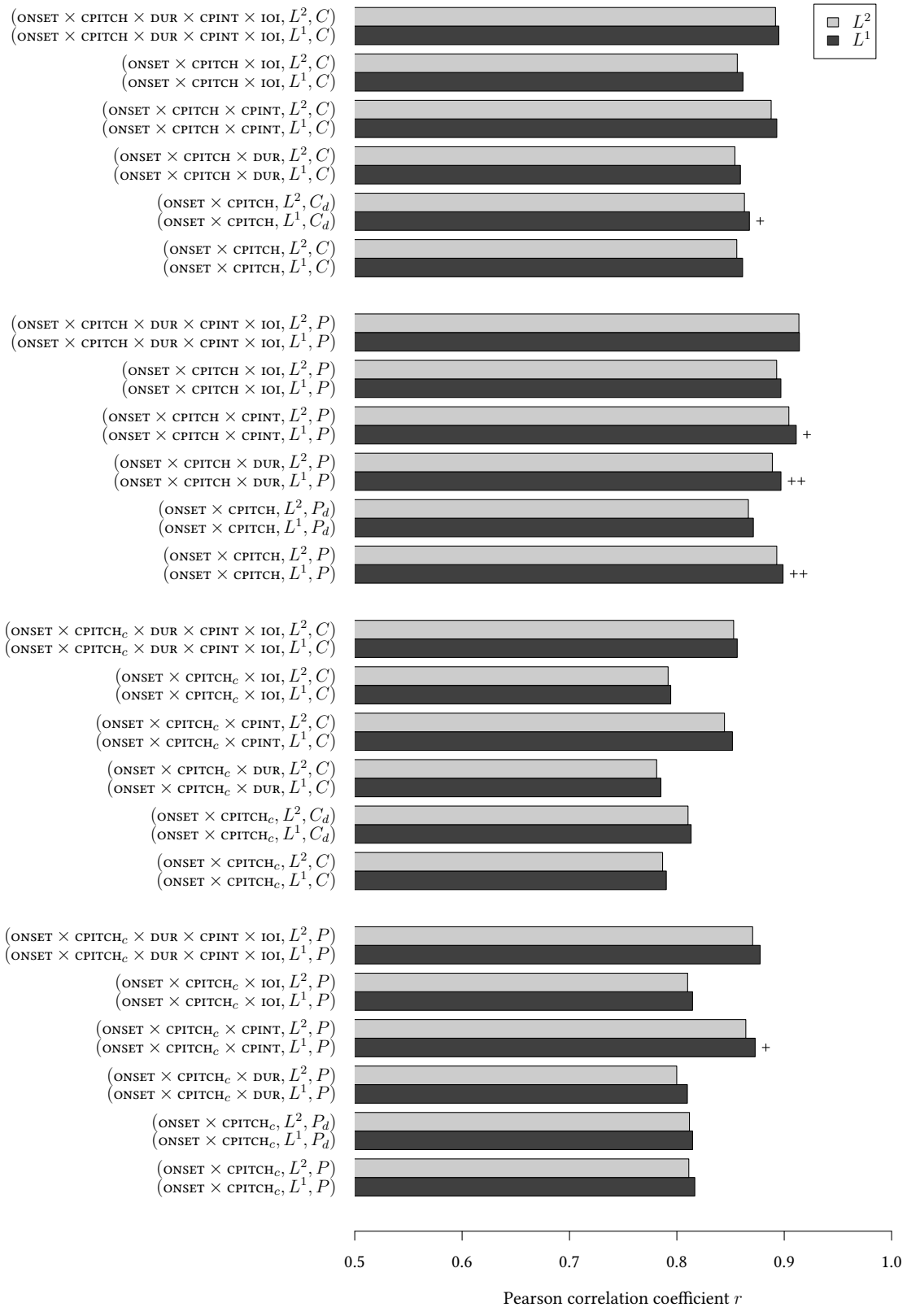



Figure 4.8: Pearson correlation between all models and human similarity ratings comparing $L^1$ and $L^2$ models. Significance in the difference between $L^1$ and $L^2$ model pairs at $\alpha$ = 0.05 and $\alpha$ = 0.01 is denoted by $+$ and $++$ respectively.

**Partial vs. complete matching**

In six out of the twelve comparisons between $L^1$ models shown in figure 4.9, partial match EMD performed significantly better than complete match EMD at $\alpha = 0.05$ level. Three out of these six, all using non-centred pitch representations, performed significantly better at $\alpha = 0.01$ level. In all other cases where a difference was not significant, partial matching slightly out-performed complete matching. This general trend in favour of partial matching lends support to our hypothesis that partial matching EMD is a more appropriate measure for this set of broadly homogeneous and roughly equal length melodic stimuli.

Using duration as weight was a slight anomaly here in that performance was virtually identical in both partial and complete matching contexts:

- $(\textsc{onset} \times \textsc{cpitch}_c, L^1, P_d)$, $r(82) = 0.815$, $p < 0.01$
- $(\textsc{onset} \times \textsc{cpitch}_c, L^1, C_d)$, $r(82) = 0.813$, $p < 0.01$
- $(\textsc{onset} \times \textsc{cpitch}, L^1, P_d)$, $r(82) = 0.871$, $p < 0.01$
- $(\textsc{onset} \times \textsc{cpitch}, L^1, C_d)$, $r(82) = 0.868$, $p < 0.01$.

It would be wrong to draw strong conclusions from these results, particularly given the wider results regarding duration below suggesting that duration is not a particularly useful discriminatory feature within these melodies. However, these results at least suggest that further research into musically meaningful EMD weights is warranted, particularly in the context of complete matching. In all the complete matching EMD models examined here except for the duration-weighted variants, the weights associated with events must sum to one, leading to a fluctuation in event weight between melodies depending on their length, which is somewhat arbitrary from a cognitive perspective. While using duration-based weights in complete matching does not result in improved performance, it at least does not follow the trend of decreased performance. Therefore, given different kinds of melodic stimuli where partial matching is known not to perform well, complete matching with musically meaningful weights may offer an alternative.

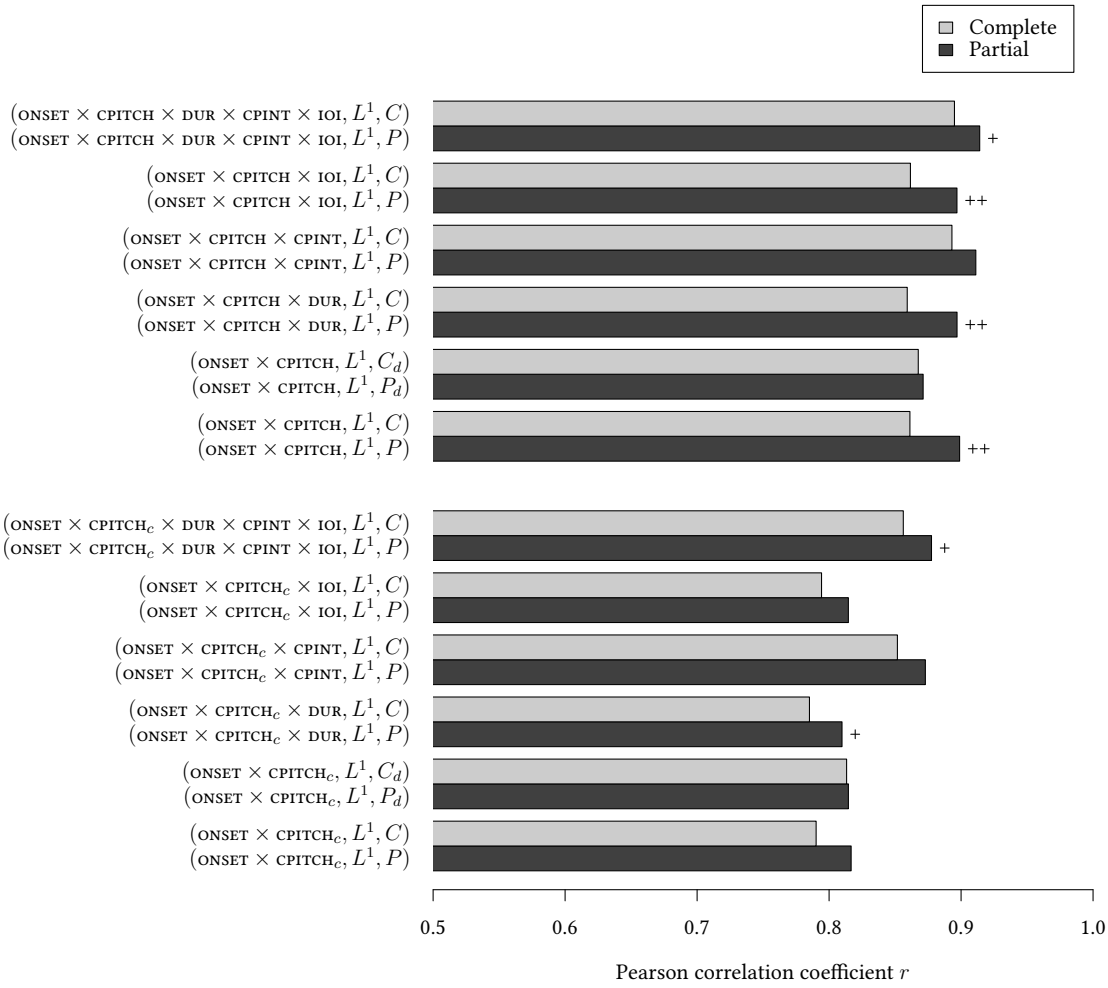



Figure 4.9: Pearson correlation between all $L^1$ models and human similarity ratings comparing partial and complete matching EMD. Significance in the difference between partial and complete matching model pairs at $\alpha$ = 0.05 and $\alpha$ = 0.01 is denoted by $+$ and $++$ respectively.

### Centred vs. original pitch height

Aligning melodies in the chromatic pitch dimension by weighted mean pitch height was found to significantly degrade performance across the board. As can be seen in figure 4.10, all $L^1$ models using original pitch height scored significantly better than their centred pitch height counterparts at $\alpha$ = 0.01 level.

For models containing a relative representation of pitch height (CPINT), the difference in performance between centred and non-centred model variants was typically less than the difference between models containing only an absolute repre-

sentation of pitch (CPITCH or $\text{CPITCH}_c$). This intuitively makes sense because CPINT is independent of transposition, and can therefore provide a degree of resilience to the distortion caused by centring absolute pitch height. Furthermore, examining the optimised salience weights (see appendix C) for centred pitch-height models containing a CPINT dimension reveals that CPINT has a salience weight of between 1.67 and 2.24 relative to the salience weight of $\text{CPITCH}_c$, which is always fixed at 1.0. Whereas when no centring is used, the salience weight associated with CPINT is between 0.57 and 0.93, again relative to the fixed salience weight of 1.0 associated with CPITCH. This suggests that when absolute pitch height information is compromised by centring—$\text{CPITCH}_c$ is used as opposed to CPITCH—relative pitch information becomes more relevant for predicting similarity.

The direction of this result is unsurprising given that all stimulus melodies are in the same key and register. Furthermore, the magnitude of the result suggests that the centring principle is inappropriate in contexts where key and register are controlled. However, a question remains as to the appropriateness of centring melodies by weighted mean pitch height in situations where key and tessitura are uncontrolled. Typke (2007) acknowledges that the centring principle offers no guarantee of an optimal alignment minimising EMD, and is only a simple and efficient method for processing melodies of differing tessiture. Our results in section 4.4 support this: some method of addressing the variation in pitch range across real-world corpora is absolutely necessary, and centring by weighted mean pitch height works well, particularly in conjunction with a relative pitch representation. However, the degree to which centring degrades the ability to predict similarity between melodies when key and tessitura are controlled strongly suggests that more musically sophisticated alignment methods are needed, possibly in conjunction with additional pitch dimensions representing aspects of tonality.

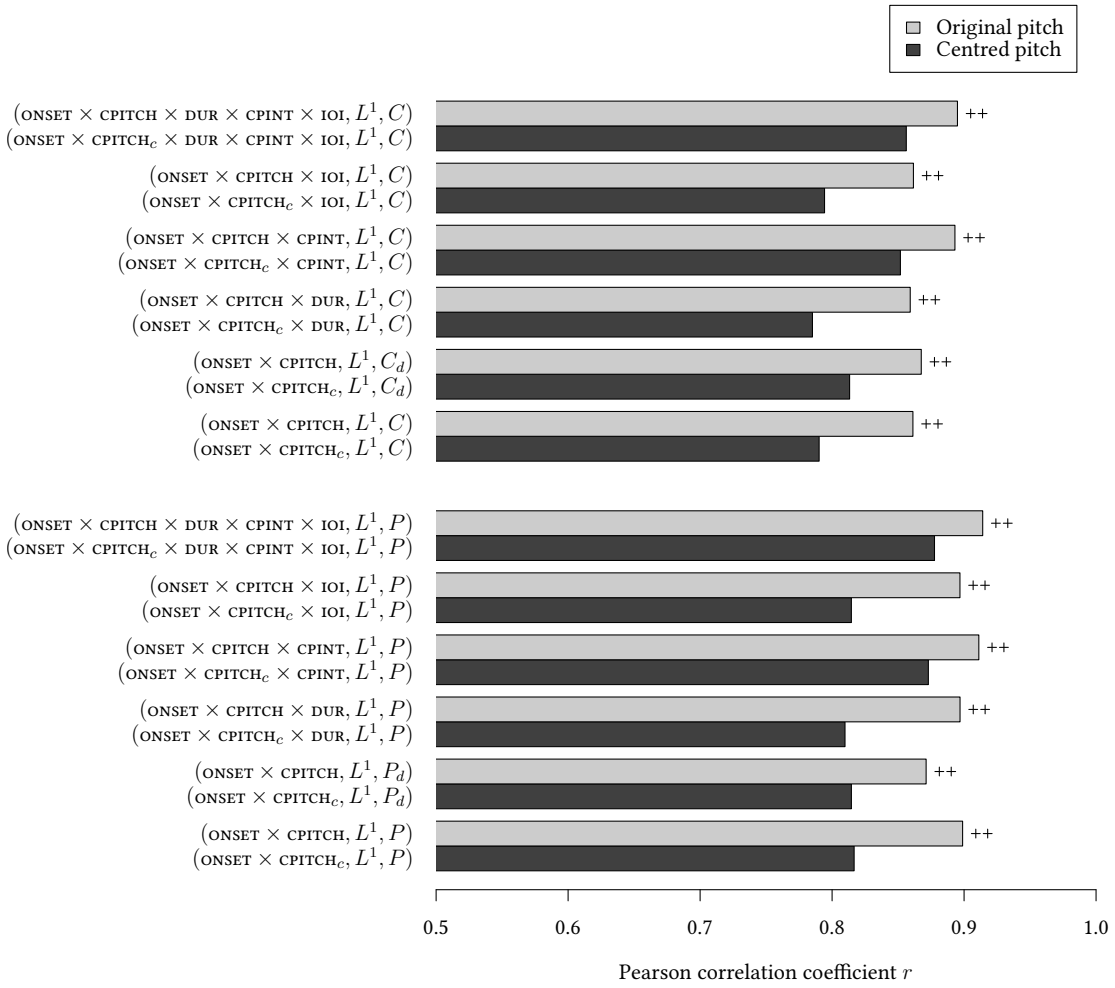



Figure 4.10: Pearson correlation between all $L^1$ models and human similarity ratings comparing centred and non-centred models. Significance in the difference between centred and non-centred model pairs at $\alpha$ = 0.01 is denoted by $++$.

### Duration as EMD weight vs. quality dimension

The inclusion of duration information across all model variants did not result in any significant improvement compared to baseline ONSET × CPITCH models. The only significant difference ($\alpha$ = 0.05) between a baseline model and a duration model was in the case of (ONSET × CPITCH, $L^1$, $P$) versus (ONSET × CPITCH, $L^1$, $P_d$), where the use of duration as EMD weight significantly decreased performance ($t(82)$ = 2.03, $p$ = 0.045). The baseline model in this case is the best performing model in this subset, which brings into question the assumption made by Typke (2007) that duration is generally appropriate as the EMD weight when predicting

similarity between short melodic stimuli.

In both complete matching conditions, using duration as EMD weight did result in a slight increase in performance compared to the respective baseline models, but did not reach significance:

- $(\textsc{onset} \times \textsc{cpitch}_c, L^1, C_d)$ $t(82) = -1.57$, $p = 0.12$
- $(\textsc{onset} \times \textsc{cpitch}, L^1, C_d)$ $t(82) = -0.45$, $p = 0.65$.

However, as shown in figure 4.11 these two models do not exhibit any improved predictive ability over their partial-matching counterparts. While inconclusive, this result at least suggests that complete-matching EMD models may benefit from weights based on musical features, but that using duration in this case does not significantly improve prediction.

Including duration as a dimension in the ground distance instead of as EMD weight resulted in almost identical performance to the baseline models. Furthermore, in all cases the optimised salience weights for DUR ($M = 0.098$, $SD = 0.05$) were less than the salience weights for ONSET ($M = 0.16$, $SD = 0.04$), with CPITCH and $\textsc{cpitch}_c$ held constant at 1. For this corpus of melodies at least, the evidence suggests that duration information is not generally useful in predicting melodic similarity using EMD-based models.

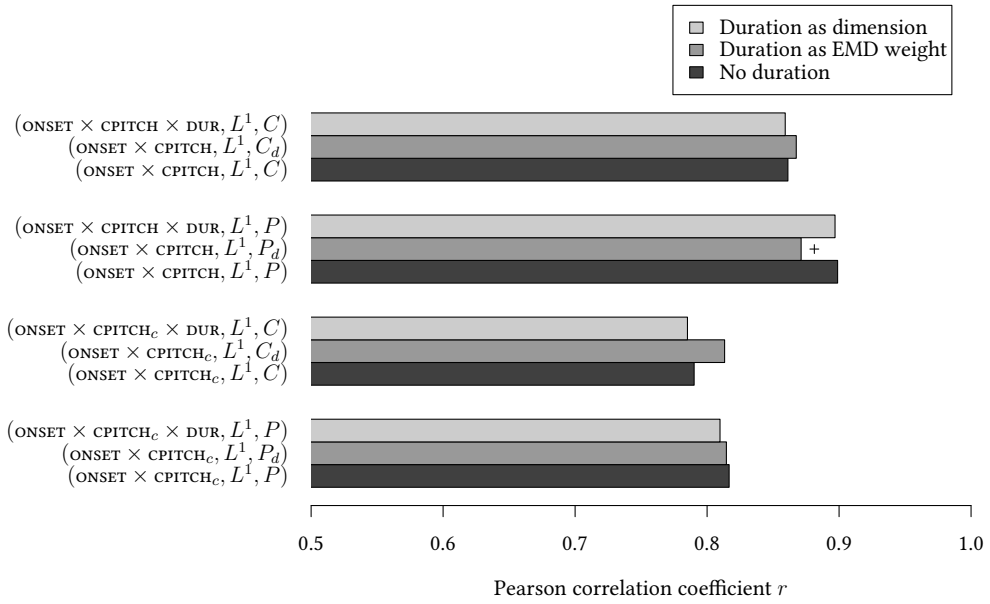



Figure 4.11: Pearson correlation between $L^1$ duration models and human similarity ratings, comparing the baseline model with duration-as-EMD-weight, and duration-as-dimension models. Significance in the difference between a duration model and the baseline model at $\alpha$ = 0.05 is denoted by $+$.

### Quality dimension comparison

The final perspective in our analysis investigates the impact of different combinations of quality dimensions on model performance. We do not include the duration-based weighting schemes, as these were shown above to not result in any significant improvement. However, both complete and partial matching conditions are included, despite the evidence suggesting that partial matching generally leads to more accurate predictions. Similarly, both centred and non-centred pitch dimensions are included as separate conditions, despite the evidence indicating that centring melodies by weighted mean pitch height decreases model accuracy for this corpus. These conditions are included here so that we may conclude which quality dimensions lead to better models within each condition. It is important to address this question so we are equipped to tackle the subsequent experiment using the MIREX 2005 data in section 4.4. We shall answer the question of which model is the optimal EMD-based model in terms of accuracy of the data provided by Müllensiefen and Frieler (2004). However, establishing a small number of optimal models corresponding to each condition will allow us to evaluate each in turn

with respect to the MIREX 2005 data.

The two best performing models in each complete/partial, centred/not-centred condition are shown in figure 4.12. The same set of quality dimensions appear consistently in the best two models for each condition. The model that most accurately predicts the human ratings, in all conditions, includes all quality dimensions investigated in this study. Table 4.1 shows the performance measures for the two overall best EMD models (partial matching and non-centred conditions), together with the best model developed by Müllensiefen and Frieler (2004), which combines a rhythmically weighted edit-distance measure (*rawedw*), with a measure based on the number of common n-grams (*ngrcoord*). Interestingly, the difference in overall fit between the geometrical EMD models and the symbolic *rawedw+ngrcoord* model is very slight, despite the approaches being fundamentally different. Further analysis of the behaviour of each model in terms of the musical characteristics of the stimuli may offer some insight into how the different approaches are able to offer comparable accounts of melodic similarity.

Table 4.1: Statistics of the best two EMD models, together with the best model developed by Müllensiefen and Frieler (2004).

| Model | Std. err | $R^2$ | Adj. $R^2$ | *F (DF)* | *p* |
|---|---|---|---|---|---|
| $(\text{ONSET} \times \text{CPITCH} \times \text{DUR} \times \text{CPINT} \times \text{IOI}, L^1, P)$ | 0.647 | 0.836 | 0.834 | 416.8 (1, 82) | < 0.01 |
| $(\text{ONSET} \times \text{CPITCH} \times \text{CPINT}, L^1, P)$ | 0.657 | 0.830 | 0.828 | 401.0 (1, 82) | < 0.01 |
| *rawedw+ngrcoord* | 0.66 | 0.830 | 0.826 | | |

The difference between the best EMD models containing all quality dimensions, and models containing only CPINT in addition to the baseline space is not significant under any condition. Applying Occam's razor, we may conclude that duration and inter-onset interval are not useful features in this dataset for these EMD models, and that a ground distance comprising only onset, pitch, and pitch interval information gives comparable performance with fewer dimensions. Furthermore, examining the mean optimised salience weights across models containing all quality dimensions in table 4.2, the DUR salience weight is close to zero, and therefore has very little impact on the computation of ground distance. The average salience weight for IOI is also low relative to pitch dimensions, but in fact slightly higher than ONSET. Given that the inclusion of IOI does not significantly improve prediction, a reasonable interpretation would be that it does not contribute any useful information beyond what is already available in ONSET. By analogy with the relationship between CPITCH, $\text{CPITCH}_c$ and CPINT discussed in section 4.3.2, if translation in time was a prominent characteristic of the stimuli, for

example, similar melodies beginning at different metrical positions, then we would expect that the time invariance afforded by IOI would become more important in modelling similarity.

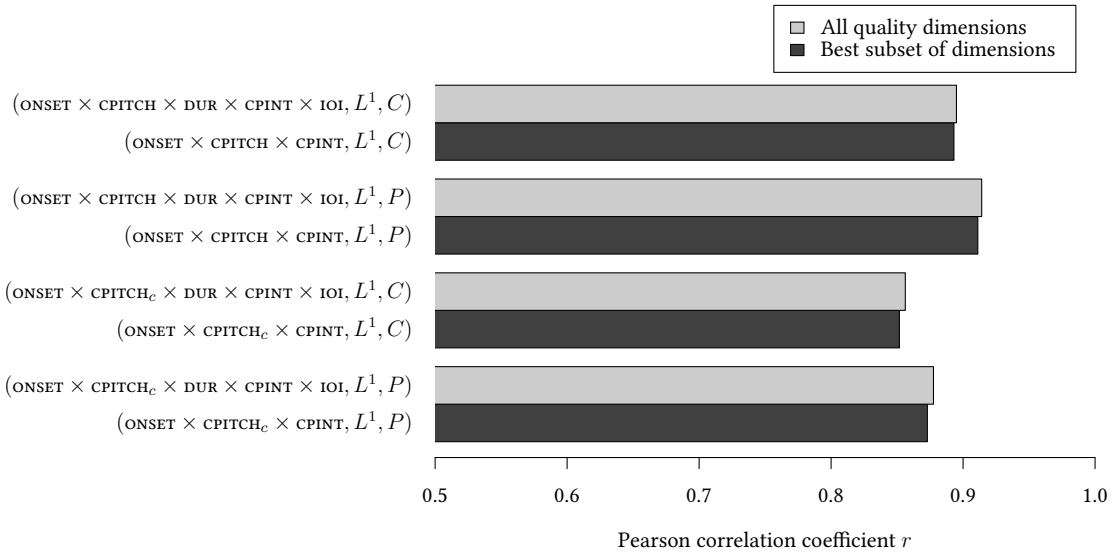



Figure 4.12: Pearson correlation between the best two performing $L^1$ models and the human similarity ratings across each partial/complete match and centred/not centred condition. There is no significance in the difference between any model pair.

Table 4.2: Average salience weights across EMD models containing all quality dimensions.

| Dimension | Mean | Standard Deviation |
|---|---|---|
| ONSET | 0.196 | 0.052 |
| CPITCH/ CPITCH$_c$ | 1.000 | 0.000 |
| DUR | 0.025 | 0.012 |
| CPINT | 1.453 | 0.448 |
| IOI | 0.286 | 0.156 |

To quantify the significance in the difference between our established best models comprised of onset, pitch and pitch interval quality dimensions, we compare each partial/complete matching and centred/not-centred variant with the other ground distance spaces defined in this study. As can be seen in figure 4.13,

onset, pitch and interval models perform significantly better in all conditions except for the partial matching non-centred condition. As discussed above, the lack of significance in this condition may be attributable to the finding that when these melodies are compared at their original pitch, the pitch dimension alone affords accurate prediction of similarity. However, when taking into account that all melodies in this dataset are in the same key and tessitura, this result is unlikely to generalise to other data when key and tessitura are variable qualities.

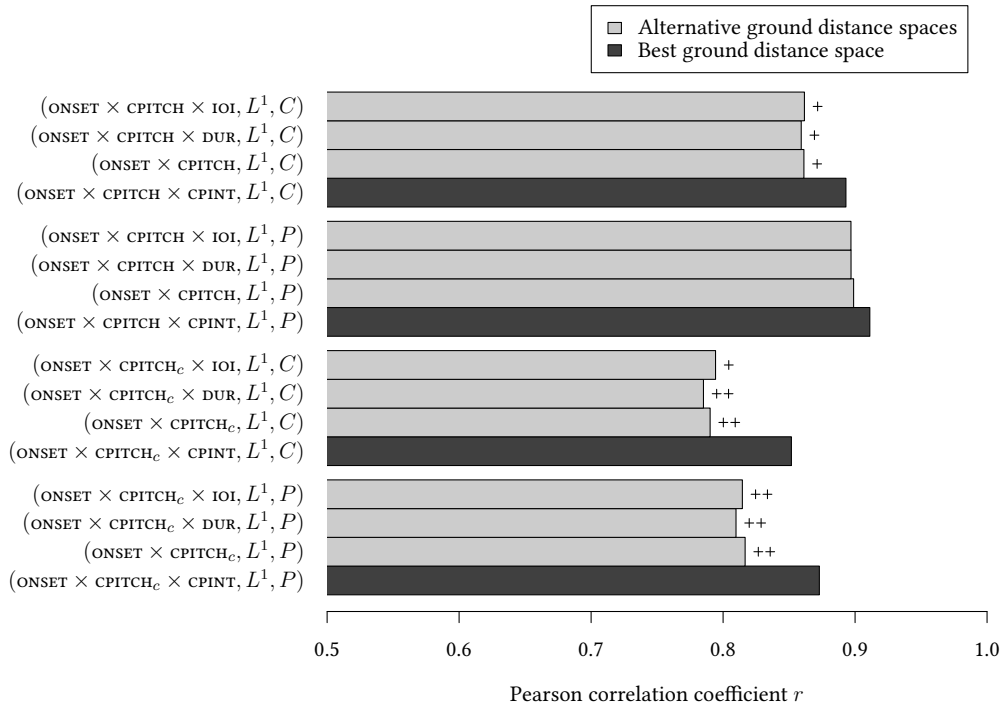



Figure 4.13: Pearson correlation between human similarity ratings and all onset, pitch and pitch interval models compared to all other ground distance space models. Significance in the difference between the best models and alternative space models at $\alpha$ = 0.05 and $\alpha$ = 0.01 is denoted by $+$ and $++$ respectively.

The complete description of the best models in each partial/complete match and centred/not-centred conditions are as follows:

- $(0.30 \cdot \textsc{onset} \times 1.0 \cdot \textsc{cpitch}_c \times 2.24 \cdot \textsc{cpint}, L^1, P)^{0.58}$
- $(0.12 \cdot \textsc{onset} \times 1.0 \cdot \textsc{cpitch} \times 0.57 \cdot \textsc{cpint}, L^1, P)^{0.56}$
- $(0.26 \cdot \textsc{onset} \times 1.0 \cdot \textsc{cpitch}_c \times 1.81 \cdot \textsc{cpint}, L^1, C)^{0.58}$
- $(0.13 \cdot \textsc{onset} \times 1.0 \cdot \textsc{cpitch} \times 0.93 \cdot \textsc{cpint}, L^1, C)^{0.56}$

These models will be evaluated further in experiment 2.

## 4.4 Experiment 2: Melodic-based music information retrieval

The purpose of experiment 2 is to evaluate the best melodic similarity models developed in experiment 1 (section 4.3) in a different context in order to test their ability to generalise to unseen data. The models in experiment 1 were optimised with respect to human similarity ratings between melodies from a carefully constructed set of stimuli intended to probe the concept of expert judgements of melodic similarity. Therefore, given the assumption that the stimuli used in experiment 1 are sufficiently representative of a notional psychological 'space of melodic similarity', and our models are not over-fitted to the stimuli, we predict that our models will be able to accurately predict melodic similarity given novel stimuli.

### 4.4.1 Original MIREX 2005 evaluation

The Music Information Retrieval Evaluation eXchange[13] (MIREX) is an annual evaluation of MIR algorithms, run in conjunction with the International Society of Music Information Retrieval (ISMIR) conference.[14] Each year a range of tasks is proposed and agreed upon by the community, and then algorithms are submitted to the International Music Information Retrieval Systems Evaluation Laboratory (IMIRSEL) based at the University of Illinois at Urbana-Champaign, where all experiments are carried out.

For the 2005 MIREX evaluation, Rainer Typke proposed a symbolic melodic similarity task based on the retrieval of short melodic passages, specifically incipits.[15] Final results can be found on the MIREX wiki.[16] Data for this task was a subset of the RISM[17] A/II collection, which is comprehensive incipit-based index of notated music from 1600. The test corpus contains 558 incipits.[18] The task consists of generating ranked lists of the most similar incipits from the corpus in response to eleven query incipits. Ground truth rankings were established prior to the experiment by Typke et al. (2005) by combining judgements of similarity

[13] http://www.music-ir.org/mirex/

[14] http://www.ismir.net/

[15] http://www.music-ir.org/mirex/wiki/2005:Symbolic_Melodic_Similarity

[16] http://www.music-ir.org/mirex/wiki/2005:Symbolic_Melodic_Similarity_Results

[17] http://www.rism.info/

[18] Some confusion exists around the exact size of the corpus used. Larger figures are indicated on the evaluation wiki page, whereas the results page indicates 558 incipits were used. Through personal correspondence with Rainer Typke (Jan–Feb 2011), Klaus Keil, director of RISM (Feb 2011) and Stephen Downie, director of IMIRSEL (Mar–Apr 2011), we believe that the corpus used in the experiment reported below is identical to that used in the original MIREX evaluation.

from musical experts. The rankings of melodies are not given as a single ordered list, but are based on grouping of similar melodies, reflecting the broad agreement between judges. Ordering within groups is irrelevant to the evaluation metric, it is rather the between-group ordering that is taken into account, which is formalised in a measure called Average Dynamic Recall (ADR) (Typke 2007, pp. 93–95). We use Typke's original implementation of ADR in our experiment (personal communication, January 11, 2011), which was also used in the MIREX evaluation.

A similar sized training corpus, with accompanying ground truth for eleven different queries, was made available for training purposes prior to the MIREX evaluation. We make no use of the accompanying ground truth for this training set in our experiment, because the salience weights of our models are pre-determined from the optimisation processes conducted within experiment 1, reported in section 4.3. However, we do make use of the training corpus itself for the purpose of extracting descriptive statistics about melodies from the RISM collection, which are then applied during the process of projecting melodies from the MIREX testing set into standardised quality dimensions.

We do not optimise the salience weights of our models again here using the MIREX training set ground truth, as in the original MIREX evaluation, because we are primarily interested in the ability of our models to generalise from training over psychologically validated data that has been carefully curated to encompass a representative diversity of melodic similarity judgements. Typke's ground truth is based on the judgements of musical experts, so can be considered equally psychologically valid for this task. However, the relatively small number of randomly selected query melodies offers less assurance against bias. Furthermore, Müllensiefen and Frieler (2004) have demonstrated that the notion of stable expert agreement on judgements about melodic similarity can exist, at least within the domain of popular melodies. If it is possible to apply models developed in this context to novel corpora, then this kind of methodology offers a potential alternative to repeating computationally expensive training processes on every new corpus.

We are interested primarily in the performance of our EMD models in comparison to the EMD-based algorithm submitted to the MIREX evaluation by Typke (2007), which incorporates EMD within a sophisticated retrieval system. Here we highlight the differences between the two approaches. Typke's algorithm uses onset time and chromatic pitch dimensions in computing the ground distance matrix, and duration is used as the EMD weight. The relative scaling of ground distance dimensions is determined by first segmenting each query into possibly overlapping segments of between six and nine consecutive notes, and then applying an evo-

lutionary algorithm to find an optimal alignment between each segment and the target melody under consideration. The query segments may be translated in both pitch and time, as well as scaled in time. The evolutionary algorithm searches for an optimal alignment which minimises the overall distance between the two point sets. After the optimal alignment is determined for each query segment, EMD is then computed for each segment with respect to the target melody. The multiple EMD measures are then combined into a single measure of similarity. The use of segmented queries in combination with the duration-based weighting scheme means that partial matching is used between each segment and the target melody. This algorithm is very computationally expensive, requiring more than 14 hours of run time in the MIREX evaluation, while the next slowest (and best performing) algorithm only required 80 seconds.[16] However, it has been designed to be robust against transposition in pitch, scaling and translation in time, and against melodies of variable length.

In contrast, our models are much simpler, consisting essentially of a single computation of EMD between each pair of melodies. No segmentation is performed, and the only alignment strategies imposed prior to computing EMD are the transposition of melodies by mean weighted pitch height for models using the $\textsc{cpitch}_c$ dimension, and ensuring all empty bars are removed from the beginning of each incipit as described in section 4.2.1. The salience weights for ground distance dimensions are pre-determined from the psychological data modelled in experiment 1, reported in section 4.3. While our approach is very simple in terms of algorithmic complexity relative to Typke's system, it is arguably more sophisticated in terms of representation. Our modelling strategy places greater emphasis on establishing a perceptually-motivated conceptual space within which ground distance is computed. All our models examined here include a relative pitch dimension, which was shown in experiment 1 to improve prediction. Furthermore, the projection of basic melodic attributes into quality dimensions is standardised according to statistics derived from the MIREX training corpus, thus affording a degree of model adaptability when presented with unseen data.

### 4.4.2 Results

The results for our best four models determined in section 4.3.2 are presented in table 4.3 alongside the results from the original MIREX 2005 evaluation. Our best performing model, using partial matching and the centred-pitch representation, achieved an ADR of 59.14% and ranked fourth out of the total eleven systems (seven original MIREX-evaluated systems, plus our four best EMD models). This

model performed slightly better compared with the complete matching centred pitch variant (57.19%, ranked fifth), both of which performed slightly better than Typke's algorithm (57.09%, here ranked sixth).

Table 4.3: MIREX 2005 results.

| Rank, | Participant | Average Dynamic Recall |
|---|---|---|
| 1 | Grachten, Arcos & Mántaras | 65.98% |
| 2 | Orio, N. | 64.96% |
| 3 | Suyoto & Uitdenbogerd | 64.18% |
| 4 | $(\textsc{onset} \times \textsc{cpitch}_c \times \textsc{cpint}, L^1, P)$ | 59.14% |
| 5 | $(\textsc{onset} \times \textsc{cpitch}_c \times \textsc{cpint}, L^1, C)$ | 57.19% |
| 6 | Typke, Wiering & Veltkamp | 57.09% |
| 7 | $(\textsc{onset} \times \textsc{cpitch} \times \textsc{cpint}, L^1, C)$ | 56.43% |
| 8 | Lemström, Mikkilä, Mäkinen & Ukkonen (P3) | 55.82% |
| 9 | $(\textsc{onset} \times \textsc{cpitch} \times \textsc{cpint}, L^1, P)$ | 54.89% |
| 10 | Lemström, Mikkilä, Mäkinen & Ukkonen (DP) | 54.27% |
| 11 | Frieler & Müllensiefen | 51.81% |

The result that both centred pitch condition models perform better than non-centred pitch models is unsurprising given that key and tessitura are in no way standardised in this dataset, as they are for the psychological data provided by Müllensiefen and Frieler (2004). Furthermore, given that we have shown in section 4.3.2 that centring melodies by mean weighted pitch height degrades performance when key and tessitura are controlled, using a musically informed pitch alignment method may improve results further.

Partial matching slightly out-performs complete matching over this data, in line with the general trend in favour of partial matching found for the psychological data. Therefore, our hypothesis from section 4.2.3 that complete matching EMD would perform better over this dataset due to the increased variation in melody length proved not to be supported. However, complete matching EMD did show to be more resilient than partial matching EMD in the non-centred pitch height condition.

The differences between our EMD-based models and Typke's model are very slight, and should not be over interpreted. However, we are encouraged that the relative simplicity of our approach, based on a cognitively-motivated representation, exhibits comparable predictive accuracy to a much more computationally expensive method.[19]

[19]Run-time for our models in this experiment is approximately 10 seconds on a 2.2GHz processor

Our models could not match the performance of the best three performing models in the original MIREX 2005 evaluation. The best three original MIREX algorithms are symbolic, string-matching algorithms: a music-theoretic based edit-distance (Grachten et al. 2005); an index built from n-grams of musical features employing standard text processing retrieval techniques (Orio 2005); and another n-gram-based index using a modulo-12 pitch representation (Suyoto and Uitdenbogerd 2005). String matching algorithms are very well suited to tasks requiring the comparison of monophonic melodies; they are well-studied in the literature and can typically be implemented very efficiently. Furthermore, as shown by Müllensiefen and Frieler (2004) string-matching models also perform very well in modelling psychological data. However, a general difficulty with string matching is extending the techniques to polyphonic music, something that is very naturally accomplished in a geometric representation. Geometrical conceptual space models also offer a philosophical framework and methodology for addressing the problem of representational semantics, and if one takes the view proposed by Gärdenfors that representation may be considered as a continuum from low-level perceptual signals up to abstract symbols, then seeking to develop new models able to operate over multiple levels of representation would seem to be a promising direction.

The fact our models could not offer superior performance to the best three systems does not invalidate our finding that employing a perceptually-motivated approach based on the conceptual space theory leads to not only a more simple EMD-based algorithm for predicting melodic similarity, but one which is also capable of generalisation to novel stimuli. Each of the best models tested in the original MIREX 2005 evaluation were trained on a training set randomly selected from the same population as the evaluation dataset, and were better able to optimise their performance to the specific task. Our aim was not exclusively to develop the best retrieval algorithm for this particular task, but more importantly to consider the more general question of how psychological understanding may inform and be incorporated into the development of an MIR system. Our models preform well, despite incorporating only a minimal subset of potential musical features, or employing alignment or segmentation methods—techniques that are know to be advantageous in modelling melodic similarity.

machine, compared to over 14 hours for Typke's system on a 1.6Ghz processor machine. Performance is not directly comparable to the original evaluation benchmarks due to the use of different hardware. Nevertheless, the difference is considerable.

## 4.5 Conclusion

The use of the EMD metric for computing distance between sets of points within the conceptual space framework represents a novel application of Gärdenfors' original formulation. However, whether the method represents a departure or development of the conceptual space theory is an open question. On the one hand it is a development enabling complex time-valued concepts to be modelled within quality dimensions. On the other hand, it is a departure because melodic concepts are not identifiable with points, or even regions, of a space. Strictly, only musical events are represented as concepts within these spaces. In order for melodies, as defined here, to be identified as points, a much higher-dimensional space is required, containing a dimension corresponding to each attribute of each event. However, we are then faced with the problem of how to represent melodies containing different numbers of events, because within a vector space formalism a fixed number of dimensions is required in order to calculate distance.

The issue of representing sequentially structured concepts within the conceptual space framework deserves further consideration. This question arises again in the following chapter concerning the geometrical representation of metre. A pure vector space formalism is possible in that case, because we are able to specify metrical concepts within finite temporal bounds, although high-dimensionality is still required. Future work might usefully reconsider the boundaries we have set here between the conceptual, and other potential underlying sub-conceptual representations. It may be the case that conceptual space representations are more appropriate for abstracted melodic features, which map to lower level, not necessarily geometrical, representations of explicit sequential structure.

# Chapter 5

# Conceptual space representations of perceived rhythmic structure

This chapter presents a formalisation of metrical-rhythmic concepts within the framework of Gärdenfors' theory of conceptual space (Gärdenfors 2000). Specifically, two conceptual space models are developed, which encapsulate salient aspects of the experience of metrically organised rhythmic structure. These models are evaluated according to their ability to discriminate between rhythmically distinctive genres of music.

Central concepts of musical timing prevalent in music notation are first discussed in order to clarify terms and concepts familiar to musicians. Then key empirical research related to temporal aspects of auditory perception is reviewed, concentrating on the theory of metre developed by London (2004; 2012). The scientific study of time perception in music aims to better understand the capacities and constraints of our perceptual mechanism and associated implications for musical experience. London's theory draws upon research from psychology, cognitive science and neuroscience, and provides the empirical foundation upon which we construct a computational theory of rhythmic similarity.

## 5.1 Notation and music theory

Many writers, from a range of disciplines, have attempted to define and explain musical rhythm. Traditionally, arguments have tended to characterise rhythm objectively, basing theories on rhythmic structure as represented in musical notation (Lerdahl and Jackendoff 1983). Such theories are open to criticism on the grounds of misrepresenting the perceptual nature of musical experience. However, the pragmatic origins of notational conventions, as a means of remembering and commu-

nicating musical ideas, has ensured that a high degree of musical understanding can be inferred by the encultured reader of notation. Furthermore, later empirical work broadly lends support to the psychological validity of many music-theoretic concepts (Gabrielsson 1993). This section reviews some of the theoretical ideas that have developed around notational concepts of musical time, before moving on to more recent empirically grounded theories.

What has become known today as Common Musical Notation (CMN), or staff notation, is a means of representing various aspects of music in symbolic form. The conventions of CMN can be traced back to musical practices of the 9th and 10th centuries (Pryer 2010). Early forms of notation were simply means to aid memory, and did not place emphasis on precisely describing either pitch or timing information. Over time, notational conventions developed, along with music theoretic concepts, which allowed for more precision in the description and communication of musical intentions. Scholarly activity, pragmatic musical concerns, as well as extra-musical factors such as the medium of inscription, were all drivers of this change towards increasing specificity. The history of Western classical music is inextricably linked with the development of notation, so much so that it can be seen as playing a large role in shaping compositional concerns and performance practice (Cook 1998; Wishart 1996).

Common musical notation is essentially a graph of pitch against time, where within each line of music, time runs from left to right along the $x$-axis, and pitch is represented from low to high on the $y$-axis. This view of musical structure is even more explicit in piano-roll notation, common in many computer-based music editing tools. Armed with a few additional concepts, notation can be seen as a reasonable analog to many of the aspects of music that are typically considered salient, at least by an encultured listener of traditionally notated Western music. The time dimension of notation is obviously of primary relevance to rhythm, and importantly encodes useful information beyond simply the temporal order of events.

The notation of musical ideas must always balance level of detail against clarity of expression. Furthermore, even for scores containing extremely high levels of detail, corresponding performances, except those mechanical in origin, will always involve varying degrees of temporal deviation from the notated values (Desain and Honing 1993). In surveying the history of rhythm, London (2009) discusses examples of both early and contemporary forms of notation that leave many aspects of musical timing to the discretion of the performer—for example, medieval systems of notation that rely on syllable lengths to co-ordinate the pace of musical lines, or Berio's Sequenza III (1965), which relies on a stopwatch for temporal organisation.

Both these forms of synchronisation are in a sense external to the music, imposing their own intrinsic characteristics on musical structure irrespective of any notion of musical flow that may be suggested by a particular succession of musical events in time. A fairly intuitive means of notating timing information is simply to rely on spatial distance within the score, where the distance between note symbols is intended to indicate the period of time between successive note onsets. Spacing and orthographic conventions generally play an important role in notation. However, the vast majority of common practice music, as well as notations of popular music, presuppose a richer conceptualisation of the structure of musical time.

Intuitively, events that occur simultaneously in a performance of a notated piece of music are aligned vertically in the score. This may be evident within a single staff of music, for example, a chord consisting of a set of notes that should be played simultaneously by a pianist, or across many staves in an orchestral score.

Perhaps the most evident imposition of structure onto the temporal domain within common musical notation is the presence of bar-lines, which segment time into tangible units of *bars* or *measures*. The duration of a bar is not usually defined in absolute terms, but instead dependent on the number of *beats* it contains, the duration of which in turn is determined by the *tempo*, typically specified in beats per minute (bpm). Within the notation, bpm can be considered an objective statement of the intended rate at which musical events should unfold over time. However, this is not necessarily in direct correspondence with perceived musical speed—a high bpm can feel slow and vice-versa. Nonetheless, in the context of a score, beats (and their grouping into bars) form idealised units of time which establish a framework within which the flow of musical time can be rationalised. A primary responsibility of an orchestral conductor is to articulate the moments of time with which notated beats are to coincide. The familiar '1-2-3-4' count-in before a band begins to play provides a similar function of indicating the intended beat structure, but which again is strongly dependent on a shared understanding of the conventions associated with a particular musical style.

Continuing with temporal aspects of musical notation, the number of beats in a bar is represented by a *time signature*, resembling a fraction, for example $\frac{4}{4}$ or $\frac{6}{8}$, and is typically present at the beginning of a score and at any other position where the number of beats per bar should change. Familiarity with the conventions of time signature notation is necessary to understand their respective musical interpretations, which in turn necessitates a preliminary definition of musical *metre*. Time signatures convey basic information pertaining to the metrical organisation of music. Metre will be discussed in more detail in section 5.2, but for the present

purpose, it can be understood as concerning the succession of strong and weak *pulses* perceptible during musical listening (at least to metrically organised music), and which to some extent is evident in the grouping of notes in a score. Metre operates on a hierarchy of levels, one of which is the level of the beat or *tactus*, as discussed above. The other levels of metrical organisation relate to subdivisions of the tactus, and to larger groupings of beats. Different time signatures have different implications for the metrical organisation of events in time.

The meaning of so called *simple* time signatures can be straightforwardly derived from the signature itself. For example, $\frac{3}{4}$ and $\frac{4}{4}$ are both simple time signatures where the denominator signifies that a *crotchet* or *quarter note* in the score represents the basic beat unit. The numerator signifies the number of crotchet beats per bar, either three or four. By convention, simple time signatures imply a duple subdivision of the beat. Therefore, the basic metrical organisation implied by a $\frac{3}{4}$ time signature is that of three beats, each divisible by two: $2 + 2 + 2$.

By contrast, *compound* time signatures imply triple subdivisions of the beat, and it is simply a matter of convention that these time signatures are notated differently. $\frac{6}{8}$ and $\frac{9}{8}$ are both examples of compound time signatures. However, they are not understood literally to mean 'six quaver (or eighth) notes' per bar, or 'nine quaver (or eighth) notes' per bar respectively. Instead, $\frac{6}{8}$ implies two beats, each divisible by three $(3 + 3)$, $\frac{9}{8}$ implies three groups of three $(3 + 3 + 3)$. So despite the mathematical equivalence of the fractions $\frac{3}{4}$ and $\frac{6}{8}$, the time signatures $\frac{3}{4}$ and $\frac{6}{8}$ refer to very different temporal structures. Time signatures give the performer an indication of the intended metrical organisation of a work. However, in themselves, time signatures can not characterise all possible levels of metrical organisation, for example, levels beyond the first subdivision of the beat, or larger groupings of bars, all of which may be readily perceptible to the listener.

The regularity of the abstract temporal framework imposed by bars and time signatures is also evident in the representation of note events themselves. In Western notation, different shapes have been used to represent notes of different length, based on a system of fixed duration ratios. The taxonomy of note shapes has evolved into a simple binary tree structure, where the duration of each note in the hierarchy is half that of the parent. The representation of ternary proportions is achieved by placing a dot after the note head, indicating that the duration should equal one and a half times the usual length. Throughout the history of notation, different symbols have been used to denote the beat. Medieval notation typically took the breve as indicating the basic tactus-level pulse. Common practice music has settled on the crotchet as the primary indicator of the tactus. The crotchet can

roughly be taken to be the mid-point in the range of standard symbols, optimising the range of expression between proportionally very short and very long notes. A pragmatic consequence of this is that subdivisions of the beat (i.e. quavers, semiquavers, and so on) can be joined together by beams, making the beat structure of the music clearly evident in the notation. Beaming together notes which are shorter in duration than the beat, implies that the sum of their individual durations should be equal to the duration of the beat. This intuition is generally upheld by the use of rests. If the sum of the durations of the shorter notes is less than the duration of the beat, then appropriate rest symbols are typically inserted to account for the difference, and to also make it explicit on which subdivision of a beat a particular note should occur.

A musical score is best thought of as a set of instructions for a performer to use in order to realise an instance of the work. A score typically provides a very high level description of the pitches and rhythms that make up a piece, together with varying amounts of information concerning dynamics, phrasing and timbre. Many of the subtitles of music are not precisely specified in the notation, or even specified at all, and must be created by the performer based upon their musical knowledge and experience. However, the pragmatic uses of notation have contributed to the evolution of a system that balances the conflicting demands of precision and clarity, and is able to provide an efficient means of communicating musical ideas. As such, notated music can provide a reasonable approximation of many of the aspects of music that are salient in musical experience, particularly when the analysis of notated music is informed by evidence from music psychology.

## 5.2 Metre as entrainment

It is necessary to draw a distinction between the different timescales that operate within music (Clarke 1987). Relationships across time may be comprehended on all levels of musical organisation, from time intervals between events lasting tens of milliseconds, to relationships between patterns of notes spanning entire works. A line between rhythm and form is usually stated as the extent of the perceptual present, which is approximately up to 10 seconds in duration (Fraisse 1978; Clarke 1999). The comprehension of form is considered to require deliberate cognitive effort involving long-term memory. Below we consider only concepts that are bounded by the temporal extent of the perceptual present, reserving larger-scale musical concepts for future research.

The work of London (2004; 2012) serves as our primary reference regarding

musical metre, and provides the basis for the computational models developed within this chapter. London provides a detailed and perceptually-motivated theory of musical metre, drawing together a range of research from music theory, musicology, psychology, and neuroscience. Importantly, the theory offers considerable generality as a result of its foundation upon basic human perceptual and physiological constraints, and provides many examples from both Western and non-Western musical traditions. The experience of music as a whole is greatly dependent on cultural context, and as such can radically differ between cultures, and even between individuals within cultures. However, considering music from the perspective of basic perceptual processes, such as the experience of periodicity, which is strongly grounded in our everyday experiences in the world, London argues that commonality across many musical practices can be found. The computational theory introduced below similarly concentrates on low-level musical concepts, addressing some of what might be considered as primitives of musical conceptualisation. This section reviews London's theory together with key supporting arguments and empirical evidence.

### 5.2.1 Rhythm versus metre

Steedman (1977, p. 555) defines metre as 'regular temporal structure'. A common distinction made in the literature is that between musical metre and rhythm, although there is debate over the extent to which they can be treated independently (Benjamin 1984; Cooper and Meyer 1960; Hasty 1997). Metre can be thought of as the grouping of perceived beats or pulses into categories, which is typically expressed as the 'regular alternation of strong and weak beats' (Lerdahl and Jackendoff 1983, p. 12). In order to clarify the meaning of "grouping", Parncutt (1994) defines two kinds of grouping structure: *periodic* and *serial*. Periodic and serial groupings represent qualitatively distinct sensations arising from a musical stimulus. Periodic grouping depends on the 'relative timing and perceptual properties of *nonadjacent* events' (Parncutt 1994, p. 412), whereas '[s]erial grouping depends primarily on the serial proximity in time, pitch, and timbre of temporally adjacent events' (Parncutt 1994, p. 412). Metre can also be thought of as concerning durationless points in time, whereas serial grouping inherently concerns the relationships between events of specific duration (Clarke 1999, p. 478).

London (2004, p. 4) defines rhythm as involving 'patterns of duration that are phenomenally present in the music'. Duration here refers not to note lengths, but to the *inter-onset interval* (IOI) between successive notes. Rhythm therefore refers to the arrangement of events in time, and in that sense can be considered as

something that exists in the world and is directly available to our sensory system.

The perceptual counterpart to rhythm is metre:

> [M]etre involves our initial perception as well as subsequent anticipation of a series of beats that we abstract from the rhythmic surface of the music as it unfolds in time. In psychological terms, rhythm involves the structure of the temporal stimulus, while metre involves our perception and cognition of such stimulus. (London 2004, p. 4)

The experience of metre can, therefore, be considered as a process of categorical perception, where the surface details of the temporal stimuli, such as the particular structure of the rhythmic pattern, or any expressive performance timing, are perceived with reference to a hierarchical organisation of regular beats. The sensation of metre is induced from a stimulus in conjunction with both innate and learned responses to periodic or quasi-periodic stimuli. Extending the notion of categorical perception, London argues that metre is a form of entrainment, that is a 'coupled oscillation or resonance' (London 2012).

To better grasp the essence of what it means to consider metre as a process of entrainment, it is useful to consider how such a process is integrated with, and beneficial to, perception in general. A vast amount of information from our environment is continuously available to our perceptual system, yet our attentional resources are finite (Kahneman 1973). Therefore, in order to make sense of the world we must be able to process sensory information efficiently, and crucially to be able to detect what is most important and deserving of attention. On the one hand, in the information theoretic sense our environment is incredibly noisy. We are continually bombarded with a whole range of information, which we must somehow process in order to make accurate predications about the behaviour of objects in the environment, as well as the effect of our own actions within the environment. Yet the world is also highly structured, and from an ecological perspective, a sensitivity to such structure is expected given the view that our perception and behaviour are largely determined by the kinds of perception and behaviour afforded by the environment (Gibson 1966). Certain patterns of events in the world afford entrainment if there is a perceptible regularity to their occurrence. Regularity or periodicity is therefore an invariant quality in perception. In the musical context, this notion can be seen reflected in the language used to describe the nature of strictly periodic rhythms, which are sometimes referred to as "static" or "stationary" rhythms. More generally, the perception of invariance in the natural world is highly suggestive of intentional behaviour, whether that might be the distinctive footfalls of a predator or potential mate. As well as the individual instincts

of self-preservation and procreation, an argument for the evolutionary adaptive quality of entrainment can also be made in terms of social interaction and cohesion. In order to interact and importantly to co-operate successfully in the world, humans must be able to synchronise movement. Synchronisation requires accurate temporal prediction in order to engage the necessary motor control prior to an anticipated timepoint: successful co-ordination cannot be based on reactivity (Trevarthen 1999-2000; Clayton et al. 2005). Crucially, the perception of temporal invariance and our capacity for entrainment allow attentional resources to be directed towards likely salient moments of time, and to therefore to better predict events in the world and act accordingly.

Returning to the musical context, London's view of metre is thus a form of sensorimotor entrainment, afforded by the temporal invariances commonly present in musical structure. For listeners, this is one mechanism by which attentional resources can be directed towards predicted salient timepoints in order to efficiently process a complex auditory stimulus. For musicians, and indeed any form of movement associated with musical stimuli, entrainment is further necessary to co-ordinate physical actions.

London (2012) provides further empirical support for his theory of metre as entrainment from recent advances in neuroscience, which shed light on the underlying biological mechanism of rhythmic perception. Neuroimaging studies have discovered patterns of neuronal activity sympathetic with metrical entrainment, providing convincing evidence that metrical perception is both stimulus driven and endogenous. Differing EEG responses to trains of identical pulses are reported by Brochard et al. (2003) and Schaefer et al. (2011) as evidence for subjective metricisation. Snyder and Large (2005) and Iversen et al. (2009) both present findings that lend support to endogenous neural responses correlating with accents that are only loosely coupled with external stimuli, and in the later study it is also demonstrated that the priming of an endogenous metre has a predictable effect on subsequent auditory responses.

The degree to which listeners are able to induce a sense of metre from a rhythmic surface has also been shown to strongly affect ability in reliably processing rhythmic information (Grube and Griffiths 2009). Where a stronger sense of metre is induced, participants were able to more accurately detect rhythmic deviations. In the same experiment, the authors also provided evidence suggesting the importance of closure at the endings of rhythmic stimuli in order for listeners to report a stronger sense of perceived rhythmicality. Open endings were shown to leave listeners feeling uncertain about the structure of rhythmic stimuli, demonstrating

how the ends of sequences can influence the perception of the whole.

Composers have long exploited our capacity to maintain a metrical context, which is possible even in the presence of conflicting musical stimuli. Syncopation is the intentional rhythmic articulation of less salient metrical timepoints, which in itself is evidence for our strong tendency for entrainment, since if we could not independently maintain a sense of metre the concept of "off-beat" would be meaningless. The notion of a continuous oscillation in attentional energy provides an account, importantly one with an empirically grounded underlying mechanism, of the commonly held view that metre concerns regular patterns of strong and weak beats.

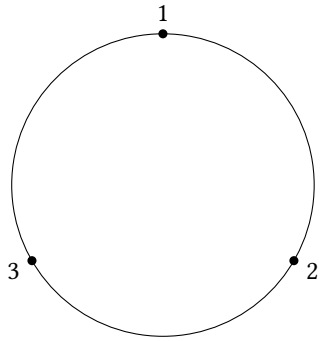
### 5.2.2 London's representation of metre

London provides the following definition of metre:

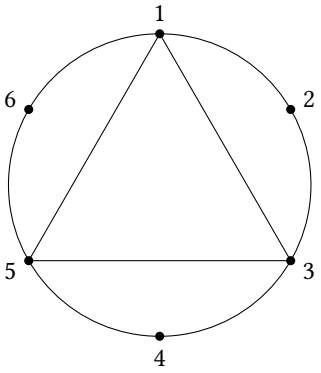
> A meter is a coordinated set of periodic temporal cycles made up of at least two peaks of sensorimotor attention. (London 2012)

The periodic nature of metre means that it can be represented graphically in the form of a circle. Following the convention developed by London (2004, pp. 64–69), time flows clockwise, and the dots on the circumference of the circle mark beats, or pulses—defined as peaks of *attentional energy*.[20] Distance on the circumference of the circle represents the relative time interval between pulses. The 12:00 position represents the downbeat—the beat marking the beginning of the cycle, and which is typically associated with the greatest concentration of attentional energy. In figure 5.1a, two levels of metrical organisation are present: the total time-span of the metre is represented by the interval around the circumference of the circle from the downbeat; and the cycle composed of the three arcs between each successive pair of dots represents a cycle of three beats. Figure 5.1b depicts a metrical structure comprising three periodic cycles: the circumference again representing the total time-span of the metre; the cycle consisting of the lines connecting beats 1, 3, and 5 forming a cycle of three beats; and the cycle consisting of the arcs between each successive pair of dots forming a cycle of six beats. Both figures 5.1a and 5.1b are labelled as representing possible metrical structures implied by a $\frac{3}{4}$ time-signature. Therefore, in terms of common music notation, the three-beat cycle in each case corresponds to the crotchet-level beat, and the six-beat cycle in the second figure corresponds to the duple subdivision of the crotchet beat into quavers.

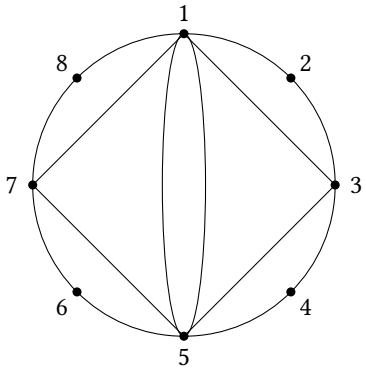
[20] The use of italic font in this section denotes terms defined in London's theory



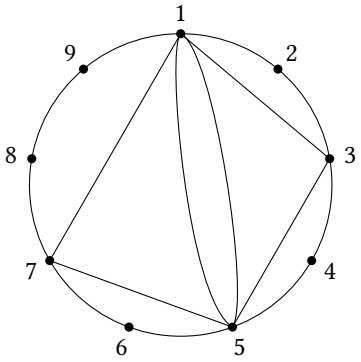

a: $\frac{3}{4}$ with no beat subdivision.




b: $\frac{3}{4}$ with quaver beat subdivision.




c: $\frac{4}{4}$ with quaver beat subdivision.




d: Non-isochronous 2-2-2-3 N-cycle grouping pattern.

Figure 5.1: Cyclical representation of metre, after London (2004, pp. 64–69).

The total number of dots around the circumference of a circle defines the *cardinality* of the metre. This cyclic component, referred to as the *N-cycle*, where N equals the cardinality, is the lowest level (fastest moving) cycle in any metrical hierarchy and includes all peaks of attentional energy. The concept of the N-cycle can be used as the basis for describing metrical structure. Figure 5.1a is therefore a 3-cycle metre, which in this case is an N-cycle component that also corresponds to the *beat cycle* or tactus. Figure 5.1b is a 6-cycle, which in this case corresponds to a subdivision of the beat-cycle. It is more common for a beat-cycle to be a *subcycle* of an N-cycle, because most metres include at least one level of tactus subdivision (London 2004, p. 35). Thus subcycles represent higher levels in the metrical hierarchy, involving longer IOIs.

Individual metrical structures, or *metrical types*, can be distinguished based on the cardinality of their N-cycle, together with the particular grouping of N-cycle beats into subcycles. Therefore, as a metrical type, figure 5.1b can be described as a

6-cycle, with a 3-cycle component comprising timepoints 1-3-5. Metres may have further levels of organisation, as shown in figure 5.1c, which contains four levels of periodic motion, and may also include non-isochronous cycles (*NI-cycles*), as in figure 5.1d, where the first subcycle consists of three short beats followed by one long beat.

Timing information is needed to specify a metre fully within this framework, giving rise to metrical types that can be individuated both in terms of tempo, as well as expressive variation due to pulse interval micro-timing variation. Drawing on the psychological literature, London defines the maximum period of the N-cycle as between 5 and 6 seconds, and the IOI between timepoints on the N-cycle as at least 100 milliseconds. This range defines the *temporal envelope* for metre (London 2004, p. 27). The maximal range of tactus IOI is from 200 ms to 2000 ms (corresponding to 30–300 bpm), with the typical range being 400–1200 ms (50–150 bpm), with a preference around 600 ms (100 bpm). The perceived character of individual metrical types changes across the range of the temporal envelope, resulting in identifiable *tempo-metrical types* (London 2004, pp. 73–76). Note that these limits are not arbitrary constraints, defined in order to simplify the representation of metre: they are empirically determined perceptual and cognitive limitations. Such quantitative understanding of perceptual phenomena is central to the present purpose of constructing perceptually valid conceptual space representations.

In addition to the above constraints on the N-cycle, London provides further constraints that apply to the internal subcycle structure of metre. In the following, London's metrical well-formedness constraints (in their revised form) are presented in full (London 2012).

**Perceptual Constraints on Levels and Cycles**

**WFC 1.1** IOIs between attentional peaks on the N-cycle must be greater than $\approx$ 100 ms.

**WFC 1.2** The beat cycle involves those attentional peaks whose IOIs fall between $\approx$ 400 ms and $\approx$ 1200 ms.

**WFC 1.3** A meter may have only one beat cycle.

**WFC 1.4** The maximum duration for any or all cycles is $\approx$ 5000 ms.

**Minimal Requirements**

**WFC 2.1** A meter must have a beat cycle.

**WFC 2.2** The beat cycle must involve at least two beats.

**WFC 2.3** The N-cycle may serve as the beat cycle.

**Intercycle Relationships**

**WFC 3.1** All cycles must have the same total period/duration.

**WFC 3.2** All cycles must be continuous.

**WFC 3.3** The N-cycle and all subcycles must begin and end at the same temporal location; that is, they must all be in phase

**WFC 3.4** Each subcycle must connect nonadjacent time points on the next lowest cycle.

**Regularity Requirements**

**WFC 4.1.1** If the IOIs on the N-cycle are non-isochronous, then the IOIs on the beat cycle must be nominally isochronous (i.e., categorically equivalent, though subject to expressive variation).

**WFC 4.1.2** If the IOIs on the N-cycle are non-isochronous, their absolute lengths must be such as to avoid ambiguities and contradictions; $(S)$ must be $> \frac{1}{2}(L)$.[21]

**WFC 4.1.3** Sequencing of NI elements on the N-cycle will remain constant from beat to beat within the cycle, maintaining maximal evenness.

**WFC 4.2.1** If the IOIs on the N-cycle are isochronous, then the beat cycle need not be.

**WFC 4.2.2** If the beat cycle is NI, then either (1) it is maximally even or (2) the cycle above the beat cycle, in most cases the half-measure cycle, must be maximally even.

The above constraints on metric well-formedness define a large space of possible metres, which London is confident encompasses the vast majority of metres present across all musical cultures (London 2004, p. 114). Importantly, the constraints allow us to exclude from consideration the much larger space of all possible hierarchical cyclic structures that do *not* correspond with a subjective experience of metre.

### 5.2.3 Prototypical and individuated metre

> We do not encounter "generic $\frac{4}{4}$" or even "$\frac{4}{4}$ at a tempo of quarter-note = 120" but a pattern of timing and dynamics that is particular to the piece, the performer, and the musical style. Therefore, to give an ecologically valid account of meter, we must move beyond a theory of tempo-metrical types to a metrical representation that involves particular timing relationships and their absolute values in a hierarchically

[21] (S) and (L) refer to categorically short and long beats respectively.

related set of metric cycles. (London 2004, p. 159)

The absolute value of timing relationships is here understood as referring to the individual intervals between the timepoints on the various cyclic levels of metrical organisation, and dynamics as the relative strength of attentional energy associated with each metrical category. London concentrates primarily on aspects of timing in the development of the theory. However, both timing and dynamics are considered equally important from the perspective of the models developed below. Within London's theory, the concept of metre now becomes something much more fluid and nuanced: the subtle and regular, with respect to the granularity of categorical perception, differences in timing between individual metrical timepoints which may be characteristic of a musical style or individual performer, are acknowledged to be part of metrical perception, rather than being considered as expressive timing deviations from an abstract music-theoretic notion of metre. Therefore, tempo-metrical types, together with particular internal timing relationships, should be understood as distinct metres in their own right. London calls this expanded view of metre the 'Many Metres Hypothesis' (London 2004, pp. 142–160), and argues that this theory offers a more parsimonious and ecologically sound account of musical expressive timing than traditional Western music theoretic notions of metre.

Similar arguments for an expanded definition of metre have also been made by researchers working with music from outside the Western musical mainstream. Polak (2010) writing on jembe music from Mali states that cyclic subpulse variation within metric types is 'inherent in the metric system', and an integral component of the rhythmic feel of the music – or as Polak rightly points out, what should more accurately be termed *metric* feel. Polak suggests that it may be 'more universally valid [...] to conceive of expressive timing as variation from metrical expectation' (Polak 2010, ¶ 149), echoing the adoption of a cognitive perspective in musical understanding as taken in this thesis.

Further argument, from both musicological and empirical perspectives, in support of an enriched definition of metre is provided by Tellef Kvifte. Kvifte's work is particularly important to consider here because it offers an alternative formulation of a key aspect of metrical well-formedness as proposed by London (2004), and one which was arrived at independently by the present author during the process of formalising and implementing London's constraints.[22] In Kvifte (2007), several prominent theories of metre are surveyed, including that of London, all of which

[22]In the revised edition of the theory (London 2012), non-isochronous N-cycles are permitted into London's WFCs.

can be can be viewed as Common Fast Pulse (CFP) models. In the terminology adopted here, the CFP is equivalent to the N-cycle, which, to reiterate, London (2004) defines as notionally isochronous, and the primary constraint on the hierarchical structure of slower grouping cycles. Kvifte is primarily concerned with non-isochronous metre, and acknowledges the utility of assuming isochrony at the fastest level of metrical organisation as a possible account of entrainment and of the subjective experience of constant tempo. Furthermore, Kvifte acknowledges from experience the importance of counting a constant fast pulse when learning to perform non-isochronous music. However, it is pointed out that this must be overcome, and the non-isochrony internalised, if one is to be able to perform well—something that seems to be intuitively mastered by non-musicians enculturated with metrically non-isochronous music. From an empirical perspective, an analysis of Norwegian *springar* dance music revealed that the range of tactus subdivision was between 15% and 85% (Kvifte 2004, p. 70), which supports the 'practically unanimous' view in the field that springar music does not as a rule have a common fastest pulse (Kvifte 2007, p. 73). Furthermore, the analysis revealed that the process of non-equal subdivision can occur at multiple metrical levels, in this case at both the tactus and sub-tactus levels (analysis was constrained to three metrical levels: the tactus, one grouping and one subdivision level).

Kvifte emphasises the necessity of keeping distinct questions of metrical categories and metrical timing. Regarding the perception of pulse interval categories, Kvifte agrees that London's model gives a plausible account. However, the point of disagreement concerns London's association of metrical timing with the concept of the isochronous N-cycle. Kvifte instead proposes a Common Slow Pulse (CSP) approach to formalising metrical structure, motivated in part by motor-mimetic theories of music cognition (Godøy 2003). In a CSP model, an isochronous higher-level grouping cycle is identified as the primary determinant of metrical macrotiming. Above this cycle, here termed the *anchor-cycle*, pulse intervals of further grouping cycles are simply multiples of anchor-cycle intervals, and perceived primarily additively. Whereas cycles below the anchor-cycle are formed by subdividing anchor-level pulse intervals into not necessarily equal intervals, and are experienced primary divisively. This model of metrical structure is very similar to that proposed by Lerdahl and Jackendoff (1983), except that the anchor-cycle is not necessarily required to also be the tactus cycle. Admitting non-equal metrical subdivisions into a formal definition of metre allows for microrhythmic details to be considered as part of metrical structure, rather than either being left unaccounted for, or potentially misleadingly being considered as random or stylistic deviation

from an abstract Western-centric theoretical conceptualisation of metre.

The inclusion of pulse interval timing variation into the concept of metre does not negate the notion of the categorical perception of time intervals. However, it does raise potential challenges to the formulation of well-formedness constraints. Typically, broad categories of equal, long and short intervals will remain distinct as a consequence of the well-formedness of hierarchical metrical grouping, irrespective of variation in absolute timings. However, as timing variations become more extreme, a particular class of metrical ambiguity impossible in London's theory becomes admissible. For example, going beyond a 1:2 subdivision ratio potentially leads to larger intervals being present on further subdividing cycles relative to the immediate grouping cycle. If the ratio is exactly 1:2, (say 200:400 ms subdivisions of a 600 ms tactus), then further equal duple subdivisions lead to identical pulse intervals being present on adjacent cycles. As Kvifte (2004) argues that metres with such extreme subdivision ratios are meaningful to those familiar with the culture of springar music, further psychological testing and formal work is required. This will not be addressed specifically here, except it suffices to say that the conceptual spaces below are capable of representing such extreme non-isochronous metres, and remain agnostic as to the precise boundaries between regions corresponding to well-formed and non-well-formed structures.

Differences in the constraints of metrical timing aside,[23] each of the above accounts argues for an understanding of metre that goes beyond abstract or idealised periodic temporal patterns. Within the formalism below, no special account is made for idealised metres. Regarding the representation of metrical structure, the formalism takes full account of expressively articulated metres: idealised metrical structures are simply structures with no variation in timing or attentional energy, and are considered prototypical of real-world patterns of entrainment. London refers to our ability to discern nuanced distinctions between familiar patterns of entrainment through exposure as a process of *individuation*.

> A listener's metric competence resides in her or his knowledge of a very large number of context-specific metrical timing patterns. The number and degree of individuation among these patterns increases with age, training, and degree of metrical enculturation.
>
> (London 2004, p. 153)

In the theory of conceptual space, Gärdenfors argues that the function of quality dimensions is to represent how concepts, grounded in perception, can vary.

[23] The revised edition of London's theory is in fact broadly compatible with the theories proposed by Polak and Kvifte.

Therefore, taking the cue from the term individuation, quality dimensions must be sought which can represent all the possible ways in which prototypical metrical concepts can be perceptually differentiated. Representing individual timing relationships and patterns of attentional energy will necessarily require conceptual spaces of higher dimensionality than those restricted to metrical prototypes as presented by Forth et al. (2010). The idea of acquiring new dimensions in order to account for differences in novel stimuli sits very naturally with the proposed geometrical model of learning and concept formation. In the spaces defined below, prototype metres are simply the centroid of regions corresponding to particular metrical types.

In the following, a symbolic formalisation of London's theory is first developed, from which a number of conceptual space representation are derived.

## 5.3 A symbolic definition of London's theory

Before defining a geometrical representation of metrical structure, it is helpful to first describe London's theory of metre formally in terms of a tree representation. This level of representation corresponds to a symbolic representation in Gärdenfors' terminology. The subsequent geometrical models of metre can then be defined as mappings of tree structures to points in conceptual spaces, and thus simplifying the definition of the geometrical representations. Furthermore, not only are tree structures intuitive ways of thinking about metrical structure, they also simplify the relationship between concrete sequences of musical events and geometrical representations of metrical concepts. Therefore, the combination of symbolic and conceptual representations are complementary. On the conceptual level, it is possible to conceptualise metre in terms of space, invoking notions of similarity in terms of distance, and types of metres in terms of geometrical regions. However, the geometry itself is not necessarily intuitive, particularly for higher-dimensional spaces. Therefore, the symbolic representation provides a means of understanding the properties of the geometrical spaces in more familiar terms.

### 5.3.1 Representational semantics of metrical trees

Tree structures offer a discrete view of musical metre, somewhat at odds with London's cyclic representation that aims to convey the continuous nature of metrical entrainment (London 2004, p. 65). "Continuous" here has two interpretations. Firstly, in a loosely mathematical sense, a model may be considered continuous if it is defined over the set of real numbers, and in effect an infinite number of ever

finely distinguished values may be used. Secondly, continuous may imply existing through time, in other words, a dynamic process. London's cyclic representations informally convey both of these qualities. The placement of N-cycle points on the circumference of the circle is not constrained to a finite set of positions, notwithstanding the constraints of metrical well-formedness. Further, the cyclic nature of the diagrams at least implies the notion of metre as a continuous process of entrainment happening over time.

Tree representations offer a more abstracted view of musical metre compared with London's cyclic representation. However, the difference is only a matter of representation, and does not constrain the theoretical understanding of metre as a dynamic process of entrainment. It will be shown that metrical trees can be mapped to points in geometrical spaces, which are themselves continuous representations in the mathematical sense. The converse is also true: all points in a conceptual space can be mapped to a tree. Whether all points and respective trees correspond to well-formed metres is another matter—but constraints can be specified either geometrically or symbolically as appropriate (since both are representations of the same psychological phenomenon) in order to maintain metrical well-formedness. Therefore, trees are a discrete symbolic representation, but only so far as they are considered here to be an abstraction of a point in a continuous space of metre. Points and regions in geometrical spaces correspond to stable (possibly ambiguous) metrical concepts, and theoretically, infinitely many metres may be represented – although as a model of metrical conceptualisation, the models bottom out at a set of just-noticeable difference regions.

Representations of metrical structure, in both symbolic and geometrical spaces, encode temporal information. However, neither refer to real time. In this sense, they are not time-continuous and dynamic. Changes in metrical experience over time can be modelled by a discrete sequence of points in a conceptual space, or equivalently as a sequence of metrical trees. However, the metrical concepts themselves are still not time-continuous. As such, the representations pursued here are best thought of as representing discrete snapshots of a process of entrainment modelled over a certain time window.

A further constraint herein assumed regarding the non-dynamic nature of metrical concepts is that only a single concept will be derived for any given musical excerpt or entire piece. Therefore, metrical structure, pulse interval timing, and the distribution of attentional energy associated with a piece of music are assumed to represent the general metrical feel of the music. The process of mapping symbolic representations of musical stimuli to metrical concepts is described in section 5.6.2.

Implicit in this constraint is that the model assumes complete knowledge of a piece of music in order to represent its metre. This constraint allow questions associated with changing metre to be excluded from the current investigation. Firstly, this is necessary in order to constrain the evaluation of metrical-rhythmic similarity to stable, bounded concepts, and avoid higher-order *sequences* of metrical concepts. Secondly, addressing metrical dynamics is very difficult without first specifying a process of metrical induction, so that it becomes possible to estimate the most likely metrical interpretation of a piece of music through time. These issues are beyond the scope of the current work, but the overall framework does not preclude their development. A further level of representation, corresponding to Gärdenfors' sub-symbolic level, may be able to represent the in-time experience of metrical entrainment, and will be the subject of future work.

### 5.3.2 Notation and definitions of trees

**Definition 5.1.** A *undirected labelled graph* $\mathcal{G} = (V, E, L)$ consists of a finite non-empty set of vertices or nodes $V$, a finite set of edges $E$, and a set of vertex labels $L$. An edge is a set $\{u, v\}$, where $u, v \in V$ and $u \neq v$.

**Definition 5.2.** Vertices $u, v \in V$ are *adjacent* if $\{u, v\} \in E$.

**Definition 5.3.** A *simple path* is a sequence of adjacent vertices, where no vertex is included more than once.

**Definition 5.4.** An undirected graph is *connected* if there exists a simple path between each pair of vertices $u, v \in V$.

**Definition 5.5.** A connected undirected graph containing no cycles is a *tree*, for which $|E| = |V| - 1$.

**Definition 5.6.** Let $\mathcal{G} = (V, E, L)$ be a labelled tree. Each node has a label, possibly empty. The function *label* : $V \rightarrow L$ returns the corresponding label $l \in L$ for a vertex $v \in V$.

**Definition 5.7.** A rooted tree is a tree in which there is a distinguished node $r \in V$, called the *root* of the tree, denoted *root*$(\mathcal{G}) = r$. There is a simple path from $r$ to every node in a tree.

**Definition 5.8.** The length of the path from the root $r$ to a node $v \in V$ is the *depth* of $v$ in $\mathcal{G}$, denoted *depth* : $V \rightarrow \mathbb{N}$. Thus the root node always has a depth of 0.

**Definition 5.9.** A *level* of a tree consists of all nodes of the same depth:
*level*$(d, V) = \{v \in V \mid$ *depth*$(v) = d\}$.

**Definition 5.10.** The *cardinality* of a level of a tree is the number of nodes at any given level: $\textit{cardinality}(d, V) = |\textit{level}(d, V)|$.

**Definition 5.11.** Any node $u \in V$ on the unique simple path from the root $r$ to $v \in V$ is an *ancestor* of $v$. If $u$ is an ancestor of $v$, then $v$ is a *descendent* of $u$. All nodes are both ancestors and descendents of themselves.

**Definition 5.12.** Node $v \in V$ is said to be a *child* of node $u \in V$ if $u$ and $v$ are adjacent and $\textit{depth}(v) = \textit{depth}(u) + 1$. In which case $u$ is the *parent* of $v$. The function $\textit{parent} : V \to V$ returns the parent given any node $v \in V$. The root $r$ is the only node without a parent, denoted $\textit{parent}(r) = \top$.

**Definition 5.13.** The set of child nodes for a given node $u \in V$ is given by $\textit{children}(u) = \{v \in V \mid \textit{parent}(v) = u\}$

**Definition 5.14.** The *degree* or *arity* of a node $u \in V$ is equal to the number of its child nodes.

**Definition 5.15.** Nodes with no child nodes, and therefore degree 0, are called *leaf nodes*.

**Definition 5.16.** All non-leaf nodes are called *internal nodes*.

**Definition 5.17.** The Boolean valued function $\textit{is-leaf} : V \to \{0, 1\}$ returns 0 (false) for all non-leaf nodes, and 1 (true) for leaf nodes. Thus, the set of leaf nodes can be denoted $\textit{leaf-nodes}(V) = \{v \in V \mid \textit{is-leaf}(v) = 1\}$.

**Definition 5.18.** The *height* of a tree is equal to the length of the longest simple path from the root $r$ to a leaf node.

**Definition 5.19.** A *perfect tree* is a tree in which every path from the root $r$ to a leaf is of equal length.

**Definition 5.20.** An *ordered tree* is a tree in which a total ordering is defined for the children of each node.

**Definition 5.21.** Given an ordered tree, child nodes can be identified in terms of their *position* relative to their parent, with respect to the defined node ordering. Node position is indexed from zero. The function $\textit{child} : \mathbb{N} \times V \to V \cup \top$ takes an index $i$ together with a node $u \in V$, and returns the $i$th child node of $u$ with respect to the node order, or undefined, $\top$, if $i$ is equal to or greater than the degree of $u$.

**Definition 5.22.** Given a node, the function $\textit{position} : V \to I \cup \top$ returns the position of the node relative to its parent, with respect to the defined node ordering, or undefined, $\top$, for the root node.

### 5.3.3 Definition of tempo-metrical trees

Given the formal definitions concerning tree structure stated in section 5.3.2, it is now possible to define instances of individuated metrical types symbolically in the form of a tree.

**Definition 5.23.** A metrical tree, $\mathcal{T}$, is a labelled, perfect, ordered tree: $\mathcal{T} = (V, E, L, O, I, A, t)$.

The sets $V$, $E$ and $L$ correspond to the above defined sets of nodes, edges and labels respectively. Each node represents a pulse in a metrical hierarchy—a peak of attentional energy—and each edge represents the hierarchical relationship between pulses. In metre-theoretic terms, each set of nodes of equal depth, that is each level of the tree, represents the pulses present within a particular metrical cycle. The measure period corresponds to level zero, that is the single element set equal to $\{r\}$. The N-cycle corresponds to the level of maximum depth, equivalent to the set of leaf nodes.

$O$ is a set of absolute timepoints. Each node is associated with a timepoint, representing the time instant of a peak of attentional energy, allowing temporal information to be encoded explicitly within a tree. The timepoint of a peak of attentional energy will be referred to as the pulse *onset*. $I$ is a set of pulse IOI values, and again each node is associated with an IOI. The IOI value associated with a node represents the time interval between the onset of the pulse the node itself represents, and the onset of the next pulse in the same metrical level. The set $A$ is a set of attentional energy values. Each leaf node is associated with a value from this set, allowing the degree of attentional energy associated with each N-cycle pulse to be encoded within a tree. The element $t$ is an integer denoting the level of the tree corresponding to the tactus cycle. Sets $O$, $I$ and $A$ are formally defined below, along with functions and constraints required for the representation of well-formed metrical structure. The node labelling procedure is then defined, which is necessary for subsequent projection of $\mathcal{T}$ into geometrical space. The variable $t$ is then discussed in section 5.3.4.

#### Onset

Elements $o \in O$ represent absolute timings of pulse onsets within a metrical structure. Time is measured relative to the downbeat in units of milliseconds.

**Definition 5.24.** The set $O$ is a set of absolute *pulse onset* timepoints, $O = \{x \in \mathbb{R} \mid 0 \leq x \leq 5900\}$. Each node $v \in V$ is associated with exactly one

element from $O$, and multiple nodes can be associated with the same timepoint. The function *onset* : $V \rightarrow O$ returns the associated onset $o \in O$ for a node $v \in V$.

As discussed in section 5.3.1, we hold the view that metre is a dynamic process of entrainment existing through time, and that our representation is a snapshot of this process modelled over some time window. It is convenient from the point of view of formalisation to consider the internal structure of a metrical concept with reference to the real-number line, and we define the onset of the root node $r$ to be equal to timepoint zero.

$$\textit{onset}(\textit{root}(\mathcal{T})) = 0 \tag{5.1}$$

The association of nodes with timepoints should not be confused with the real-time duration over which a stable concept of metre may be experienced.

The upper limit for values in $O$ is determined by the temporal envelope of metre, that is the minimum and maximum time intervals that afford metrical entrainment. London (2004) states the range of metrical pulse IOI as between 100 ms to approximately 5 or 6 s (WFC 1.1 and WFC 1.4, p. 101). The upper limit of 6 s is used here to ensure the representation covers the greatest possible range of plausible metres, while acknowledging the possibility that not all permissible structures may necessarily be representations of valid metrical entrainment for all or even any listeners. The upper limit for onset values is therefore 5900 ms, since no onset may be greater than this in order for the minimum and maximum pulse IOI limits to be respected.

The association of each node in $\mathcal{T}$ with a timepoint provides a convenient and intuitive means by which to define node ordering.

$$u < v \text{ iff } \textit{onset}(u) < \textit{onset}(v), \quad \forall u, v \in V \tag{5.2}$$

Defining node order in terms of pulse onset provides a total ordering over all child nodes because pulses within the same metrical cycle must be separated in time by at least 100 ms. The availability of a total ordering over child nodes will be used subsequently to define node labels, giving a total ordering over all nodes in $\mathcal{T}$, which is necessary for projection into geometric space.

London's graphical representation of metre, discussed in section 5.2.2, requires that all peaks of attentional energy be present on the N-cycle, with subcycles forming higher-level grouping over this sequence of pulses. To ensure that tree representations of metre conform to this structure, the following constraint requires

that the onset of every internal node be equal to the onset of its first child node.

$$\textit{onset}(v) = \textit{onset}(\textit{child}(0, v)), \quad \forall v \in V \setminus \textit{leaf-nodes}(V) \tag{5.3}$$

Nodes representing pulses at multiple levels of $\mathcal{T}$ that are also associated with the same timepoint are said to be *temporally coincidental*. Therefore, the root node $r$ together with all temporally coincidental nodes collectively represent the metrical downbeat, and are associated with timepoint zero.

**Pulse IOI**

Elements $i \in I$ represent time intervals between pulse onsets, specifically the time interval between one pulse onset and the next pulse onset represented at the same metrical level. Time is measured in units of milliseconds.

**Definition 5.25.** The set $I$ is a set of *pulse IOI* values, $I = \{x \in \mathbb{R} \mid 100 \leq x \leq 6000\}$. Each node $v \in V$ is associated with exactly one element from $I$, and multiple nodes can be associated with the same time interval. The function $\textit{p-ioi} : V \to I$ returns the corresponding pulse IOI $i \in I$ for a node $v \in V$.

As discussed in the previous definition of $O$, values in $I$ correspond to the range of the temporal envelope of metre. The association of each node with a pulse IOI is intended to represent the anticipatory nature of metrical entrainment: the expectation associated with a peak of attentional energy toward the occurrence of the next. The pulse IOI associated with the final node of any level of $\mathcal{T}$ represents the time interval towards the next downbeat, emphasising the cyclic nature of metrical experience.

We define the sum of the pulse IOIs of all child nodes to be equal to the pulse IOI of their parent.

$$\textit{p-ioi}(v) = \sum_{w \in \textit{children}(v)} \textit{p-ioi}(w), \quad \forall v \in V \setminus \textit{leaf-nodes}(V) \tag{5.4}$$

It is therefore trivially true that the sum of all pulse intervals on each level of $\mathcal{T}$ are equal. The above constraint, together with constraints (5.1) and (5.3) satisfy the following three of London's requirements of metrical well-formedness: WFC 3.1 that all cycles must have the same total duration; WFC 3.2 that all cycles must be continuous; and WFC 3.3 that all cycles must be in phase (p. 102).

**Attentional energy**

Elements $a \in A$ represent the degree of attentional energy associated with each N-cycle timepoint.

**Definition 5.26.** The set $A$ is a set of *attentional energy* values, $A = \{x \in \mathbb{R} \mid 0 \leq x \leq 1\}$. Each leaf node $v \in$ *leaf-nodes*$(V)$ is associated with exactly one element from $A$, and multiple nodes can be associated with the same attentional energy value. The attentional energy value associated with a leaf node can be obtained directly from a leaf node, or from any temporally coincidental internal node, by the function *a-energy* : $V \to A$ (5.5).

Attentional energy values are only associated with N-cycle timepoints, and therefore higher-level groupings may not take on distinguished attentional energy values. The function *a-energy*$(.)$ is defined such that the attentional energy value associated with a leaf node can be obtained given any temporally coincidental internal node. Within this definition, the function *child*$(0, .)$ returns the left-most child node of an internal node, and is called recursively until the corresponding leaf node is found.

$$\begin{aligned} \textit{a-energy} : V &\to A \\ v &\mapsto \begin{cases} a & \text{if } \textit{is-leaf}(v) = 1 \\ \textit{a-energy}(\textit{child}(0, v)) & \text{otherwise} \end{cases} \end{aligned} \tag{5.5}$$

Attentional energy values represent a hypothetical weighting of metrical pulses. In actuality, the degree of attentional energy associated with any pulse may be a combination of many factors, both endogenous and stimulus driven. Neurological studies have shown evidence that both endogenous and stimulus accents have a significant influence on patterns of neurological activation (Snyder and Large 2005; Iversen et al. 2009). Given such evidence, together with findings from cognitive scientific modelling of musical expectation (Pearce and Wiggins 2006), it is reasonable to expect that all aspects of perceived musical structure—whether concerning rhythm, melody, harmony, or timbre—may be involved in shaping patterns of attentional energy. A complete model of attentional energy is beyond the present scope, and instead we take a pragmatic approach. In the context of prototypical metres, all attentional energy values will default to 1. A reasonable alternative to this might be a GTTM-style metrical accent based on the number of timepoints across all cycles that are temporally aligned with each N-cycle timepoint (Lerdahl and Jackendoff 1983). However, further evidence should be sought before incorporating such an assumption into the model. Individuated

attentional energy values for metrical concepts derived from musical passages is simply a function of the number of events that are perceptually coincidental with each N-cycle timepoint. The mapping from symbolic representations of music to metrical concepts is discussed further in section 5.6.2.

#### Constraints on metrical tree structure

Metrical concepts are bounded in time, and therefore in the number of cycles they can contain.

**Definition 5.27.** Node depth $D = \{x \in \mathbb{N} \mid 0 \leq x \leq 5\}$ in $\mathcal{T}$ is denoted $\mathit{depth} : V \to D$.

The maximum depth of any node in $\mathcal{T}$ is thus defined as 5, because no well-formed metrical structure within the bounds of the temporal envelope of metre, can consist of more than six hierarchical cycles, as shown in (5.6). Such perceptual constraints are in no sense physical laws, and extending the range arbitrarily by a few milliseconds would admit structures of seven cycles. However, temporal factors may not be the only source of constraint on the ability to perceive greater numbers of metrical levels. Intuitively, the greater the number of levels, the more complex a metrical concept becomes, requiring greater cognitive resources. Furthermore, the greater the number of levels, the greater the number of distinct metrical categories are available with which to form a metrical interpretation of surface rhythmic patterns. From an information theoretic perspective, presumably the greater number of categories diminishes in utility beyond a certain point. Nonetheless, such questions are here left for future work, and for the purpose of formalisation, an absolute range of the metrical envelope is assumed. However, it should be noted that the conceptual space representations necessarily extend beyond the fixed range defined for metrical trees, so metrical structures beyond this range *can* be represented at the conceptual level, although they fall into regions that are beyond the typical bounds of metrical entrainment, and are therefore not considered well-formed metrical structures.

$$\begin{aligned} \arg\max_{n\in\mathbb{N}}, 100 \cdot 2^n &\leq 6000 \\ &= 5 \end{aligned} \tag{5.6}$$

The height of $\mathcal{T}$ has the usual definition of the length of the longest path from the root to a leaf, and we constrain the minimum height to one.

$$\begin{aligned} \mathit{height} : \mathcal{T} &\to D \setminus \{0\} \\ V(\mathcal{T}) &\mapsto \max_{v\in V} \{\mathit{depth}(v)\} \end{aligned} \tag{5.7}$$

The minimum height of $\mathcal{T}$ is one because London's definition of metre requires at least two levels of coordinated periodic motion, as a consequence of WFC 2.2 (p. 101) requiring that the tactus cycle contain at least two beats. In a metre comprising only two cycles, the N-cycle therefore must be the tactus, permissible by WFC 2.3, with the additional slower cycle grouping tactus beats into typically twos or threes.

Each node can have either zero, two or three child nodes, following Lerdahl and Jackendoff's metrical well-formedness rule (MWFR 3): 'At each metrical level, strong beats are spaced either two or three beats apart' (Lerdahl and Jackendoff 1983, p. 69).

$$\begin{aligned} \textit{arity} : V &\rightarrow \{0, 2, 3\} \\ v &\mapsto \begin{cases} 0 & \text{if } \textit{is-leaf}(v) = 1 \\ |\textit{children}(v)| & \text{otherwise} \end{cases} \end{aligned} \tag{5.8}$$

As defined above, nodes of arity zero are leaf nodes. All other internal nodes in $\mathcal{T}$ are required to be of arity two or three, corresponding to duple or triple metrical subdivisions respectively. For every internal node, this constraint on arity accords with London's WFC 3.4, that each subcycle must connect nonadjacent timepoints. The arity of $\mathcal{T}$ is equal to the maximum arity of its constituent nodes. Thus, $\mathcal{T}$ is always of arity two or three, depending on whether triple subdivisions are present in the metre or not.

This constraint on metrical grouping structure is more restrictive than that stated in London's theory, which allows attentional groups of larger multiples. Lerdahl and Jackendoff (1983) proposed their well-formedness rule in the context of Western classical music, and therefore it should not be assumed to be applicable across all musical cultures. Notwithstanding this caveat, in order to preserve uniformity within the present formalism, we here assume that all groupings of pulses greater than three can be modelled as hierarchical groupings of twos and threes. This concession to formality may well be refuted by empirical evidence, but that question will not be pursued further here.

#### Abstraction of sequential structure

Two further functions of $\mathcal{T}$, required for the subsequent geometrical mappings, must be defined. Each function summarises a structural feature of $\mathcal{T}$. Except for prototypical metres, all instances of $\mathcal{T}$ represent varying degrees of expressive variation, both in timing and attentional energy. To completely represent this

expressive variation requires a representation of the internal sequential structure of each cycle, to be defined in section 5.5. However, for the simpler geometrical representation of purely periodic structure, to be defined in section 5.4, an abstraction of the sequential structure of each cycle is required. Therefore, equations (5.9) and (5.10) define the mean pulse IOI and attentional energy values respectively for each cycle, in terms of node depth. The use of the mean here assumes that variation is normally distributed, and in psychological terms, that it is appropriate to consider the mean as being prototypical of a perceptual category. At least in the case of timing, there is evidence to suggest that rhythmic categorisation does not align completely with, or symmetrically about, the mean of expressively varied IOI category timings (Desain and Honing 2003; Repp et al. in press).

$$\textit{mean-p-ioi}(d, V) = \frac{\sum_{w \in \textit{level}(d,V)} \textit{p-ioi}(w)}{\textit{cardinality}(d, V)} \tag{5.9}$$

$$\textit{mean-a-energy}(d, V) = \frac{\sum_{w \in \textit{level}(d,V)} \textit{a-energy}(w)}{\textit{cardinality}(d, V)} \tag{5.10}$$

**Node labels**

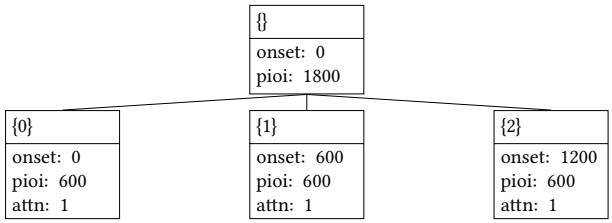
A total ordering over all nodes in $\mathcal{T}$ is required in order to define a mapping from $\mathcal{T}$ into geometrical space. This is a distinct notion to the temporal ordering of pulses as defined above in terms of pulse onset, which only provides a total ordering within individual levels of $\mathcal{T}$. A total ordering over all nodes in $\mathcal{T}$ may be arbitrary, so long as it is complete. Given that we have an existing definition of child node order based on temporal sequence, a total ordering over all nodes in $\mathcal{T}$ can simply be defined in terms of node labels generated by considering the unique path to each node from the root.

Each node $v \in V$ is labelled by a unique $n$-tuple $\textit{label}(v) \in L$, where $n = \textit{depth}(v)$, according to the following recursive function. The symbol $\|$ is used here to denote tuple concatenation.

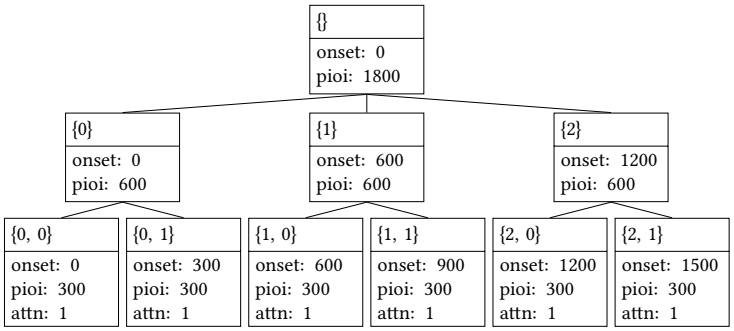
$$\textit{label}(v) = \begin{cases} \langle\rangle & \text{if } v = \textit{root}(T) \\ \textit{label}(\textit{parent}(v)) \| \langle \textit{position}(v) \rangle & \text{otherwise} \end{cases} \tag{5.11}$$

The root is thus labelled $\langle\rangle$, and child nodes of the root are labelled respectively: $\langle 0 \rangle$, $\langle 1 \rangle$, and $\langle 2 \rangle$. Labels for each subsequent child node are generated in the same manner by appending the child's position relative to its parent to the parent's label:

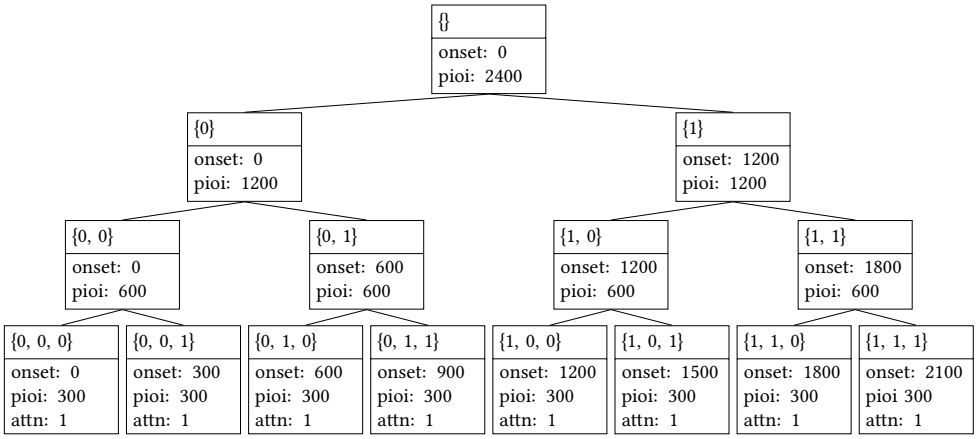
$\langle 0, 0 \rangle$ , $\langle 0, 1 \rangle$ , $\langle 0, 2 \rangle$ . . . $\langle 0, 0, 0 \rangle$ , $\langle 0, 0, 1 \rangle$ , $\langle 0, 0, 2 \rangle$ etc. Tree representations corresponding to instances of the metres in figure 5.1 are shown in figure 5.2. The total ordering over $\mathcal{T}$ is then defined as the lexicographic ordering of node labels.



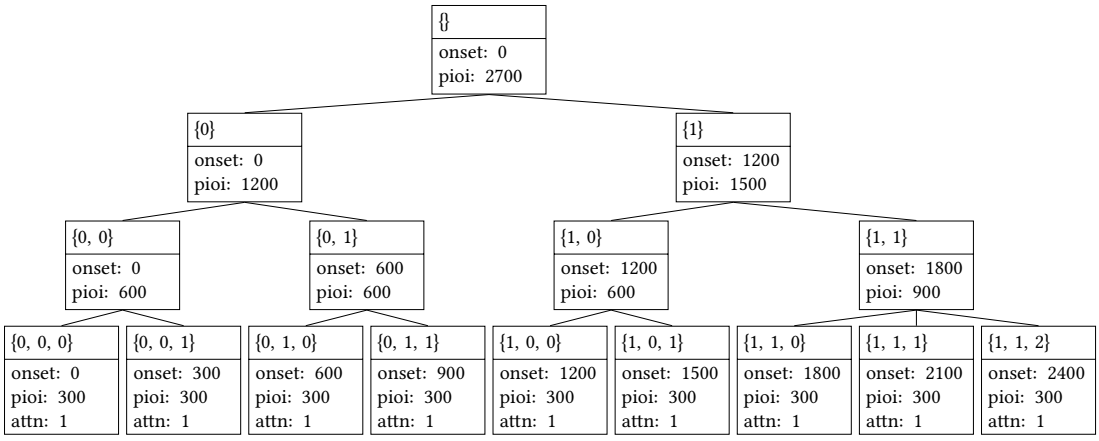

a: $\frac{3}{4}$ with no beat subdivision.




b: $\frac{3}{4}$ with quaver beat subdivision.




c: $\frac{4}{4}$ with quaver beat subdivision.




d: Non-isochronous 2-2-2-3 N-cycle grouping pattern.

Figure 5.2: Tree representation of metre, corresponding to the prototypical metrical patterns shown in figure 5.1. The upper part of each node contains the node label, the lower part contains the pulse onset and IOI. Each leaf node additionally contains an attentional energy value.

### 5.3.4 Representation of the tactus within metrical trees

Unlike the measure period or N-cycle, which are associated with root and leaf nodes respectively, the tactus cycle cannot be defined in terms of tree structure. A number of cycles may be candidate tactus cycles in perception, and the tactus cycle is not necessarily the most salient cycle in terms of pulse-period salience (Parncutt 1994, p. 438), as evident in entrainment at fast and slow tempi across the range of tactus IOI from 200 ms to 2000 ms (London 2004, p. 27). Therefore, it is necessary to define the tactus level, $t$, explicitly.

**Definition 5.28.** Set $D' = D \setminus \{0, 5\}$ is the set of permissible tactus levels. The element $t \in D'$ denotes the tactus level in $\mathcal{T}$.

Furthermore, we constrain the IOI between tactus pulses to be between 200 ms and 2000 ms (5.12).

$$\forall v \in \textit{level}(t, V) \quad 200 \leq \textit{p-ioi}(v) \leq 2000 \tag{5.12}$$

Following London, by defining $t$ thus, we state that one cycle is always distinguished in metrical perception (WFC 2.1, p. 101), and that only a single cycle can hold this focus in attention at any given moment (WFC 1.3). The tactus cycle cannot be at depth zero, because at least one grouping cycle must exist above the tactus (WFC 2.2). It also cannot be at depth five because that would entail a measure period cycle of greater than 6000 ms, given the 200 ms tactus lower bound. Levels above the tactus level are referred to as *grouping cycles*, and levels below as *subdividing cycles*. Therefore, given a metrical tree of maximal height five, the tactus cycle may be denoted as any level from one to four. If $t$ = 1, the metre is said to contain one grouping cycle, and four subdividing cycles. At the other extreme, if $t$ = 4, then the metre contains four grouping cycles and a single subdividing cycle. Tables 5.1 and 5.2 show two maximal height metrical structures at both the lower and upper limits of the metrical temporal envelope (mean pulse IOI values of each cycle are shown, as defined by *mean-p-ioi* (5.9)).

Both metrical structures in tables 5.1 and 5.2 have duple relationships between the pulse IOIs on each metrical cycle, which afford the greatest number of cycles to be defined within the metrical envelope. The maximally subdivided metre is the fastest possible well-formed metrical structure with four levels of subdivision. A musical interpretation of this metre is that of a duple tactus metre of very slow tempo (37.5 bpm), with four levels of duple subdivision. Notated in conventional Western notation, with crotchet as tactus beat, the N-cycle, *level*$(5, V)$, pulses

Table 5.1: Mean pulse IOIs for cycles in a metre with a maximal number of subdividing cycles.

| Depth | Cycle | Mean pulse IOI |
|---|---|---|
| 0 | grouping 1 | 3200 |
| 1 | tactus | 1600 |
| 2 | subdividing 1 | 800 |
| 3 | subdividing 2 | 400 |
| 4 | subdividing 3 | 200 |
| 5 | subdivision 4 | 100 |

Table 5.2: Mean pulse IOIs for cycles in a metre with a maximal number of grouping cycles.

| Depth | Cycle | Mean pulse IOI |
|---|---|---|
| 0 | grouping 4 | 6000 |
| 1 | grouping 3 | 3000 |
| 2 | grouping 2 | 1500 |
| 3 | grouping 1 | 750 |
| 4 | tactus | 375 |
| 5 | subdividing 1 | 187.5 |

would correspond to hemidemisemiquavers, and the measure period, *level*$(0, V)$, a minim.

The maximally grouped metre is the slowest well-formed metrical structure with four levels of grouping. This structure corresponds to a metre with 16 tactus beats at 160 bpm. Each tactus beat is recursively grouped into pairs, with one level of duple subdivision. Again in common notation with crotchet equals tactus, the measure period would correspond to twice the length of a breve, or the medieval duple longa, while the N-cycle, would correspond to a quaver.

Arguably such use of notation would not be an efficient use of established conventions, and other means could be employed to convey extremely slow or fast passages more effectively. But moreover, it is questionable whether such metrical structure actually persist in cognition, despite being permissible within the formalisation. Although it is difficult to imagine music that would induce precisely these extreme metrical structures spontaneously, is it not inconceivable that music containing such periodic components *could* be experienced within these metrical frameworks, even if only with intentional effort. Furthermore, on the one hand, these metres are extremely simple, being composed solely of equal duple relation-

ships between the pulses of adjacent cycles. On the other hand, the height of the structures makes them relatively complex. It is difficult to conceive of a 32-cycle N-cycle consisting of 100 ms pulse IOIs, or a 6 second long measure period with the same clarity as a moderately paced 3- or 4-cycle tactus. This is not surprising given what is known about human perception and reproduction of time intervals. However, such cycles are assumed here to be part of the concept of metre, important in the context of the specific combination of periodic components in which they exist, even if their exact functional role in the structure is not foremost in perception. The key point is that there is no easy distinction to be made between the clearly conscious aspects of metrical perception, and perhaps pre-conscious correlates, that nonetheless contribute to our experience of metre.

The metrical level to which we attend as the tactus may of course change, depending on: tempo fluctuations, in which case attention is likely to tend towards the centre of the range of pulse salience; rhythmic accentuation (McKinney and Moelants 2006); or simply through a conscious shifting of attention to faster or slower periodicities. Such 'flipping' of attention between metrical cycles is modelled by distinct trees. Even when the node structure of a metrical hierarchy does not change, modelled in $\mathcal{T}$ by the elements $(V, E, L, O, I, A)$, another cycle can still become foregrounded in attention and become the tactus. This is represented in the model as a change in $t$. In terms of the conceptual level representation, a change in $t$ results in the metre being represented by a different point in the space, and the vector between the old and new points represents the phenomenon of shifting tactus levels. By assuming that the absolute depth of the tactus cycle is given, we avoid issues of tempo and metrical induction. However, future work should investigate the integration of the conceptual models developed herein with existing models of beat tracking and metrical induction.

## 5.4 Conceptual space of periodic metrical structure

This formalisation defines a high-level conceptualisation of metre, concerning only properties of metrical periodicity. Despite the primary periodic nature of metre, sequentiality also has an important role, particularly when considering metres comprising patterns of non-equally spaced beats, or the question of how attentional energy is distributed *within* a periodic pattern. The conceptual space described in this section concerns only the periodic structure of metre, and abstracts away the sequential structure internal to periodic patterns. A conceptual space which also represents sequence within metrical cycles is addressed in section 5.5.

### 5.4.1 Domains of metrical periodicity

We now construct a conceptual space representing metrical periodicity, in terms of the above tree formalisation of London's theory of metre. An earlier definition of this space was published by Forth et al. (2010). Since publication, a number of questions raised by this initial formalisation have been addressed. The principle objection to the previous formalisation is its permitting of undefined values in certain dimensions, thus compromising some of the benefits of a purely geometrical model (see further comment in appendix D). In the revised formalisation, undefined values are no longer necessary, making the model simpler, and more useful. The improved model is defined below.

In the following, quality dimensions are defined to represent the variable perceptual qualities of concepts. In order to model the perceptual qualities of metrical structure, it is necessary to define multiple dimensions of the same type, which are grouped together into a domain of that type. Therefore, a domain is simply a multidimensional space whose regions correspond to Gärdenfors' notion of natural properties. We shall define three domains here, composed of typed dimensions, each of which is a subspace of the total space. Therefore, regions across the total space correspond to Gärdenfors' notion of a natural concept, which in our case is a representation of perceived metrical structure.

To avoid repetition below, we first define the aspects of the representation that are applicable to all dimensions and domains. All spaces are normed vector spaces supporting the usual operations of vector addition and scalar multiplication. As in chapter 4, $L^1$ and $L^2$ norms are considered, and treated as factors in the subsequent evaluation. However, as we are now dealing with identifiable domains within the conceptual space, we have the flexibility to treat domains hierarchically and to define different norms within different domains. Therefore, we also consider the case of $L^2$ *within* domains, and $L^1$ *between* domains, denoted $L^1+L^2$. The structure of this hierarchical space follows Gärdenfors' recommendation of using Euclidean distance for integral qualities (as we have within domains) and city-block distance for separable qualities (as we have between domains). To compute this distance, the distance between points within each $L^2$ normed domain is first calculated, and then these distances are treated as points in a meta-level $L^1$ space, to the effect that the overall distance is an additive combination of the within-domain distances.

To address the issue of different dimensional scales, all distances within domains are min-max normalised (Jain et al. 2005), meaning that distances are linearly scaled to the range [0,1] with respect to the longest vector permissible within

the domain.[24]

Gärdenfors' allows for salience weights to be attached to each dimension, as well as to each domain within a conceptual space. Here we are only concerned with salience weights of domains. Within each domain, all dimensions are considered to have an implicit salience weight of one. Analogous to the model fitting process in chapter 4, the salience weights for each domain will be the free parameters of the models to be optimised as part of the evaluation.

#### MEAN_P_IOI

First, we define a MEAN_P_IOI quality dimension to represent the mean IOI between attentional timepoints of a metric cycle, measured in milliseconds, and defined over the range 12.5–48000 ms.[25]

$$\text{MEAN\_P\_IOI} = \{x \in \mathbb{R} \mid 12.5 \leq x \leq 48000\} \tag{5.13}$$

As defined in section 5.3.3, a metrical tree can be composed of at most six cycles, with up to four possible cycles above or below the tactus at any instant depending on the level of $\mathcal{T}$ representing the tactus. In order to be able to represent the mean pulse IOI of all possible hierarchical groupings of cycles, we construct a 9-dimensional domain $\text{MEAN\_P\_IOI}^9$. Each dimension corresponds to a potential metrical cycle arranged with respect to $\mathcal{T}$ from the top down and relative to the tactus $t(\mathcal{T})$.

The dimensional structure of all domains is defined with reference to the tactus, since this cycle is defined for all metres. Individual dimensions of a domain are denoted with a suffix integer, starting from zero, contained in square brackets. The distinguished dimension $\text{MEAN\_P\_IOI}^9[4]$ represents the mean IOI of the tactus cycle. Dimensions 0–3 represent grouping cycles above the tactus, and dimensions 5–8 represent tactus subdivisions, of increasing depth with respect to $\mathcal{T}$. Algorithm `MAP_T_MEAN_PIOI` (5.1) defines the mapping from $\mathcal{T}$ to a point in the domain of $\text{MEAN\_P\_IOI}^9$.

[24] "Permissible" here means that defined constraints on metrical structure are taken into account, so distances are normalised according to the longest vector within the region of a domain corresponding to well-formed metrical properties.

[25] The lower and upper bounds here are greater than for well-formed metrical pulse IOIs because the geometry explicitly represents four grouping levels, and four subdivision levels, regardless of which conform to well-formed metrical cycles. Therefore, 12.5 ms corresponds to a fourth level of subdivision below the tactus in the case of a minimal N-cycle at subdivision level one ($\frac{100}{2^3}$). 48000 ms corresponds to a fourth level of grouping above the tactus in the case of a maximal measure period at grouping level one ($6000 \cdot 2^3$).

**Algorithm 5.1** MAP_T_MEAN_PIOI

**Require:** $\mathcal{T}$

1: MEAN_P_IOI$^9 \leftarrow 0$ // initialise array with 0
2: $h \leftarrow$ *height*$(\mathcal{T})$
3: $i \leftarrow 4 - t(\mathcal{T})$ // absolute measure-period depth
4: **for** $d \leftarrow 0$ **to** $h$ **do** // map levels of $\mathcal{T}$
5: MEAN_P_IOI$^9[d + i] \leftarrow$ *mean-p-ioi*$(d, V(\mathcal{T}))$
6: **end for**
7:
8: **for** $j \leftarrow 2$ **to** $0$ **do** // fill unspecified grouping cycles
9: **if** MEAN_P_IOI$^9[j] = 0$ **then**
10: MEAN_P_IOI$^9[j] \leftarrow$ MEAN_P_IOI$^9[j + 1] * 2$
11: **end if**
12: **end for**
13:
14: **for** $j \leftarrow 5$ **to** $8$ **do** // fill unspecified subdivision cycles
15: **if** MEAN_P_IOI$^9[j] = 0$ **then**
16: MEAN_P_IOI$^9[j] \leftarrow$ MEAN_P_IOI$^9[j - 1]/2$
17: **end if**
18: **end for**
19: **return** MEAN_P_IOI$^9$

Where a cycle is not present in the metre, an implicit duple grouping or subdivision cycle is assumed, depending on whether it is above or below the tactus respectively. Lines 8–12 ensure that unspecified groupings cycles above the first grouping cycle, which is always specified because at least one grouping cycle must exist above the tactus (WFC 2.2, p. 101), are filled with implicit duple groupings. Lines 14–18 ensure that unspecified subdivision cycles below the tactus are filled with implicit duple subdivisions. Duple grouping and subdivision cycles are assumed over triple cycles because a number of studies have shown that duple relationships are dominant in perception in the absence of cues to the contrary (Desain and Honing 2003), and both adults and children are more able to accurately reproduce rhythms with a duple metre, as opposed to triple (Drake 1993). Furthermore, duple groupings and subdivisions are more likely to fall within the perceptual time frame of metre, and therefore are more likely to be applicable across a wider range of stimuli.

It is possible to formalise well-formedness constraints of metrical structure at both symbolic and geometrical levels of representation. To serves as an example of how a mapping between multiple levels of representation can offer complementary perspectives on the same underling concept, the logical expression in (5.14) imposes a constraint on MEAN_P_IOI$^9$ such that the ratio between values of neigh-

bouring dimensions must be between 2:1 and 3:1 inclusively. In geometrical terms, this defines regions of well-formedness within the domain, which is a valuable property of the space that could be exploited within applications exploring, or otherwise utilising, a spatial representation of metrical structure. In symbolic terms, this constraint simply makes explicit the implications of the symbolic definition of node arity (5.8), which ensures that all internal nodes in $\mathcal{T}$ are of arity two or three, and constraint (5.4), which requires that the pulse IOI value associated with each parent node be equal to the sum of pulse IOI values associated with its children. The geometric constraint does not place any additional restriction on metrical structure beyond what is already required by the symbolic formalism stated in section 5.3.3. None of these constraints prevent non-integer ratios, within the range 2:1–3:1, between neighbouring dimensional values in $\text{MEAN\_P\_IOI}^9$. A non-integer relationships between $\text{MEAN\_P\_IOI}^9$ dimensional values simply represents a non-isochronous beat grouping structure. The domain of $\text{MEAN\_P\_IOI}^9$ is not expressive enough to represent the sequential structure of non-isochronous grouping structures. For example, a cycle consisting of a short-short-long sequence of beats is represented by the same dimensional value as a short-long-short cycle, assuming the same pulse IOI values, because the mean pulse IOI value across each cycle are equal. However, both cycles are considered distinct from a short-long-long cycle involving the same pulse IOI values because the mean pulse IOI value across the cycle will be different due to the relative number of short versus long beats.

$$\forall i \in [1..8] \quad 2 \cdot \text{MEAN\_P\_IOI}^9[i] \leq \text{MEAN\_P\_IOI}^9[i-1] \leq 3 \cdot \text{MEAN\_P\_IOI}^9[i] \quad (5.14)$$

As concrete examples of mapping metrical structure to the geometrical domain of $\text{MEAN\_P\_IOI}^9$, consider $\frac{3}{4}$, as represented in figure 5.2a, and the non-isochronous metre with a 2-2-2-3 N-cycle grouping pattern in figure 5.2d. The $\frac{3}{4}$ metre has two cycles: a 3-cycle N-cycle, representing an isochronous 600 ms tactus cycle, and a single grouping cycle, the measure period. As a point in the domain of $\text{MEAN\_P\_IOI}^9$, this metre is represented as follows.

$$\tfrac{3}{4} \text{ (3-cycle)} \quad \langle 14400, 7200, 3600, \mathbf{1800}, \mathbf{600}, 300, 150, 75, 37.5 \rangle$$

As with the subsequent example, bold values indicate cycles explicitly present in the metre, and non-bold values indicate implicit cycles filled in by algorithm `MAP_T_MEAN_PIOI` (5.1). The dimensional values of the domain can be seen to satisfy constraint (5.14).

The non-isochronous metre based on a 2-2-2-3 N-cycle grouping pattern con-

tains four cycles, with a tactus cycle consisting of a short-short-short-long pulse IOI sequence: $\langle 600, 600, 600, 900\rangle$. As a point in MEAN_P_IOI$^9$, this metre is represented as follows.

2-2-2-3 N-cycle metre $\quad \langle 10800, 5400, \mathbf{2700}, \mathbf{1350}, \mathbf{675}, \mathbf{300}, 150, 75, 37.5\rangle$

The ratio between the value of MEAN_P_IOI$^9$[4] (= 675) and MEAN_P_IOI$^9$[5] (= 300) is 2.25, as a consequence of the tactus cycle grouping N-cycle beats into both twos and threes.

An important consequence of filling in unspecified cycles with implicit duple cycles is that within this domain, there is no way to differentiate between particular metrical types. For example, the $\frac{3}{4}$ metre in figure 5.2b, which is a $\frac{3}{4}$ metre with an explicit quaver subdivision cycle, is represented by exactly the same point in this domain as the $\frac{3}{4}$ metre without an additional duple subdivision cycle. Likewise for grouping cycles: both conventional $\frac{2}{4}$ and $\frac{4}{4}$ with the same mean tactus pulse IOI are represented by the same point in the space. Therefore, MEAN_P_IOI$^9$ represents a broad range of synchronised periodicities related to any particular metrical concept. This turns out to be a valuable property of the space, as intuitively, such metres are highly similar. However, the subtle differences in these distinct patterns of entrainment must be accounted for. Therefore, the actual prominence of these periodicities in perception is represented by the following domain.

#### MEAN_A_ENERGY

The quality dimension MEAN_A_ENERGY represents the mean attentional energy of the timepoints within a given metric cycle, in the unit range $[0, 1]$:

$$\text{MEAN\_A\_ENERGY} = \{x \in \mathbb{R} \mid 0 \leq x \leq 1\} \tag{5.15}$$

Analogous to MEAN_P_IOI$^9$, we define a 9-dimensional domain MEAN_A_ENERGY$^9$ in order to represent the mean attentional energy given to each metrical cycle in all possible hierarchical organisations relative to the tactus, represented by the distinguished dimension MEAN_A_ENERGY$^9$[4]. Algorithm MAP_T_MEAN_AENERGY (5.2) defines the projection from $\mathcal{T}$ to a point in this domain.

Taking again the examples of the prototypical $\frac{3}{4}$ metres from figure 5.2, the difference between the 3-cycle and the 6-cycle variants are represented as follows.

$\frac{3}{4}$ (3-cycle) $\quad \langle 0, 0, 0, 1, 1, 0, 0, 0, 0\rangle$
$\frac{3}{4}$ (6-cycle) $\quad \langle 0, 0, 0, 1, 1, 1, 0, 0, 0\rangle$

**Algorithm 5.2** MAP_T_MEAN_AENERGY

**Require:** $\mathcal{T}$

1: MEAN_A_ENERGY$^9 \leftarrow 0$ // initialise array with $0$.
2: $h \leftarrow$ *height*$(\mathcal{T})$
3: $i \leftarrow 4 - t(\mathcal{T})$ // absolute measure-period depth
4: **for** $d \leftarrow 0$ **to** $h$ **do** // map levels of $\mathcal{T}$
5: MEAN_A_ENERGY$^9[d + i] \leftarrow$ *mean-a-energy*$(d, V(\mathcal{T}))$
6: **end for**
7: **return** MEAN_A_ENERGY$^9$

In the case of prototypical metres, where all attentional energy values default to one, values in MEAN_A_ENERGY dimensions are either one if a cycle is present in a metre, or zero otherwise. Therefore, for prototypical metres, this domain can be thought of as simply a binary vector indicating the presence or absence of metrical levels in a metre. However, the more general definition of mean cycle energy is useful when considering metrical instances with varying attentional energy. Furthermore, it is essential that the range of MEAN_A_ENERGY dimensions be continuous between zero and one to permit various geometric operations, such as generating a series of interpolations between two metres, or calculating a centroid of a set of metres. The centroid between the above 3-cycle and 6-cycle metres is simply:

$$\tfrac{3}{4} \text{ (6-cycle)} \quad \langle 0, 0, 0, 1, 1, 0.5, 0, 0, 0 \rangle$$

Metrical well-formedness requires that this metre is still structurally a 6-cycle, but that the 6-cycle itself is less prominent in attention. Exactly *how* it is less prominent, in terms of the sequence of accentuated timepoints, cannot be represented in this space because only an average over the whole cycle is represented. However, a higher-dimensional space capable of representing individuated metres based on sequential cycle structure is defined in section 5.5.

An interesting, and potentially testable consequence of using continuous values to represent attentional energy is that traditional metrical concepts, such as the concepts represented by time signatures, correspond to regions within this space, and are not single discrete points. Intuitively, this is attractive as in perception there is often not a clear distinction between structurally similar metres, for example, $\frac{2}{4}$ and $\frac{4}{4}$ metres at the same tempo. The difference in this case lies primarily in the strength of the second grouping cycle above the tactus, represented in dimension MEAN_A_ENERGY$^9$[2]. A reasonable hypothesis might assume that when the mean attentional energy in this dimension is less-than 0.5, the metre is likely to be considered $\frac{2}{4}$, but $\frac{4}{4}$ when greater-than 0.5, perhaps subject to hysteresis if one

or other metre is primed in perception, or a suitable rhythmic stimulus indicative of one prototypical metre were to gradually transform into one indicative of the other.

#### C_RATIO

The final domain necessary to specify a conceptual space of metrical periodicity represents the hierarchical relationship between cycles, independent of tempo. Let C_RATIO be a quality dimension defined over the range $[2, 3]$:

$$\text{C\_RATIO} = \{x \in \mathbb{R} \mid 2 \leq x \leq 3\} \tag{5.16}$$

In terms of levels in $\mathcal{T}$, the C_RATIO value of a cycle of depth $d$ is equal to the cardinality of level $d+1$ divided by the cardinality of level $d$. We therefore construct a domain of C_RATIO$^8$, with one fewer dimensions than the previous two domains because cardinality ratio is undefined for the N-cycle. The dimensions in C_RATIO$^8$ again correspond to the top-down ordering of levels in $\mathcal{T}$, and C_RATIO$^8[4]$ represents the cardinality ratio of the tactus. Algorithm MAP_T_MEAN_CRATIO (5.3) defines the mapping from $\mathcal{T}$ to a point in C_RATIO$^8$. Analogously to MEAN_P_IOI$^9$, unspecified levels of $\mathcal{T}$ are assumed to be implicit duple groupings or subdivision, which is achieved by initialising all dimensional values to 2 prior to mapping levels of $\mathcal{T}$.[26]

**Algorithm 5.3** MAP_T_MEAN_CRATIO

**Require**: $\mathcal{T}$

1: C_RATIO$^8 \leftarrow 2$ // initialise array with 2
2: $h \leftarrow \mathit{height}(\mathcal{T}) - 1$
3: $i \leftarrow 4 - t(\mathcal{T})$ // absolute measure-period depth
4: **for** $d \leftarrow 0$ **to** $h$ **do** // map levels of $\mathcal{T}$
5: C_RATIO$^8[d+i] \leftarrow \frac{\mathit{cardinality}(d+1, V(\mathcal{T}))}{\mathit{cardinality}(d, V(\mathcal{T}))}$
6: **end for**
7: **return** C_RATIO$^8$

The $\frac{3}{4}$ metre from figure 5.2a, and the non-isochronous 2-2-2-3 N-cycle metre from figure 5.2d are respectively mapped to the following points in C_RATIO$^8$.

| | |
|---|---|
| $\frac{3}{4}$ (3-cycle) | $\langle 2, 2, 2, \mathbf{3}, 2, 2, 2, 2 \rangle$ |
| 2-2-2-3 N-cycle metre | $\langle 2, 2, \mathbf{2}, \mathbf{2}, \mathbf{2.25}, 2, 2, 2 \rangle$ |

[26] Equivalently, a domain representing the hierarchical relationship between cycles could be defined as a mapping from MEAN_P_IOI$^9$. We use C_RATIO$^8$ here for simplicity. However, the development of a conceptual space formalism within which geometrical constructs can be defined hierarchically, akin to the multiple viewpoint formalism for symbol sequences (Conklin and Witten 1995), would greatly simplify the construction of complex multi-domain conceptual spaces.

Dimensional values in bold again denote values derived from explicit cycles present in a metre, of which there is one fewer than the total number of cycles due to the definition of cardinality-ratio as a relational quality between adjacent cycles.

#### METRE-P

Each domain defined above is combined into a single conceptual space to produce a multi-faceted geometric representation of metrical periodicity. A minimal conceptual space, METRE-P, capturing London's notion of periodic flow of attentional energy, can be constructed thus.

$$\text{METRE-P} \quad = \quad \text{MEAN\_P\_IOI}^{9} \times \text{MEAN\_A\_ENERGY}^{9} \times \text{C\_RATIO}^{8} \tag{5.17}$$

The domains of METRE-P, together with their salience weights, represent variably salient orthogonal qualities of metrical entrainment. The domain of $\text{MEAN\_P\_IOI}^{9}$ represents the absolute frequency of the oscillatory components of a metre. Within this domain, metres comprising the same oscillatory components are considered identical. The domain of $\text{MEAN\_A\_ENERGY}^{9}$ represents the absolute level of attentional energy associated with each oscillatory component. Within this domain, metres with the same distribution of attentional energy across cycles, or that simply have the same number of grouping and subdivision cycles in the case of prototypical metres, are considered identical. The domain of $\text{C\_RATIO}^{8}$ represents the structural relationship between cycles, analogous to $\text{MEAN\_P\_IOI}^{9}$ except in relative terms. Within this domain, metres with the same hierarchical structure are identical, irrespective of tempo. The domain of $\text{C\_RATIO}^{9}$ is directly computable from $\text{MEAN\_P\_IOI}^{9}$. However, the facets of metrical similarity modelled in terms of distances within each domain are distinct, and both must be explicitly represented in the space in order to model both absolute and relative qualities of metrical timing.

### 5.4.2 Discussion

The geometrical requirement of the space that unspecified cycles in $\mathcal{T}$ are assumed to be implicit duple cycles avoids the need for undefined values within the formalisation. However, it may be argued that as only six hierarchically coordinated duple cycles can fit within the temporal envelope of metre, and a lesser number when triple relationships between cycles are present, that the values representing some, if not all, implicit cycles in $\text{MEAN\_P\_IOI}^{9}$ are meaningless to the concept of

metre. Indeed, the range of MEAN_P_IOI dimensions is substantially larger than the temporal extent of metrical entrainment. This requirement of the space is necessary from the perspective of the vector space formalism, but it is unsatisfactory from a cognitive perspective. One possible compromise could involve dimensional salience weights, which here are all assumed to be one. Dimensional saliences could be defined as a function of the dimensional values, in line with evidence of tempo preference (Moelants 2002). As such, when values exceed perceptual limits, corresponding salience weights would tend towards zero, effectively projecting out uninformative dimensions. However, this possible development will be left for future work.

## 5.5 Conceptual space of sequential metrical structure

The general conceptual space approach to representation remains the same as in the previous model of metrical periodicity, but here the concepts are represented at a lower level of abstraction. Regions of this space do not only represent the periodic qualities of metre, but also the sequential structure within metric cycles.

### 5.5.1 Domains of metrical sequence

Two domains, in the Gärdenfors sense of sets of integral dimensions, are posited as spaces within which two types of properties pertaining to the sequential structure of individuated tempo-metrical types are represented. The general vector space assumptions outlined in section 5.4.1 are also assumed here.

#### P_IOI

We begin again by defining the quality dimensions of each domain. The dimension P_IOI represents the IOI between attentional timepoints on a metrical cycle, measured in milliseconds, and defined over the range 0–48000.

$$\text{P\_IOI} = \{x \in \mathbb{R} \mid 0 \leq x \leq 48000\} \tag{5.18}$$

To represent the sequential structure of a cycle explicitly requires the availability of a dimension for every possible hierarchically determined categorical timepoint. In terms of $\mathcal{T}$, this means a distinct dimension for all possible paths leading from the root $r$ to a node, for all possible values of $t$. Ignoring the implications of $t$ initially and just considering $\mathcal{T}$ as an ordinary tree, representing the pulse IOI of $r$ would only require a single P_IOI dimension. Representing the level below the

root requires a minimum of three dimensions: one for each of the three possible children. In general, each level of depth $d$ requires $3^d$ dimensions to represent the full range of hierarchically determined categorical timepoints.[27]

$\mathcal{T}$ has a maximum height of five, but recall that the tactus cycle $t$ may be defined as any level from one to four, depending on pulse rate, creating a maximum of either four subdividing cycles (plus one grouping cycle), or four grouping cycles (plus one potential subdividing cycle). Thus, the geometry must be able to represent all nine distinct levels, as in the previous domains of MEAN_P_IOI$^9$ and MEAN_A_ENERGY$^9$ in section 5.4.1. Therefore, to represent the sequential structure of all possible hierarchical arrangements of metrical cycles relative to the tactus requires a domain of $\sum_{n=0}^{8} 3^n = 9841$ dimensions. This is a very high number of dimensions, particularly for a notional conceptual space. However, as discussed in section 5.4.2, only a subset of dimensions ever represent perceptually salient qualities at any one time. In this space, the minimal subset of dimensions needed to represent a well formed metrical structure, the most basic 2-cycle metre, consists of only four P_IOI dimensions. Only three of which correspond to explicit pulse IOI values in the symbolic definition of metre: the measure period; the first tactus beat; and the second tactus beat. The fourth P_IOI dimension, taking a value of zero, represents the non-presence of a potential third tactus beat, and ensures that the complete hierarchical structure of a simple two-cycle duple metre is explicitly represented. However, the high-dimensionality of the entire domain is necessary within the vector space formalism we have adopted, which requires that all theoretically possible dimensions be defined and take on some value consistent with the perceptual qualities of the concept represented.

Dimensions in P_IOI$^{9841}$ are arranged in label-order relative to $t(\mathcal{T})$, starting with grouping cycle 4, down to subdividing cycle 4. Algorithm `MAP_T_PIOI` (5.4) is the top-level procedure that maps instances of $\mathcal{T}$ to a point in the domain of P_IOI$^{9841}$. Lines 2–8 take each node $v \in V(\mathcal{T})$, and assign the pulse IOI value $i \in I(\mathcal{T})$ associated with node $v$ to a dimension in P_IOI$^{9841}$. Calculating the dimension offset for each node depends on two functions. Function *abs-depth-offset*$(.)$ (5.19) computes the dimension offset in P_IOI$^{9841}$ corre-

[27] We have also developed two further models of metrical structure where duple and triple subdivisions are represented by distinguished dimensions. These spaces require respectively $4^n$ and $5^n$ dimensions to represent each metrical level. Full consideration of these spaces here is beyond the current scope, but future work could evaluate the different notions of embedded similarity, and in particular, investigate the different characteristics of trajectories through each space.

sponding to the depth of node $v$ in $\mathcal{T}$, relative to the tactus $t(\mathcal{T})$.

$$\begin{aligned} \textit{abs-depth-offset} : \mathbb{N} &\rightarrow \mathbb{N} \\ n &\mapsto \begin{cases} 0 & \text{if } n = 0 \\ \sum_{i=1}^{n} 3^{i-1} & \text{otherwise} \end{cases} \end{aligned} \tag{5.19}$$

Function *within-cycle-offset*(.) (5.20) takes the label $l \in L(\mathcal{T})$ associated with each node $v$, and from that computes an offset based on the unique path from the root $r$ to $v$ represented by the corresponding label $l$,

$$\begin{aligned} \textit{within-cycle-offset} : L &\rightarrow \mathbb{N} \\ l &\mapsto \begin{cases} 0 & \text{if } l = \langle\rangle \\ \sum_{i=1}^{n} l_i \cdot 3^{n-i} & \text{otherwise} \end{cases} \end{aligned} \tag{5.20}$$

where $n$ is equal to the number of elements in label $l$, and $l_i$ is the $i$th element of $l$ indexed from 1. The values computed by each of these functions are summed to produce the unique offset in P_IOI[9841] corresponding to the categorical timepoint represented by each node.

Lines 2–8 of algorithm `MAP_T_PIOI` only set the value of dimensions corresponding to the nodes defined within any instance of $\mathcal{T}$. However, as with MEAN_P_IOI[9] in section 5.4.1, all dimensions must take a value consistent with a well-formed hierarchical structure spanning the entire space in order to preserve the correctness of the geometry. Again, implicit duple groupings and subdivisions are assumed. Therefore, line 9 of algorithm `MAP_T_PIOI` calls auxiliary algorithm `FILL_PIOI_DOMAIN` (5.5), which fills the space with implicit duple relationships between dimensional values based on the structure of the explicit metrical hierarchy.

**Algorithm 5.4** `MAP_T_PIOI`

**Require:** $\mathcal{T}$

1: $\text{P\_IOI}^{9841} \leftarrow 0$ // initialise array with 0
2: **for** $v \in V(\mathcal{T})$ **do** // map all nodes in $\mathcal{T}$
3: $a \leftarrow 4 - t(\mathcal{T}) + \textit{depth}(v)$ // absolute depth of $v$ relative to $t(\mathcal{T})$
4: $i \leftarrow \textit{abs-depth-offset}(a)$
5: $l \leftarrow \textit{label}(v)$
6: $j \leftarrow \textit{within-cycle-offset}(l)$
7: $\text{P\_IOI}^{9841}[i + j] \leftarrow \textit{p-ioi}(v)$
8: **end for**
9: $\text{P\_IOI}^{9841} \leftarrow$ `FILL_PIOI_DOMAIN`$(\mathcal{T}, \text{P\_IOI}^{9841})$ // call algorithm 5.5
10: **return** $\text{P\_IOI}^{9841}$

Algorithm `FILL_PIOI_DOMAIN` includes two distinct stages for filling-in implicit duple relationship across the space. In tree-theoretic terms, lines 1–10 begin with dimensions corresponding to explicit N-cycle pulse IOI values, and iterate over all subsequent dimensions corresponding to further subdividing pulse IOIs, assigning implicit duple subdivisions. This results in a notional metrical tree that has the theoretical maximum of four subdividing cycles below the tactus. Lines 12–24 proceed by iteratively creating an additional duple grouping cycle, and then copying the entire underlying tree structure so that the well-formedness is maintained below the newly created grouping cycle. The process continues until the structure reaches the theoretical maximum of four grouping cycle above the tactus. Line 19 calls auxiliary algorithm `REPEAT_SUB_SEQ` (5.6), which performs the copying of values across dimensions. It is presented separately here purely for notational clarity.

**Algorithm 5.5** FILL_PIOI_DOMAIN

**Require:** $\mathcal{T}$
**Require:** $\text{P\_IOI}^{9841}$
1: $a \leftarrow 4 - t(\mathcal{T}) + \textit{height}(\mathcal{T})$ // N-cycle depth relative to $t(\mathcal{T})$
2: $i \leftarrow \textit{abs-depth-offset}(a)$ // parent node offset
3: $j \leftarrow \textit{abs-depth-offset}(a+1)$ // child node offset
4: **while** $j < 9841$ **do** // fill implicit duple subdivisions
5: **for** $k \leftarrow 1$ **to** $2$ **do**
6: $\text{P\_IOI}^{9841}[j] \leftarrow \text{P\_IOI}^{9841}[i]/2$
7: $j \leftarrow j + k$
8: **end for**
9: $i \leftarrow i + 1$
10: **end while**
11:
12: $b \leftarrow 4 - t(\mathcal{T})$ // absolute measure-period depth
13: **while** $b > 0$ **do**
14: $i \leftarrow \textit{abs-depth-offset}(b-1)$ // parent node offset
15: $j \leftarrow \textit{abs-depth-offset}(b)$ // child node offset
16: $\text{P\_IOI}^{9841}[i] \leftarrow 2 * \text{P\_IOI}^{9841}[j]$ // fill next implicit duple grouping
17: $c \leftarrow 0$
18: **while** $j < 9841$ **do** // fill subdivisions relative to new grouping
19: REPEAT_SUB_SEQ($\text{P\_IOI}^{9841}, j, 3^c$) // call algorithm 5.6
20: $c \leftarrow c + 1$
21: $j \leftarrow j + 3^{c+b-1}$
22: **end while**
23: $b \leftarrow b - 1$
24: **end while**
25: **return** $\text{P\_IOI}^{9841}$

**Algorithm 5.6** REPEAT_SUB_SEQ

**Require:** $\text{P\_IOI}^{9841}$
**Require:** a // start index of sub-sequence to copy
**Require:** b // length of sub-sequence to copy
1: $i \leftarrow a$
2: **while** $i < a + b$ **do**
3: $\text{P\_IOI}^{9841}[i+b] \leftarrow \text{P\_IOI}^{9841}[i]$
4: $i \leftarrow i + 1$
5: **end while**

#### A_ENERGY

Next we define a dimension A_ENERGY, to represent the attentional energy associated with a categorical metrical timepoint. Each A_ENERGY dimension can take a positive real value in the range $[0, 1]$, where 0 represents no attentional energy,

and 1 the maximum emphasis for a given piece of music.

$$\textsc{a\_energy} = \{x \in \mathbb{R} \mid 0 \leq x \leq 1\} \tag{5.21}$$

Attentional energy values are only defined for n-cycle timepoints. Therefore $3^8$ individual dimensions are required to represent all theoretically possible categorical timepoints. The dimensions in $\textsc{a\_energy}^{6561}$ are arranged again in label-order with respect to $\mathcal{T}$. Offset values indexing $\textsc{a\_energy}^{6561}$ are calculated according to function *abs-within-n-cycle-offset*$(.,.)$ (5.22),

$$\begin{aligned} \textit{abs-within-n-cycle-offset} : L \times D' &\rightarrow \mathbb{N} \\ (l, d) &\mapsto \begin{cases} 0 & \text{if } l = \langle\rangle \\ \sum_{i=1}^{n} l_i \cdot 3^{4+d-i} & \text{otherwise} \end{cases} \end{aligned} \tag{5.22}$$

where $n$ is equal to the number of elements in label $l$, $l_i$ is the $i$th element of $l$ indexed from 1, and $d$ is the depth of the tactus cycle.

The mapping from $\mathcal{T}$ to $\textsc{a\_energy}^{6561}$ is defined in algorithm MAP_T_AENERGY (5.7). All values in $\textsc{a\_energy}^{6561}$ are defined as zero, unless assigned a value by the algorithm.

**Algorithm 5.7** MAP_T_AENERGY

**Require**: $\mathcal{T}$

1: $\textsc{a\_energy}^{6561} \leftarrow 0$ // initialise array with 0
2: **for** $v \in$ *leaf-nodes*$(V(\mathcal{T}))$ **do**
3: $l \leftarrow$ *label*$(v)$
4: $d \leftarrow t(\mathcal{T})$ // tactus depth in $\mathcal{T}$
5: $j \leftarrow$ *abs-within-n-cycle-offset*$(l, d)$
6: $\textsc{a\_energy}^{6561}[j] \leftarrow$ *a-energy*$(v)$
7: **end for**
8: **return** $\textsc{a\_energy}^{6561}$

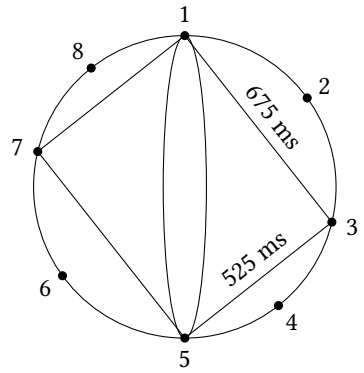
#### METRE-S

The conceptual space of METRE-S is then simply the Cartesian product of the domains $\textsc{p\_ioi}^{9841}$ and $\textsc{a\_energy}^{6561}$.

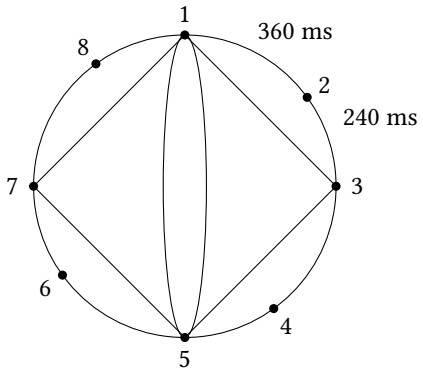
$$\textsc{metre-s} \quad = \quad \textsc{p\_ioi}^{9841} \times \textsc{a\_energy}^{6561} \tag{5.23}$$

Points within $\textsc{p\_ioi}^{9841}$ represent tempo-metrical types individuated by variation in the micro-timing of component pulse IOIs. Thus, regions of the space can

be identified as corresponding to various tempo-metrical types, the centroids of which correspond to mechanically precise prototypical instances. Departing from the centroid of each region are located instances of the same broad concept, but which vary in the absolute timings of their internal periodic components. Taking the example of $\frac{4}{4}$, regions around the centroid of mechanical $\frac{4}{4}$ might be identified as corresponding to a four beat metre of alternating short and long tactus beats, as shown in figure 5.3a, or a swing metre with an isochronous 4-cycle tactus, but unequal tactus subdivisions, as depicted in figure 5.3b.



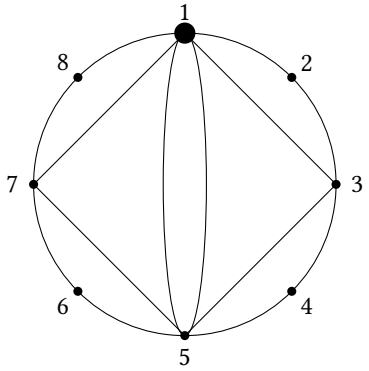

a: 8-cycle with alternating long-short tactus beats.



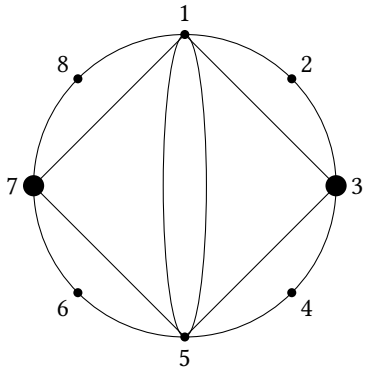

b: 8-cycle with a non-isochronous swing N-cycle.

Figure 5.3: Two examples of 8-cycle 100 bpm metres with individuated microtiming.

Regions in A_ENERGY$^{6561}$ represent metrical types individuated by the level of attentional energy associated with each metrical timepoint. For example, a region corresponding to general $\frac{4}{4}$-ness can be identified containing all $\frac{4}{4}$ metrical type instances distinguishable by the relative strength of the emphasis associated with each N-cycle point. Further sub-regions of this region might be identified as $\frac{4}{4}$ metrical concepts with strong downbeat emphasis, or strong second and fourth tactus beat emphasis, as depicted in figure 5.4a and figure 5.4b respectively by the size of the dots marking N-cycle timepoints.




a: 8-cycle with emphasised downbeat.




b: 8-cycle with emphasised second and fourth tactus beats.

Figure 5.4: Two examples of 8-cycle metres with different patterns of attentional energy.

## 5.6 Evaluation

To evaluate the validity of the notion of similarity embedded within the conceptual spaces defined previously, a genre classification task is performed. The aim here is not to evaluate the models as genre classifiers per se, but rather to evaluate the extent to which each space captures the nuances of metrical-rhythmic structure. Six genres, each with a recognisably distinct rhythmic feel are selected for the experiment: bolero, cha-cha, pasodoble, rumba, swing and waltz. If the structure of the spaces adequately captures salient aspects of metrical-rhythmic conceptualisation, and distances correspond with a notion of musical similarity, then distinctive regions should correspond with each genre.

### 5.6.1 Method

Three general models are evaluated in a series of genre classification tasks, employing $k$-nearest-neighbour classification over symbolic representations of music from a dataset of genre-labelled MIDI files (see section 5.6.2). A baseline model,

TEMPO, provides classification based purely on the tempo specified in each MIDI file. A tempo representation was chosen following previous results indicating that tempo is an important feature in discriminating between dance music styles (Gouyon et al. 2004). We then evaluate the classification accuracy achieved using METRE-P and METRE-S spaces respectively, with each of the norm conditions: $L^1$, $L^2$ and $L^1 + L^2$.

Each experimental run follows a standard classification task methodology. The method employed in each test is identical except for an additional optimisation step for METRE-P and METRE-S models in order to optimise the salience weightings of each domain. The optimisation stage is not necessary in the baseline model TEMPO as each piece of music is represented simply by a single tempo feature, or to use conceptual spaces terminology, a point in a 1-dimensional space of tempo.

The $k$-nearest-neighbour ($k$NN) algorithm is used as the classifier in each experiment. $k$NN is a simple classification scheme that assigns an unseen object to a class based on the class of the $k$ closest objects to it in a space. Since $k$NN is based on a notion of distance in a space, it is a very straightforward way of evaluating the conceptual space models herein developed.

One issue with $k$NN concerns the value of $k$, the optimal value of which cannot be determined a priori. Each model is evaluated using the values $k$ = 1, $k$ = 3 and $k$ = 4. Each unseen point is assigned a genre label based on the labels of its $k$ nearest neighbours by majority vote. Ties are decided in favour of the genre label associated with the point(s) of smallest mean distance from the testing point.

To minimise bias, a stratified 10x10-fold cross-validation (10x10cv) scheme is followed for each model evaluation. Standard 10-fold cross-validation splits the dataset randomly into 10 disjoint subsets. Each subset in turn is used as a testing set, while the remaining nine form a training set, which in the case of $k$NN is the set used to classify each testing point. Classification accuracy is recorded for the test set of each fold against the genre labels provided in the dataset. Stratified cross-validation is applied here, which ensures that a representative proportion of pieces from each genre is present in each subset. The same random folds are used in each model evaluation.

The partition used within a single 10-fold cross validation run can have a significant effect on the performance of a classifier, particularly for relatively small datasets. This can also lead to the misuse of hypothesis tests intended to justify the relative performance of a number of models over the same dataset (Salzberg 1997; Dietterich 1998; Nadeau and Bengio 2003; Bouckaert 2004). A number of recommendations for pairings of sampling schemes with appropriate hypothesis tests

are made in the previously cited literature. On the basis of replicability, Bouckaert (2004) recommends the use of the paired *t*-test on samples generated from a 10x10 sorted-runs sampling scheme. 10x10 sorted-runs sampling involves repeating standard 10-fold cross validation ten times, with each run using a different random split of the dataset. Each run yields ten accuracy measures, one for each fold, which are sorted. The mean accuracy for each fold, across all sorted runs is then computed, resulting in a sample of $n = 10$. This sample is taken to be representative of the accuracy of the classifier.

Bouckaert (2004) considers six sampling schemes, including conventional 10-fold cross-validation, and only average sorted runs is shown to have an acceptable Type 1 error of less than 5% as well as reasonable power. It is also shown that samples resulting from sorted-runs sampling do not heavily violate independence assumptions of commonly used hypothesis tests, of which the paired *t*-test is shown to be superior to the sign test, and marginally better than the rank sum (Wilcoxon's) test. One downside of employing 10x10 sorted runs sampling is the additional processing time, and for cases where this is prohibitive, methods such as 5x2-fold cross validation (Dietterich 1998) or the resampled *t*-test (Nadeau and Bengio 2003) may be appropriate. Following Bouckaert (2004), we apply a 10x10 sorted-runs sampling scheme, and compare differences in performance using a paired *t*-test.

The evaluation of METRE-P and METRE-S involve an additional optimisation step, carried out independently over each 10x10cv training set, in order to determine the optimal salience weights of the conceptual domains. Simulated annealing, a stochastic search algorithm, is employed as the optimisation method. (Kirkpatrick et al. 1983, see also section 4.3.1). To avoid overfitting, inner 10-fold cross-validation is performed. The training fold in each 10x10cv run is itself partitioned into 10 disjoint subsets, forming 10 inner training-testing folds. For each inner fold, 200 iterations of simulated annealing are performed, maximising classification accuracy. The mean of the optimised salience weights discovered within each inner CV run are then applied during the classification of the outer testing set.

### 5.6.2 Data

The dataset consists of MIDI encoded songs taken from the Geerdes pop music database.[28] Each full length song has been professionally transcribed and may be considered an accurate symbolic representation of the commercially released

[28] http://www.midimusic.de/

audio. MIDI files are used to avoid potential noise that may be introduced by processing audio recordings of music directly. Furthermore, to avoid issues of metrical induction, notated tempi and time-signatures, together with event onsets, are used to construct metrical instances. For the purpose of this study we concentrate on percussion tracks only, encoded as MIDI channel 10. The entire database contains over 14000 songs, predominantly covering mainstream pop music. The six dance genres selected represent the largest subset ($n$ = 195) of dance genres meeting the following criteria:

1. constant tempo;
2. constant time-signature; and
3. contain a MIDI channel 10 percussion track.

Strict event quantisation is not required of the corpus, as a quantisation process is carried out during the generation of attentional energy profiles (described below). Nonetheless, 51 out of the total 195 MIDI files are strictly quantised, and thus contain no expressive performance timing. Tempo ranges across the corpus from 70 bpm to 225 bpm ($M$ = 135.19, $SD$ = 33.34). The number of pieces belonging to each genre are listed in table 5.3. The full listing of pieces used is available in appendix E.

Table 5.3: Overview of the genre classification dataset.

| Genre | $n$ |
|---|---|
| Bolero | 35 |
| Cha-cha | 37 |
| Pasodoble | 40 |
| Rumba | 40 |
| Swing | 33 |
| Waltz | 10 |
| Total | 195 |

Metrical concepts represented in both METRE-P and METRE-S spaces represent stable patterns of periodic attentional energy. Therefore, each piece must be mapped to a single point in each space, representing its characteristic metrical-rhythmic structure. This is achieved by first constructing a symbolic metrical tree, $\mathcal{T}$, for each piece of music, and then projecting each tree to a point in geometrical space according to the definitions in sections 5.4.1 and 5.5.1.

The hierarchical structure of a metrical tree $\mathcal{T}$ is simply determined by the notated time-signature. As the test data consists of professionally encoded MIDI files, we are able to rely on conventional interpretations of the encoded time-signatures

to indicate basic metrical structure. All time-signatures in the dataset indicate isochronous tactus cycles ($\frac{2}{4}$, $\frac{3}{4}$, $\frac{4}{4}$, $\frac{6}{4}$, and $\frac{6}{8}$), the cardinality of which is determined by the associated number of beats per bar.

The only potentially ambiguous time-signature present in this dataset is $\frac{6}{4}$, the grouping structure of which could be interpreted as 3+3 or 2+2+2 tactus beats. On manual inspection, the piece has a strong waltz feel, and the grouping structure was manually annotated as 3+3 (i.e. a 6-cycle tactus cycle with a grouping structure akin to two bars of $\frac{3}{4}$).

The first level of tactus subdivision is assumed to be either duple or triple, depending on whether the time-signature indicates simple or compound time. A second level of subdivision is assumed to be duple, providing it does not fall below the 100 ms pulse IOI threshold. Tempo is extracted from each MIDI file, from which beat duration in milliseconds is derived, and is assumed to indicate the tactus pulse rate. All metrical timing values are derived from the timing of the tactus cycle, in conformance with metrical well-formedness. Metrical structures instantiated in this manner are therefore prototypical in terms of their timing.

The attentional energy values associated with each N-cycle timepoint are intended as a representation of a purely cognitive phenomenon. Whilst the relative emphasis given to individual timepoints within metres is intuitively a salient concept for differentiating amongst experiences of metrical music, as discussed in section 5.3.3 there are potentially many higher-level musical and lower-level auditory factors one might hypothesise as being involved in establishing such a concept. The pragmatic alternative employed here is to derive attentional energy values directly from the musical surface representation, defining a measure based on the number of musical event onsets that coincide with each metrical N-cycle timepoint. 'Coincident' is used here so as to not imply any particular interpretation of the causal relationship between metrical entrainment and sequences of musical events. Although for the purpose of this evaluation attentional energy is directly linked to event stimuli, the cognitive representations themselves in principle imply no distinction between metrical entrainment experienced in conjunction with physical musical events and entirely imagined musical experiences with no physical correlate.

The implication of the non-dynamic nature of the formalism pursued here is that attentional energy represents an averaged notion of rhythmic articulation—similar (but not exactly the same, as will become evident below) to a histogram of musical events with bins centred on N-cycle pulse onsets modulo the measure period. For musical extracts shorter than the extent of the perceptual present,

this definition is reasonable because the representations are not aimed at encoding explicit sequential rhythmic structure, but rather associated periodic patterns of attentional energy. For longer extracts this assumption becomes questionable, particularly if rhythmic structure varies considerably, when arguably metrical attentional energy is also likely to shift. Such dynamic processes could be modelled as time-dependant paths within conceptual space, but this is left for future work.

The N-cycle provides the categorical timepoints, represented by set $O$, with which musical events coincide, modulo the measure period. Events falling within a symmetrical, empirically-determined time window of each N-cycle timepoint are considered temporally coincidental. We take $\pm 20$ ms of a pulse onset to be a reasonable lower bound for this time window, based on evidence provided by Hirsh (1959) indicating that this is the minimum IOI necessary for listeners to reliably discern the correct ordering of two successive onsets. We assume an upper bound of $\pm 50$ ms, as this would allow the unambiguous association of event onsets with pulse onsets at the fastest pulse IOI rate of $100$ ms. A time window of $\pm 30$ ms is used in this experiment, which amounts to 95% of all events being assigned to a metrical category. Future work should address the question of generating representations of temporally individuated metrical concepts sensitive to the micro-timing of musical events. Such work would also allow for more realistic assumptions to be made concerning the boundaries between metrical categories, in line with findings such as those reported by Desain and Honing (2003) and Repp et al. (in press).

For each piece, the number of events coinciding with each timepoint is counted, and the totals normalised to unit range by dividing by the greatest value. Normalisation allows distributions of attentional energy to be comparable across different pieces of music irrespective of the number of events they contain. The normalised values are taken as indicative of attentional energy. Events that do not coincide with an N-cycle timepoint are assumed to be non-metrical, or possibly strong instances of performance variation. Therefore, within this framework rhythmic figures such as triplets in simple-time metres do not contribute to the calculation of attentional energy. This is reasonable given that our representation aims to capture metrical concepts, rather than rhythmic articulation that may play against an established metrical framework.[29]

[29] A simple extension to the representations tested here would be to allow duple and triple relationships to be represented simultaneously, within distinguished dimensions. This would allow 'against the metre' accent to be represented as part of a more holistic metrical-rhythmic concept.

### 5.6.3 Results

For each reported result, the null hypothesis assumed, $H_0$, is that the performance between model pairs $A$ and $B$ is the same. Following Bouckaert (2004), we use a paired *t*-test to compute a *t*-statistic $Z$, and calculate the *p*-value as the probability $p(Z)$ that $Z$, or less, is observed assuming $H_0$ is true. $H_0$ is rejected if $p(Z) < \alpha/2$ or $p(Z) > 1 - \alpha/2$. The former indicates that model $B$ outperforms $A$, the latter that $A$ outperforms $B$.

Before considering the classification performance of the individual representations, we first consider the overall results with respect to the nearest neighbour parameter $k$, to assess the impact on classification. For all models, classification accuracy improved as $k$ increased. The difference in improvement was significant ($\alpha$ = 0.01) according to paired *t*-tests between all corresponding pairs of models in $k$ = 1 and $k$ = 3 runs. The difference between $k$ = 3 and $k$ = 4 was less pronounced and only significant for the three METRE-S models. On further inspection, the relative performance of all models across the three parameter values of $k$ remained consistent, and did not alter the significance of the differences between competing models. Therefore, we can be confident that the parameterisation of the classification algorithm is not biasing the comparative evaluation. All following results will be reported for $k$ = 3. Optimised domain salience weights, complete 10x10 sorted-runs accuracy data and overall confusion matrices from each METRE-P and METRE-S $k$ = 3 classification run can be found in appendix F.

Table 5.4 presents the mean classification accuracy for each model over the 10x10 cross validation scheme. The TEMPO model, providing classification based purely on given tempo, has reasonable accuracy given the simplicity of the representation, and in comparison to the naïve baseline of 21.51%, based on simply assigning each piece to the genre with the largest number of examples in the dataset. All conceptual space models perform significantly better than TEMPO ($\alpha$ = 0.001), as we would hope given the relative complexity of the geometrical representations.

Figure 5.5 presents the accuracy over the 10x10cv classification runs of METRE-P and METRE-S side-by-side conditioned on the vector space norm. From the plot, the effect of distance metric does not appear strong, similar to the finding with melodic similarity in chapter 4. However, within each norm condition, there is a distinct improvement in the accuracy achieved by METRE-S compared with that of METRE-P.

Table 5.4: Mean classification accuracy over 10x10 cross validation.

| Model | Mean Accuracy % | Std |
|---|---|---|
| (TEMPO, $k$ = 3) | 48.39 | 9.96 |
| (METRE-P, $L^1$, $k$ = 3) | 74.95 | 8.89 |
| (METRE-P, $L^1 + L^2$, $k$ = 3) | 76.46 | 8.63 |
| (METRE-P, $L^2$, $k$ = 3) | 76.32 | 8.83 |
| (METRE-S, $L^1$, $k$ = 3) | 78.54 | 8.11 |
| (METRE-S, $L^1 + L^2$, $k$ = 3) | 80.80 | 8.01 |
| (METRE-S, $L^2$, $k$ = 3) | 79.40 | 8.96 |

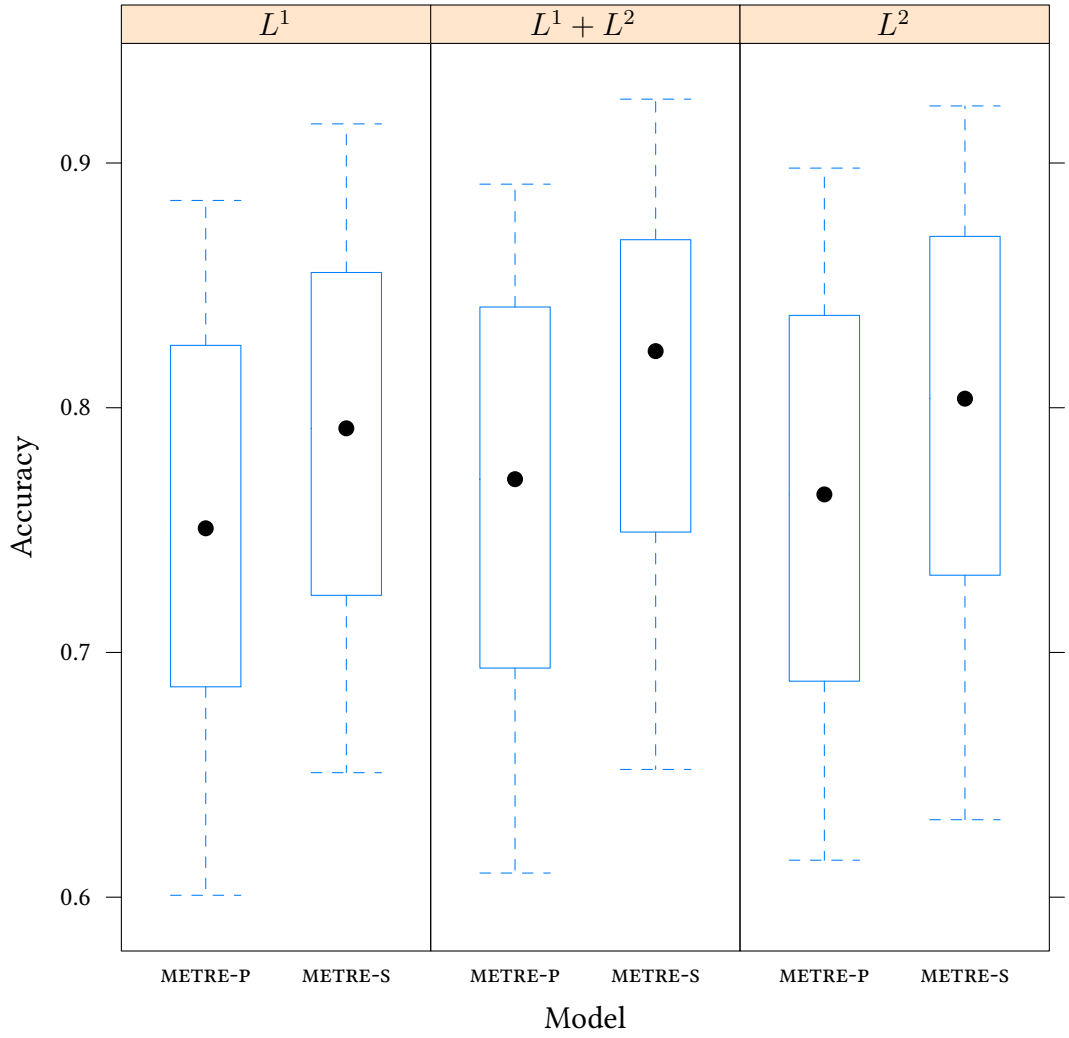



Figure 5.5: Pairwise comparison of METRE-P and METRE-S accuracy conditioned on distance metric.

All pairwise comparisons between METRE-P and METRE-S, within norm conditions, are significant, as shown in table 5.5. Significance codes are used to indicate the level at which the difference between model performance is considered significant: *** $\alpha$ = 0.001; ** $\alpha$ = 0.01; and * $\alpha$ = 0.1. As noted above, for a difference to be considered significant at level $\alpha$, the *p*-value must be less than $\alpha/2$, or greater than $1 - \alpha/2$.

Table 5.5: METRE-P VS. METRE-S within norm conditions.

| Models compared | Norm | Paired *t* statistic | Model $B$ outperforms $A$ |
|---|---|---|---|
| METRE-P VS. METRE-S | $L^1$ | $t(9)$ = -10.57, p < 0.001 | *** |
| METRE-P VS. METRE-S | $L^1 + L^2$ | $t(9)$ = -11.55, p < 0.001 | *** |
| METRE-P VS. METRE-S | $L^2$ | $t(9)$ = -8.02, p < 0.001 | *** |

Looking in more detail at the effect of the vector space norms, table 5.6 shows the pairwise comparison of METRE-P accuracy in each norm condition. For METRE-P, the difference between $L^1$ and $L^2$ was the most significant, in favour of $L^2$. The difference between $L^1$ and $L^1 + L^2$ was also significant at $\alpha = 0.01$.[30] The difference between both METRE-P models using the $L^2$ norm, either exclusively or only within domains, was not significant. Therefore, although the $L^1 + L^2$ METRE-P model was the best performing METRE-P model overall, we cannot conclude from these results that it is a more appropriate configuration than $L^2$. However, the results do suggest that using a Euclidean norm is more advantageous than city-block for this geometric representation.

Table 5.6: Comparison of norms within METRE-P.

| Norms compared | Paired *t* statistic | Model $B$ outperforms $A$ |
|---|---|---|
| $L^1$ vs. $L^1 + L^2$ | $t(9)$ = -4.45, $p$ < 0.001 | ** |
| $L^1$ vs. $L^2$ | $t(9)$ = -8.66, $p$ < 0.001 | *** |
| $L^2$ vs. $L^1 + L^2$ | $t(9)$ = -0.52, $p$ = 0.308 | |

In the case of METRE-S, each pairwise comparison of accuracy in the different norm conditions is significant at the lowest significance level $\alpha$ = 0.1, as shown

[30] The *p*-value for the difference in METRE-P $L^1$ vs. $L^1 + L^2$ accuracy to a greater number of decimal places is 0.0008. Therefore, this difference is not significant at $\alpha$ = 0.001 because 0.0008 > 0.001/2.

in table 5.7. In contrast to METRE-P, the difference in performance between $L^1$ and $L^2$, in favour of $L^2$, is the least significant difference here, indicating that $L^2$ does not afford improvement over $L^1$ to the same degree for METRE-S as for METRE-P. Furthermore, the accuracy afforded by the $L^1 + L^2$ norm is significantly improved over that afforded by both $L^1$ ($\alpha$ = 0.001) and $L^2$ ($\alpha$ = 0.01), indicating that a hierarchical vector space where Euclidean distance is used within domains, and city-block between domains, is optimal in this case.

Table 5.7: Comparison of norms within METRE-S.

| Norms compared | Paired $t$ statistic | Model $B$ outperforms $A$ |
|---|---|---|
| $L^1$ vs. $L^1 + L^2$ | $t(9)$ = -5.82, $p$ < 0.001 | *** |
| $L^1$ vs. $L^2$ | $t(9)$ = -2.10, $p$ = 0.032 | * |
| $L^2$ vs. $L^1 + L^2$ | $t(9)$ = -3.57, $p$ = 0.003 | ** |

## 5.7 Conclusion

The ability of the classification models to discriminate between stylistic dance rhythms purely on the basis of metrical-rhythmic structure lends some support to the validity of our proposed conceptual space representations of metre. Furthermore, the labelling of compositions and styles of music as points and regions in geometrical space is itself potentially valuable information, as the labels themselves are very generic, and yet afford a degree of interpretation in terms of spatial metaphors. However, the dataset used within the evaluation was small by comparison to typical MIR standards, and what more, this evaluation method can only be considered a proxy for a controlled psychological evaluation, which must be conducted before any firm claims concerning the validity of the models can be made.

Additional data was collected during the classification task, available in appendix F, which may shed light on which of the components of the representations are most useful in achieving the clustering. Furthermore, the confusion matrices are deserving of musicological consideration, which may be able to offer novel perspectives on the distribution of musical concepts within the conceptual spaces. The visualisation of semantic distance embedded within METRE-P and METRE-S between a small number of prototypical metrical structures can be found in appendix G. These visualisations hint at a possibly fruitful avenue of research

in seeking to develop mappings between the very explicit and high-dimensional representations of metrical structure developed here, and spaces that are much more readily comprehensible, and thus more compatible with Gärdenfors' theory of conceptual space.

Not withstanding future psychological investigation, the proposed representations raise some interesting questions. Posing a representation in terms of geometry affords particular ways of thinking about and manipulating the objects of the representation. We have only considered the more familiar concepts of metre in terms of the spaces defined. Yet the definition of the spaces explicitly aims to encompass all theoretically possible metrical structures that can be derived from the first principles of perceptual and physiological limitations. Familiar metrical structures represent relatively sparse regions of the spaces, which begs the question of what, if anything, do other regions of the space correspond to perceptually. It is trivial to compute the centroid between two metrical concepts, but it is not obvious whether such a metrical structure is perceptually equidistant. This is particularly apparent in pulse IOI dimensions, when departure from isochrony may manifest in ways that may not even be considered metrical at all.[31]

[31]The method of filtered point-symmetry developed by Plotkin (2010) would make for an interesting comparison with ideas of trajectory and transformation in conceptual spaces.

# Chapter 6

# Conclusions

Three geometrical approaches to musical representation have been presented, each addressing an important issue in the application of computational techniques to furthering understanding of music and musical processes. Within each topic area considered, an effort was made to incorporate cognitive principles in defining the questions that were to be addressed, and where possible, in the development of theoretical frameworks and consequent practical implementations and methods of evaluation. The premise for this work is simple: music is foremost a psychological phenomenon, and seeking to study music with computational means can not only benefit from an informed psychological underpinning, but also contribute significantly to furthering our understanding of music.

The contributions of this work are as follows:

**Metre space**

- A symbolic formalisation of a prominent psychological theory of metre, making it amenable to computation.
- Two mappings from the symbolic-level representation of metrical structure to geometrical representations, providing spaces embedding different qualities hypothesised to be salient in the perception and cognition associated with metrical entrainment.

**Melodic similarity**

- A method of predicting human judgements of melodic similarity employing the Earth Mover's Distance metric over a novel perceptually-motivated ground distance space.

**Pattern discovery**

- A proposed search heuristic for identifying salient musical patterns amongst the exhaustive set of patterns discovered by the SIATEC algorithm.

The application of cognitively-motivated representations in modelling music and musical behaviour demonstrates a method for approaching music research. Only a small number of musical questions have been addressed here, but they may serve as examples for future work. Furthermore, by employing Gärdenfors' framework of conceptual space to the modelling of complex concepts, this work is also a contribution to the theory of conceptual representation.

The application of the conceptual spaces framework to problems of music representation has proved to be informative, and the affordances of geometrical methods attractive for the modelling of similarity. The problem of melodic similarity reported in this dissertation utilises a very simple geometric representation, and yet the ability to scale dimensions, transforming the embedded notion of similarity, offers a simple and intuitive means of modelling a highly subjective concept. However, limits to the potential usefulness of geometrical models, at least those of the form developed here, are evident. Simple linear dimensions of absolute time and pitch work well for local comparisons, but do not account well for longer-term relationships across time or larger pitch ranges. Involving dimensions modelling the relative, first-order relationships between perceptual qualities proved beneficial, and it would seem likely that more cognitively informed geometrical structures, such as multidimensional domains of tonality or metrical time, may assist further in the modelling and understanding of melodic similarity.

The modelling of time in particular, and more generally sequence, within geometric representations is deserving of further consideration. Statistical models can be very effective in modelling sequential information, and have a particular virtue of being able to learn incrementally, constructing abstractions over linear sequences in time. This virtue is lacking in the present models of melodic similarity, and the spaces of metrical structure, both of which assume fixed spaces within which musical concepts can be identified. Further research considering the respective strengths and weaknesses, and possible integration, of geometrical and statistical approaches is certainly warranted.

Specific implications arising from the developed conceptual space models of metre concern the psychological investigation of rhythmic similarity. The spaces themselves represent testable theories of perceptual similarity, and interesting work awaits to be done investigating the fit, or not as the case may be, between

the geometrical properties of the spaces and human perception.

In the area of computational creativity, the subject that initially motivated this thesis, the potential for future work investigating the application of perceptually-grounded representations within systems of artificial creative agents is great. A major problem that must be addressed within systems simulating creative behaviour is how to equip agents with sufficiently rich representations of objects from the domain in which they operate. In a musical context, the developed spaces of metrical-rhythmic structures not only provide a rich representation of important musical concepts, but also one in which traversal of the space can be defined in musically meaningful terms. For example, we have shown that regions of the metre spaces can be correlated with the rhythmic characteristics of different musical genres. Therefore, this knowledge could inform the generation of new music appropriate to a particular style, or even to explore the boundaries between genre regions in the creation of hybrid styles.

Both perceptual and engineering challenges are presented by the potential for developing sub-symbolic levels of representation that could underpin the geometrical spaces of metre. An acknowledged limitation of the developed spaces is their dislocation from real-time perceptual input. It may prove possible to connect signal processing techniques designed for beat tracking and metrical induction with the conceptual spaces of metre, thus grounding the models in physical musical stimuli. In turn, the conceptual spaces could offer additional scope for improving the performance of existing signal processing methods by leveraging the conceptual knowledge embedded within the geometry. Similar challenges lie in the opposite direction, in seeking lower-dimensional projections of the spaces that, while may have to sacrifice the current high level of explicitness, may potentially be considerably more intuitive.

A related issue, at a different temporal level, concerns the representation of larger-scale musical structures. The `SIA` family of algorithms offers methods for discovering larger-scale patterns across symbolically represented musical works. Relationships between patterns are expressed as vectors in a space. Therefore, future work in this area might usefully consider whether these vectors may be grounded in a conceptual space as a means of characterising the relationships between patterns within musically salient terms of reference. Similarly, within the spaces of metrical structure we have not considered the representation of trajectories through space, which given the potential for the hierarchical construction of conceptual spaces advocated by Gärdenfors, may well consist of further higher-level geometrical constructs.

# Appendix A

# Notational conventions

| | |
|---|---|
| $S = \{\ldots\}$ | the set $S$ |
| $S \times S'$ | the Cartesian product of $S$ and $S'$ |
| $\lvert S \rvert$ | the cardinality of $S$ |
| $\emptyset$ | the empty set |
| $\mathbb{R}$ | real numbers |
| $\mathbb{R}^+$ | positive real numbers |
| $\mathbb{R}^k$ | $k$-dimensional real vector space |
| $\mathbb{Z}$ | integer numbers |
| $\mathbb{Z}^+$ | positive integer numbers |
| $\mathbb{N}$ | non-negative integer numbers |
| $[x, y]$ | inclusive real-number interval between $x$ and $y$ |
| $[x..y]$ | inclusive integer-number interval between $x$ and $y$ |
| $\mathbf{v} = \langle\ldots\rangle$ | the vector $\mathbf{v}$ |
| $\mathbf{M} = [m_{ij}]$ | the matrix $\mathbf{M}$ |
| $\mathbf{m}_i^j = \langle e_1, e_2, \ldots, e_j \rangle$ | the ordered sequence of length $j \in \mathbb{Z}^+$, indexed by $i \leq j$ |
| $\parallel$ | tuple concatenation: $\langle 0, 1\rangle \parallel \langle 2, 3\rangle \rightarrow \langle 0, 1, 2, 3\rangle$ |
| $\top$ | the symbol denoting undefined |

# Appendix B

# Müllensiefen and Frieler (2004) melodic similarity dataset

Table B.1: Müllensiefen and Frieler (2004) dataset used in the evaluation of EMD-based models of melodic similarity. Source: Müllensiefen (2004).

| Artist | Title | Composer | Year | Source |
|---|---|---|---|---|
| Demis Roussos | Goodby My Love, Goodybye | Mario Panas | 1973 | Hits der 70er, KDM-Verlag, 2000 |
| Backstreet Boys | As Long As You Love me | Martin Sandberg | 1997 | Hits der 80er und 90er, KDM-Verlag, 1999 |
| Wolfgang Petry | Augen zu und durch | Petry, Valance and Ackermann | 1997 | Hits der 80er und 90er, KDM-Verlag, 1999 |
| Passion Fruit | Wonderland | | 2000 | Transcription by Müllensiefen |
| Kosmonova | Danse avec moi | | 2000 | Transcription by Müllensiefen |
| Wes Montgomery | Bumpin' on sunset | Wes Montgomery | 1966 | Transcription by Pogoda |
| Aquagen | Summer is calling | | 2002 | Transcription by Müllensiefen |
| The Beatles | From me to you | Paul McCartney | 1963 | The New Beatles Complete, Wise Publications, 1992 |
| Die Nilsen Brothers | Aber Dich gibt's nur einmal für mich | Pit | 1965 | 100 Hits in C-Dur, Musikverlag Monika Hildner |
| Die Kolibris | Die Hände zum Himmel | W. van Nimwegen | 1998 | 100 Hits in C-Dur, Musikverlag Monika Hildner |
| Bing Crosby | Swanee River | Traditional | | Hits und Songs, Edition Metropol Köln, 1982 |
| | Sailor | Russian Traditional | | From Sloboda and Parker (1985), after O'Toole (1974) |
| Peter Maffay | Du | Peter Orloff | 1969 | Peter Maffay: Heute vor 30 Jahren, Bosworth Edition, 2001 |
| Backstreet Boys | I want it that way | Andreas Carlson | 1999 | Hits der 80er und 90er, KDM-Verlag, 1999 |

# Appendix C

# Optimised EMD model parameters

Table C.1: Optimised EMD model parameters.

| model | ONSET | CPITCH$_c$ | CPITCH | DUR | CPINT | IOI | exponent |
|---|---|---|---|---|---|---|---|
| (ONSET × CPITCH$_c$, $L^1$, $P$) | 0.21 | 1.00 | | | | | 0.28 |
| (ONSET × CPITCH, $L^1$, $P$) | 0.11 | | 1.00 | | | | 0.30 |
| (ONSET × CPITCH$_c$, $L^1$, $C$) | 0.21 | 1.00 | | | | | 0.39 |
| (ONSET × CPITCH, $L^1$, $C$) | 0.13 | | 1.00 | | | | 0.39 |
| (ONSET × CPITCH$_c$, $L^2$, $P$) | 0.24 | 1.00 | | | | | 0.28 |
| (ONSET × CPITCH, $L^2$, $P$) | 0.12 | | 1.00 | | | | 0.30 |
| (ONSET × CPITCH$_c$, $L^2$, $C$) | 0.24 | 1.00 | | | | | 0.40 |
| (ONSET × CPITCH, $L^2$, $C$) | 0.15 | | 1.00 | | | | 0.39 |
| (ONSET × CPITCH$_c$, $L^1$, $P_d$) | 0.24 | 1.00 | | | | | 0.33 |
| (ONSET × CPITCH, $L^1$, $P_d$) | 0.17 | | 1.00 | | | | 0.33 |
| (ONSET × CPITCH$_c$, $L^1$, $C_d$) | 0.22 | 1.00 | | | | | 0.35 |
| (ONSET × CPITCH, $L^1$, $C_d$) | 0.16 | | 1.00 | | | | 0.35 |
| (ONSET × CPITCH$_c$, $L^2$, $P_d$) | 0.26 | 1.00 | | | | | 0.33 |
| (ONSET × CPITCH, $L^2$, $P_d$) | 0.18 | | 1.00 | | | | 0.32 |
| (ONSET × CPITCH$_c$, $L^2$, $C_d$) | 0.24 | 1.00 | | | | | 0.37 |
| (ONSET × CPITCH, $L^2$, $C_d$) | 0.17 | | 1.00 | | | | 0.35 |
| (ONSET × CPITCH$_c$ × DUR, $L^1$, $P$) | 0.17 | 1.00 | | 0.12 | | | 0.37 |
| (ONSET × CPITCH × DUR, $L^1$, $P$) | 0.10 | | 1.00 | 0.03 | | | 0.32 |
| (ONSET × CPITCH$_c$ × DUR, $L^1$, $C$) | 0.18 | 1.00 | | 0.14 | | | 0.46 |
| (ONSET × CPITCH × DUR, $L^1$, $C$) | 0.13 | | 1.00 | 0.06 | | | 0.41 |
| (ONSET × CPITCH$_c$ × DUR, $L^2$, $P$) | 0.24 | 1.00 | | 0.15 | | | 0.39 |
| (ONSET × CPITCH × DUR, $L^2$, $P$) | 0.12 | | 1.00 | 0.03 | | | 0.33 |
| (ONSET × CPITCH$_c$ × DUR, $L^2$, $C$) | 0.21 | 1.00 | | 0.17 | | | 0.49 |
| (ONSET × CPITCH × DUR, $L^2$, $C$) | 0.15 | | 1.00 | 0.08 | | | 0.41 |
| (ONSET × CPITCH$_c$ × CPINT, $L^1$, $P$) | 0.30 | 1.00 | | | 2.24 | | 0.45 |
| (ONSET × CPITCH × CPINT, $L^1$, $P$) | 0.12 | | 1.00 | | 0.57 | | 0.41 |
| (ONSET × CPITCH$_c$ × CPINT, $L^1$, $C$) | 0.26 | 1.00 | | | 1.81 | | 0.63 |
| (ONSET × CPITCH × CPINT, $L^1$, $C$) | 0.13 | | 1.00 | | 0.93 | | 0.58 |
| (ONSET × CPITCH$_c$ × CPINT, $L^2$, $P$) | 0.41 | 1.00 | | | 2.68 | | 0.44 |
| (ONSET × CPITCH × CPINT, $L^2$, $P$) | 0.23 | | 1.00 | | 1.18 | | 0.42 |
| (ONSET × CPITCH$_c$ × CPINT, $L^2$, $C$) | 0.30 | 1.00 | | | 2.09 | | 0.62 |
| (ONSET × CPITCH × CPINT, $L^2$, $C$) | 0.14 | | 1.00 | | 1.02 | | 0.59 |
| (ONSET × CPITCH$_c$ × IOI, $L^1$, $P$) | 0.26 | 1.00 | | | | 0.12 | 0.34 |
| (ONSET × CPITCH × IOI, $L^1$, $P$) | 0.11 | | 1.00 | | | 0.03 | 0.33 |
| (ONSET × CPITCH$_c$ × IOI, $L^1$, $C$) | 0.24 | 1.00 | | | | 0.17 | 0.42 |
| (ONSET × CPITCH × IOI, $L^1$, $C$) | 0.12 | | 1.00 | | | 0.05 | 0.40 |
| (ONSET × CPITCH$_c$ × IOI, $L^2$, $P$) | 0.31 | 1.00 | | | | 0.10 | 0.33 |
| (ONSET × CPITCH × IOI, $L^2$, $P$) | 0.18 | | 1.00 | | | 0.02 | 0.31 |
| (ONSET × CPITCH$_c$ × IOI, $L^2$, $C$) | 0.28 | 1.00 | | | | 0.26 | 0.47 |
| (ONSET × CPITCH × IOI, $L^2$, $C$) | 0.15 | | 1.00 | | | 0.07 | 0.41 |
| (ONSET × CPITCH$_c$ × DUR × CPINT × IOI, $L^1$, $P$) | 0.24 | 1.00 | | 0.03 | 1.85 | 0.26 | 0.51 |
| (ONSET × CPITCH × DUR × CPINT × IOI, $L^1$, $P$) | 0.13 | | 1.00 | 0.01 | 0.82 | 0.15 | 0.47 |
| (ONSET × CPITCH$_c$ × DUR × CPINT × IOI, $L^1$, $C$) | 0.24 | 1.00 | | 0.02 | 1.67 | 0.18 | 0.59 |
| (ONSET × CPITCH × DUR × CPINT × IOI, $L^1$, $C$) | 0.12 | | 1.00 | 0.02 | 0.91 | 0.14 | 0.56 |
| (ONSET × CPITCH$_c$ × DUR × CPINT × IOI, $L^2$, $P$) | 0.25 | 1.00 | | 0.05 | 1.98 | 0.56 | 0.59 |
| (ONSET × CPITCH × DUR × CPINT × IOI, $L^2$, $P$) | 0.19 | | 1.00 | 0.03 | 1.32 | 0.31 | 0.50 |
| (ONSET × CPITCH$_c$ × DUR × CPINT × IOI, $L^2$, $C$) | 0.25 | 1.00 | | 0.03 | 1.98 | 0.52 | 0.69 |
| (ONSET × CPITCH × DUR × CPINT × IOI, $L^2$, $C$) | 0.15 | | 1.00 | 0.01 | 1.09 | 0.17 | 0.62 |

# Appendix D

# Undefined values in conceptual space

One shortcoming from a geometrical perspective of the conceptual space formalisation of metre by Forth et al. (2010) is the presence of undefined values. The motivation behind allowing undefined values was that if a periodic component is notionally not present in the metrical concept, no numeric value within the defined dimensions is appropriate to represent its absence. Therefore, stepping outside a purely geometrical framework, an undefined value was permitted, and an appropriate algebra defined allowing arithmetic operations between defined and undefined values. This algebra states that the difference between defined and undefined values is always some constant $\epsilon$. This allows a distances to be calculated between points in the space which at least takes into account a simple notion of difference between defined and undefined values. However, this is arguably at odds with Gärdenfors' geometric notion of quality dimensions, because there is no meaningful interpretation of betweenness between defined and undefined dimensional values. In a sense the meaning of an undefined value is orthogonal to the meaning represented by the dimension.

To give a more concrete illustration, the notion of the centroid between two points where some values are undefined is problematic. Taking a single P_IOI dimension as an example, if the centroid between 100 ms and an undefined value, $\top$, were defined as undefined, the distance between each value from this "centroid" is not equal, it is $\epsilon$ and zero respectively, and not a point equidistant from both. Defining the centroid in terms of the distance $\epsilon$, perhaps $\frac{\epsilon}{2}$ is equally unsatisfactory, and indeed meaningless because the value no longer represents a time interval, and again the distance between each point and the "centroid" are not equal. Therefore, in order to pursue a purely geometrical representation of metre, an alternative formalisation that does not require undefined values was sought.

# Appendix E

# Geerdes genre classification dataset

Table E.1: Songs used in the evaluation of the conceptual spaces of metrical-rhythmical structure.

| Artist | Title | Genre |
|---|---|---|
| Los Panchos | Contigo En La Distancia | bolero |
| Guerra, Juan Luis | Burbujas De Amor | bolero |
| Ana Belén | Lia | bolero |
| Moncho | Callate | bolero |
| Manzanero, Armando | Que Pasa | bolero |
| Ana Belén | La Mentira | bolero |
| Carrillo, Alvaro | Sabor A Mi | bolero |
| Duarte, Ernesto | Como Fué | bolero |
| Ronstadt, Linda | Quiereme Mucho | bolero |
| Flippers | Der letzte Bolero | bolero |
| Rodriguez, Silvio | Dos Gardenias | bolero |
| Ana Belén & Banderas, Antonio | No Ser Por Que Te Quiero | bolero |
| Victor Manuel | Me Asalto La Primavera | bolero |
| El Consorcio | Camino Verde | bolero |
| Luis Miguel | Inolvidable | bolero |
| Luis Miguel | Mucho Corazón | bolero |
| Machin, Antonio | Corazón Loco | bolero |
| Machin, Antonio | Mira Que Eres Linda | bolero |
| Luis Miguel | Usted | bolero |
| Tamara | Si Nos Dejan | bolero |
| Escobar, Manolo | Boda Blanca | bolero |
| Aguilar, Pepe | Perdoname | bolero |
| Quezada, Milly & Fernandito | Pideme | bolero |
| Estefan, Gloria | Como Me Duele Perderte | bolero |
| Durcal, Rocio | Infiel | bolero |
| El Coyote & Su Banda Tierra Sa | Te Soñé | bolero |
| Aguilar, Pepe | Que sepan todos | bolero |
| Payador, Luis | Dos Besos | bolero |
| Durcal, Rocio | Sombras Nada Mas | bolero |
| Grupo Palomo | No Me Conoces Aún | bolero |
| Zaa, Charlie | Flor Sin Retoño | bolero |
| Tamara | Como Me Gusta | bolero |
| Ainhoa & Beth | Piensa En Mi | bolero |
| Luis Miguel | La Gloria Eres Tu | bolero |
| Estefan, Gloria | Hoy | bolero |
| Modern Romance | Cherry Pink | chacha |
| Perez Prado | Cerezo Rosa | chacha |
| Ana Belén | Derroche | chacha |
| Ronstadt, Linda | Perfidia | chacha |
| Ronstadt, Linda | Piel Canela | chacha |
| Egues, E. | El Bodeguero | chacha |
| Perez Prado | Macarenas (Mambo/Cha-Cha) | chacha |
| Cobos, Luis | Perfidia | chacha |
| Orquestra Mondragón | El Huevo De Colón | chacha |
| Orquestra Plateria | Ligia Helena | chacha |

Continued on next page

Table E.1: Songs used in the evaluation of the conceptual spaces of metrical-rhythmical structure.

| Artist | Title | Genre |
|---|---|---|
| Azucar Moreno | De Lo Que Te Has Perdido | chacha |
| Ronstadt, Linda | Piensa En Mí (Cha-Cha-Version) | chacha |
| Mercader, Frank | Echame A Mi La Culpa | chacha |
| El Consorcio | Cachito Mio | chacha |
| Dann, Georgie | La Gallina Cha Cha Cha | chacha |
| El Consorcio | La Espinita | chacha |
| Aránega, Albert | Me Lo Dijo Adela | chacha |
| Fernades, José | Casa Separa | chacha |
| Rodríguez, José Luis (El Puma) | Esta Mujer Me Mata | chacha |
| Dann, Georgie | Macumba | chacha |
| Victor Manuel | Si Ella No Me Quisiera | chacha |
| Rosana | Pa' Calor | chacha |
| Pimpinela | Caliente, Caliente | chacha |
| Durcal, Rocio | Poquito Olvido | chacha |
| Orquestra Plateria | L'home Dibuixat | chacha |
| Ben Sa Tumba & Son Orchestre | La Banana (El Unico Fruto Del Amor) | chacha |
| Presuntos Implicados | Vereda Tropical | chacha |
| Santana | El Farol | chacha |
| Santana | Primavera | chacha |
| Emmanuel | Corazón De Melao | chacha |
| Iglesias, Julio | Gozar La Vida | chacha |
| La Mosca Tse Tse | Cha cha cha | chacha |
| Moncho & Dyango | Son Cuatro Dias | chacha |
| Orquesta Encantada | Noches De Ipacaraí | chacha |
| Thalia | Tu Y Yo | chacha |
| Banda Del Capitan Canalla | La Loba Feroz | chacha |
| Bublé, Michael | Sway | chacha |
| Traditional | España Cañi (Span.Pasodoble) | pasodoble |
| Escobar, Manolo | Que Viva España | pasodoble |
| Mariano, Luis | Valencia | pasodoble |
| Escobar, Manolo | Solo Te Pido | pasodoble |
| Orquestra Maravella | Islas Canarias | pasodoble |
| Voskuylen, Henry van | Costa Del Sol | pasodoble |
| Jurado, Rocío | Viva El Pasodoble | pasodoble |
| Pascual, Gustavo | Paquito El Chocolatero | pasodoble |
| Pasodoble Popular | Ragón Falez Pasodoble | pasodoble |
| Reina, Juanita | Francisco Alegre | pasodoble |
| Castellanos, C. | La Morena De Mi Copla | pasodoble |
| Carosone, Renato | Torero Y Olé | pasodoble |
| Portela, Raul | Lisboa Antiga | pasodoble |
| Escobar, Manolo | Te Llaman Perla Preciosa | pasodoble |
| Piquer, Conchita | Bandera Roja Y Gualda | pasodoble |
| Valderrama, Juanito | El Emigrante | pasodoble |
| Santiago, Maria José | Quien Dijo Que El Amor No Está De Moda | pasodoble |
| El Consorcio | El Cha Ca Cha Del Tren | pasodoble |
| Orquestra Maravella | El Gallito | pasodoble |
| Escobar, Manolo | Mujeres Y Vino | pasodoble |
| Huelva, Perlita de | Desafío Torero | pasodoble |
| Orquestra Maravella | El Gato Montés | pasodoble |
| Mª Jesús-A.Aránega | En Er Mundo | pasodoble |
| Escobar, Manolo | Madrecita Maria Del Carmen | pasodoble |
| Maria Jesús | Campanera | pasodoble |
| Saldo, Roberto | Sombreros Et Mantilles | pasodoble |
| Escobar, Manolo | Niña Bonita | pasodoble |
| Escobar, Manolo | Mi Barco Velero | pasodoble |
| Madrid, Julio | Tu pelo | pasodoble |
| Farina, Rafael | Salamanca | pasodoble |
| Pimpinela | Pasodoble Te Quiero | pasodoble |
| Traditional | La Passada | pasodoble |
| Cano, Carlos | Chiclanera | pasodoble |
| Tony Bruins Orchestre | Fiesta En La Caleta | pasodoble |
| Buxeda, Ely | Agarrate Saxo | pasodoble |
| Buxeda, Ely | Vaya Saxo | pasodoble |
| Buxeda, Ely | El Saxo Humano | pasodoble |
| Buxeda, Ely | Mi Arma | pasodoble |

Continued on next page

Table E.1: Songs used in the evaluation of the conceptual spaces of metrical-rhythmical structure.

| Artist | Title | Genre |
|---|---|---|
| Last, James | Viva España | pasodoble |
| Conde, Alejandro | 80 Primaveras | pasodoble |
| Gipsy Kings | Baila Me | rumba |
| Gipsy Kings | Sin Ella | rumba |
| Orellana, Raúl | Gipsy Rhythm | rumba |
| Los Payos | Maria Isabel | rumba |
| Rosario | Escucha Primo | rumba |
| Serrat, Joan Manuel | Tocar Madera | rumba |
| Los Valldemosa | Vuelo 502 | rumba |
| Escobar, Manolo | Mi Carro | rumba |
| Veneno, Kiko | Te Echo De Menos | rumba |
| Raphael | Escándalo | rumba |
| Los Manolos | Para Ser Rumbero | rumba |
| Peret | Borriquito | rumba |
| Académica Palanca | Me Llaman Mala Persona | rumba |
| Gipsy Kings | Escucha Me | rumba |
| De Ville, Willy | Demasiado Corazón | rumba |
| Los Del Rio | Aurora | rumba |
| Los Manolos | Una Aventura | rumba |
| Ay Ay Ay | No Trovo Casa Meva | rumba |
| Los Machucambos | Porompompero | rumba |
| Los Del Rio | Clodomiro El Ñajo | rumba |
| Iglesias, Julio | Agua Dulce, Agua Salá | rumba |
| Flores, Antonio | Alba | rumba |
| Los Del Rio | Hey Macarena | rumba |
| Iglesias, Julio | Baila Morena | rumba |
| Azucar Moreno | Moliendo Café | rumba |
| Peret | Gitana Hechicera | rumba |
| Camela | Sueños Inalcanzables | rumba |
| Azucar Moreno | Solo Se Vive Una Vez | rumba |
| Rumba 3 | No Sé No Sé | rumba |
| Rumba 3 | Tengo Lo Que Quiero | rumba |
| Camela | ¿Qué He Conseguido? | rumba |
| Camela | Vivir Por Vivir | rumba |
| Camela | Vuelve Junto A Mi | rumba |
| Requiebros | Caballo De Mis Deseos | rumba |
| Ketama | No Estamos Locos | rumba |
| Rumba 3 | Perdido Amor | rumba |
| Tonino | Trakatra | rumba |
| Niña Pastori | Tu Me Camelas | rumba |
| Niña Pastori | Ese Gitano | rumba |
| Camela | Corazón Indomable | rumba |
| Miller, Glenn | In The Mood | swing |
| Armstrong, Louis | Hello Dolly | swing |
| Miller, Glenn | Chattanooga Choo Choo | swing |
| Miller, Glenn | Tuxedo Junction | swing |
| Miller, Glenn | Take The -A- Train | swing |
| Ellington, Duke | Satin Doll | swing |
| Miller, Glenn | Getting Sentimental Over You | swing |
| Jonasz, M. | Mister Swing | swing |
| Arlen, Harold | It's Only Paper Moon | swing |
| Bernie, Ben | Sweet Georgia Brown | swing |
| Porter, Cole | Night And Day | swing |
| Sinatra, Frank | They Can't Take That Away ... | swing |
| Evans & Reaves | Lady Of Spain :Swing V. | swing |
| McHugh, Jim | On The Sunny Side Of The Street | swing |
| Crosby, Bing | You Must Have Been A Beautiful Baby | swing |
| Armstrong, L. & Fitzgerald, E. | Cheek To Cheek | swing |
| Monroe, Marilyn | I'm Gonna File My Claim | swing |
| Boone, Pat | Crazy Train | swing |
| James, Harry | Memories Of You | swing |
| Caroll, Diannah | Old Friends | swing |
| Davis Jr, Sammy | If My Friends Could See ... | swing |
| Paaske, Erik | Godt Man Er Faerig Med Det | swing |
| Sinatra, Frank | Mack The Knife | swing |

Continued on next page

Table E.1: Songs used in the evaluation of the conceptual spaces of metrical-rhythmical structure.

| Artist | Title | Genre |
|---|---|---|
| Helmer Olesens Orkester | On A Slow Boat To China | swing |
| Miller, Glenn | String Of Pearls | swing |
| James, Etta | At Last | swing |
| Fitzgerald, Ella | Blue Skies | swing |
| Guardiola, José | Mackie El Navaja | swing |
| Haley, Bill | Mambo Rock | swing |
| Kuhn, Paul | Die Farbe der Liebe | swing |
| Grup Cosmos | Buona Sera | swing |
| Los Cinco Latinos | Un Telegrama | swing |
| Barber, Chris | Take Me Back To New Orleans | swing |
| Underhållningsorkestern | Luffarfröjd | waltz |
| Shadows | Autumn | waltz |
| Berry, Dave | Mama | waltz |
| Black, Cilla | Anyone Who Had A Heart | waltz |
| Albrecht, Gaby | Einmal mit dir | waltz |
| Macias, Enrico | Mon Coeur D'Attache | waltz |
| Freber, Jean | Pigalle (Mussette) | waltz |
| Righteous Brothers | Ebb Tide | waltz |
| Peterman, Monic | Nostalgie | waltz |
| Valente, Caterina | Dich werd ich nie vergessen | waltz |

# Appendix F

# Conceptual space genre classification data

## F.1 METRE-P optimised salience weights ($k$ = 3)

Table F.1: Optimised METRE-P domain salience weights for $k$ = 3: mean over all runs of 10x10cv.

| Model | MEAN_P_IOI | MEAN_A_ENERGY | C_RATIO |
|---|---|---|---|
| (METRE-P, $L^1$, $k$ = 3) | 1.000 | 0.794 | 0.704 |
| (METRE-P, $L^1 + L^2$, $k$ = 3) | 1.000 | 0.683 | 0.618 |
| (METRE-P, $L^2$, $k$ = 3) | 1.000 | 0.627 | 0.660 |

## F.2 METRE-S optimised salience weights ($k$ = 3)

Table F.2: Optimised METRE-S domain salience weights for $k$ = 3: mean over all runs of 10x10cv.

| Model | P_IOI | A_ENERGY |
|---|---|---|
| (METRE-S, $L^1$, $k$ = 3) | 1.000 | 0.653 |
| (METRE-S, $L^1 + L^2$, $k$ = 3) | 1.000 | 0.499 |
| (METRE-S, $L^2$, $k$ = 3) | 1.000 | 0.510 |

## F.3 TEMPO classifier results ($k$ = 3)

Table F.3: Accuracy for (TEMPO, $k$ = 3) over sorted 10x10 cross validation.

| | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 | Fold 6 | Fold 7 | Fold 8 | Fold 9 | Fold 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Run 1 | 0.26 | 0.39 | 0.45 | 0.48 | 0.50 | 0.50 | 0.52 | 0.53 | 0.57 | 0.67 |
| Run 2 | 0.35 | 0.39 | 0.43 | 0.44 | 0.47 | 0.47 | 0.50 | 0.57 | 0.60 | 0.67 |
| Run 3 | 0.33 | 0.38 | 0.42 | 0.43 | 0.50 | 0.55 | 0.56 | 0.58 | 0.61 | 0.62 |
| Run 4 | 0.37 | 0.38 | 0.40 | 0.47 | 0.48 | 0.50 | 0.50 | 0.50 | 0.52 | 0.78 |
| Run 5 | 0.26 | 0.33 | 0.38 | 0.45 | 0.50 | 0.52 | 0.56 | 0.60 | 0.61 | 0.68 |
| Run 6 | 0.28 | 0.42 | 0.43 | 0.44 | 0.44 | 0.47 | 0.50 | 0.55 | 0.57 | 0.62 |
| Run 7 | 0.29 | 0.39 | 0.43 | 0.47 | 0.48 | 0.50 | 0.55 | 0.56 | 0.58 | 0.65 |
| Run 8 | 0.28 | 0.38 | 0.43 | 0.44 | 0.45 | 0.47 | 0.48 | 0.60 | 0.67 | 0.68 |
| Run 9 | 0.33 | 0.33 | 0.38 | 0.44 | 0.45 | 0.47 | 0.48 | 0.53 | 0.55 | 0.62 |
| Run 10 | 0.19 | 0.42 | 0.43 | 0.47 | 0.48 | 0.50 | 0.55 | 0.55 | 0.56 | 0.61 |
| Mean | 0.29 | 0.38 | 0.42 | 0.46 | 0.47 | 0.50 | 0.52 | 0.56 | 0.58 | 0.66 |

Table F.4: Confusion matrix for (TEMPO, $k$ = 3) over 10x10 cross validation. Rows refer to labelled genre and columns to predicted genre.

| | Bolero | Cha-cha | Pasodoble | Rumba | Swing | Waltz |
|---|---|---|---|---|---|---|
| Bolero | **145** | 36 | 14 | 71 | 72 | 12 |
| Cha-cha | 10 | **258** | 67 | 0 | 26 | 9 |
| Pasodoble | 9 | 88 | **285** | 18 | 0 | 0 |
| Rumba | 83 | 35 | 56 | **142** | 74 | 10 |
| Swing | 76 | 34 | 30 | 68 | **111** | 11 |
| Waltz | 18 | 15 | 0 | 37 | 28 | **2** |

## F.4 METRE-P classifier results ($k$ = 3)

Table F.5: Accuracy for (METRE-P, $L^1$, $k$ = 3) over sorted 10x10 cross validation.

| | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 | Fold 6 | Fold 7 | Fold 8 | Fold 9 | Fold 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Run 1 | 0.57 | 0.71 | 0.72 | 0.75 | 0.76 | 0.78 | 0.79 | 0.80 | 0.83 | 0.84 |
| Run 2 | 0.60 | 0.63 | 0.67 | 0.67 | 0.72 | 0.80 | 0.83 | 0.84 | 0.86 | 0.89 |
| Run 3 | 0.65 | 0.67 | 0.67 | 0.67 | 0.71 | 0.78 | 0.80 | 0.83 | 0.84 | 0.84 |
| Run 4 | 0.55 | 0.61 | 0.67 | 0.68 | 0.74 | 0.80 | 0.83 | 0.86 | 0.86 | 0.89 |
| Run 5 | 0.63 | 0.67 | 0.68 | 0.71 | 0.71 | 0.72 | 0.78 | 0.83 | 0.90 | 0.90 |
| Run 6 | 0.56 | 0.56 | 0.67 | 0.67 | 0.68 | 0.75 | 0.83 | 0.84 | 0.85 | 0.95 |
| Run 7 | 0.67 | 0.68 | 0.70 | 0.71 | 0.74 | 0.75 | 0.76 | 0.78 | 0.89 | 0.89 |
| Run 8 | 0.57 | 0.60 | 0.71 | 0.74 | 0.75 | 0.78 | 0.78 | 0.81 | 0.83 | 0.84 |
| Run 9 | 0.61 | 0.63 | 0.65 | 0.68 | 0.76 | 0.78 | 0.83 | 0.85 | 0.86 | 0.86 |
| Run 10 | 0.60 | 0.63 | 0.72 | 0.72 | 0.74 | 0.76 | 0.80 | 0.81 | 0.81 | 0.94 |
| Mean | 0.60 | 0.64 | 0.69 | 0.70 | 0.73 | 0.77 | 0.80 | 0.83 | 0.85 | 0.88 |

Table F.6: Confusion matrix for (METRE-P, $L^1$, $k$ = 3) over 10x10 cross validation. Rows refer to labelled genre and columns to predicted genre.

| | Bolero | Cha-cha | Pasodoble | Rumba | Swing | Waltz |
|---|---|---|---|---|---|---|
| Bolero | **224** | 46 | 19 | 30 | 31 | 0 |
| Cha-cha | 23 | **327** | 0 | 10 | 10 | 0 |
| Pasodoble | 3 | 40 | **309** | 29 | 9 | 10 |
| Rumba | 42 | 42 | 61 | **248** | 7 | 0 |
| Swing | 14 | 29 | 1 | 28 | **258** | 0 |
| Waltz | 0 | 0 | 0 | 0 | 6 | **94** |

Table F.7: Accuracy for (METRE-P, $L^1+L^2$, $k$ = 3) over sorted 10x10 cross validation.

| | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 | Fold 6 | Fold 7 | Fold 8 | Fold 9 | Fold 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Run 1 | 0.67 | 0.68 | 0.71 | 0.72 | 0.76 | 0.76 | 0.80 | 0.83 | 0.84 | 0.85 |
| Run 2 | 0.50 | 0.68 | 0.71 | 0.76 | 0.78 | 0.78 | 0.81 | 0.84 | 0.89 | 0.90 |
| Run 3 | 0.67 | 0.71 | 0.72 | 0.75 | 0.76 | 0.78 | 0.78 | 0.80 | 0.84 | 0.84 |
| Run 4 | 0.61 | 0.65 | 0.67 | 0.74 | 0.78 | 0.83 | 0.86 | 0.89 | 0.90 | 0.95 |
| Run 5 | 0.58 | 0.67 | 0.71 | 0.71 | 0.72 | 0.74 | 0.76 | 0.83 | 0.90 | 0.90 |
| Run 6 | 0.61 | 0.62 | 0.67 | 0.71 | 0.75 | 0.79 | 0.79 | 0.83 | 0.85 | 0.90 |
| Run 7 | 0.63 | 0.63 | 0.67 | 0.76 | 0.78 | 0.80 | 0.80 | 0.86 | 0.89 | 0.89 |
| Run 8 | 0.57 | 0.70 | 0.74 | 0.78 | 0.78 | 0.80 | 0.81 | 0.84 | 0.86 | 0.89 |
| Run 9 | 0.61 | 0.63 | 0.65 | 0.68 | 0.71 | 0.78 | 0.81 | 0.83 | 0.86 | 0.90 |
| Run 10 | 0.65 | 0.67 | 0.68 | 0.71 | 0.76 | 0.78 | 0.80 | 0.84 | 0.86 | 0.89 |
| Mean | 0.61 | 0.66 | 0.69 | 0.73 | 0.76 | 0.78 | 0.80 | 0.84 | 0.87 | 0.89 |

Table F.8: Confusion matrix for (METRE-P, $L^1 + L^2$, $k$ = 3) over 10x10 cross validation. Rows refer to labelled genre and columns to predicted genre.

| | Bolero | Cha-cha | Pasodoble | Rumba | Swing | Waltz |
|---|---|---|---|---|---|---|
| Bolero | **228** | 43 | 22 | 26 | 31 | 0 |
| Cha-cha | 29 | **329** | 8 | 3 | 1 | 0 |
| Pasodoble | 3 | 54 | **294** | 30 | 9 | 10 |
| Rumba | 21 | 38 | 64 | **273** | 4 | 0 |
| Swing | 16 | 22 | 3 | 20 | **269** | 0 |
| Waltz | 0 | 0 | 0 | 0 | 2 | **98** |

Table F.9: Accuracy for (METRE-P, $L^2$, $k$ = 3) over sorted 10x10 cross validation.

| | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 | Fold 6 | Fold 7 | Fold 8 | Fold 9 | Fold 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Run 1 | 0.67 | 0.71 | 0.71 | 0.72 | 0.76 | 0.79 | 0.80 | 0.83 | 0.85 | 0.89 |
| Run 2 | 0.55 | 0.63 | 0.71 | 0.76 | 0.76 | 0.78 | 0.83 | 0.84 | 0.89 | 0.90 |
| Run 3 | 0.67 | 0.70 | 0.71 | 0.76 | 0.78 | 0.79 | 0.80 | 0.81 | 0.83 | 0.95 |
| Run 4 | 0.63 | 0.65 | 0.67 | 0.67 | 0.67 | 0.83 | 0.86 | 0.86 | 0.90 | 0.95 |
| Run 5 | 0.62 | 0.63 | 0.67 | 0.67 | 0.72 | 0.76 | 0.83 | 0.89 | 0.90 | 0.90 |
| Run 6 | 0.61 | 0.67 | 0.67 | 0.76 | 0.79 | 0.79 | 0.80 | 0.80 | 0.83 | 0.90 |
| Run 7 | 0.57 | 0.63 | 0.68 | 0.75 | 0.76 | 0.78 | 0.85 | 0.86 | 0.89 | 0.89 |
| Run 8 | 0.57 | 0.65 | 0.72 | 0.74 | 0.78 | 0.81 | 0.81 | 0.83 | 0.84 | 0.85 |
| Run 9 | 0.63 | 0.65 | 0.67 | 0.68 | 0.71 | 0.78 | 0.83 | 0.85 | 0.86 | 0.86 |
| Run 10 | 0.63 | 0.67 | 0.67 | 0.70 | 0.71 | 0.74 | 0.76 | 0.80 | 0.86 | 0.89 |
| Mean | 0.62 | 0.66 | 0.69 | 0.72 | 0.74 | 0.78 | 0.82 | 0.84 | 0.87 | 0.90 |

Table F.10: Confusion matrix for (METRE-P, $L^2$, $k$ = 3) over 10x10 cross validation. Rows refer to labelled genre and columns to predicted genre.

| | Bolero | Cha-cha | Pasodoble | Rumba | Swing | Waltz |
|---|---|---|---|---|---|---|
| Bolero | **219** | 58 | 12 | 29 | 32 | 0 |
| Cha-cha | 27 | **321** | 9 | 9 | 4 | 0 |
| Pasodoble | 4 | 57 | **289** | 30 | 10 | 10 |
| Rumba | 5 | 35 | 67 | **282** | 11 | 0 |
| Swing | 11 | 20 | 8 | 13 | **278** | 0 |
| Waltz | 0 | 0 | 0 | 0 | 1 | **99** |

## F.5 METRE-S classifier results ($k$ = 3)

Table F.11: Accuracy for (METRE-S, $L^1$, $k$ = 3) over sorted 10x10 cross validation.

| | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 | Fold 6 | Fold 7 | Fold 8 | Fold 9 | Fold 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Run 1 | 0.71 | 0.72 | 0.76 | 0.76 | 0.78 | 0.80 | 0.80 | 0.84 | 0.84 | 0.94 |
| Run 2 | 0.65 | 0.68 | 0.71 | 0.72 | 0.79 | 0.81 | 0.83 | 0.85 | 0.86 | 0.94 |
| Run 3 | 0.70 | 0.72 | 0.75 | 0.76 | 0.81 | 0.81 | 0.83 | 0.84 | 0.84 | 0.89 |
| Run 4 | 0.67 | 0.67 | 0.70 | 0.71 | 0.74 | 0.85 | 0.86 | 0.89 | 0.89 | 0.89 |
| Run 5 | 0.62 | 0.68 | 0.71 | 0.75 | 0.78 | 0.79 | 0.83 | 0.85 | 0.86 | 0.89 |
| Run 6 | 0.67 | 0.67 | 0.72 | 0.75 | 0.76 | 0.78 | 0.79 | 0.84 | 0.86 | 0.90 |
| Run 7 | 0.58 | 0.62 | 0.68 | 0.78 | 0.80 | 0.81 | 0.86 | 0.89 | 0.90 | 0.94 |
| Run 8 | 0.65 | 0.67 | 0.72 | 0.76 | 0.79 | 0.83 | 0.84 | 0.85 | 0.89 | 0.90 |
| Run 9 | 0.63 | 0.67 | 0.75 | 0.76 | 0.79 | 0.83 | 0.86 | 0.89 | 0.90 | 0.90 |
| Run 10 | 0.63 | 0.70 | 0.71 | 0.72 | 0.74 | 0.75 | 0.78 | 0.81 | 0.86 | 0.94 |
| Mean | 0.65 | 0.68 | 0.72 | 0.75 | 0.78 | 0.81 | 0.83 | 0.86 | 0.87 | 0.92 |

Table F.12: Confusion matrix for (METRE-S, $L^1$, $k$ = 3) over 10x10 cross validation. Rows refer to labelled genre and columns to predicted genre.

| | Bolero | Cha-cha | Pasodoble | Rumba | Swing | Waltz |
|---|---|---|---|---|---|---|
| Bolero | **226** | 43 | 9 | 25 | 47 | 0 |
| Cha-cha | 21 | **296** | 25 | 20 | 8 | 0 |
| Pasodoble | 7 | 47 | **316** | 10 | 10 | 10 |
| Rumba | 19 | 26 | 38 | **314** | 3 | 0 |
| Swing | 10 | 10 | 12 | 20 | **278** | 0 |
| Waltz | 0 | 0 | 0 | 0 | 0 | **100** |

Table F.13: Accuracy for (METRE-S, $L^1 + L^2$, $k$ = 3) over sorted 10x10 cross validation.

| | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 | Fold 6 | Fold 7 | Fold 8 | Fold 9 | Fold 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Run 1 | 0.71 | 0.72 | 0.76 | 0.78 | 0.80 | 0.84 | 0.85 | 0.86 | 0.89 | 0.95 |
| Run 2 | 0.68 | 0.70 | 0.71 | 0.81 | 0.83 | 0.84 | 0.86 | 0.89 | 0.90 | 0.94 |
| Run 3 | 0.75 | 0.75 | 0.76 | 0.81 | 0.81 | 0.83 | 0.83 | 0.89 | 0.89 | 0.89 |
| Run 4 | 0.61 | 0.65 | 0.71 | 0.76 | 0.84 | 0.85 | 0.86 | 0.89 | 0.94 | 0.95 |
| Run 5 | 0.63 | 0.71 | 0.75 | 0.76 | 0.83 | 0.83 | 0.84 | 0.86 | 0.89 | 0.90 |
| Run 6 | 0.61 | 0.76 | 0.78 | 0.78 | 0.79 | 0.80 | 0.81 | 0.81 | 0.84 | 1.00 |
| Run 7 | 0.63 | 0.67 | 0.75 | 0.76 | 0.78 | 0.79 | 0.85 | 0.89 | 0.90 | 0.94 |
| Run 8 | 0.57 | 0.75 | 0.79 | 0.83 | 0.84 | 0.85 | 0.86 | 0.86 | 0.89 | 0.89 |
| Run 9 | 0.67 | 0.71 | 0.75 | 0.79 | 0.83 | 0.84 | 0.89 | 0.90 | 0.90 | 0.90 |
| Run 10 | 0.65 | 0.68 | 0.72 | 0.79 | 0.81 | 0.81 | 0.83 | 0.85 | 0.86 | 0.89 |
| Mean | 0.65 | 0.71 | 0.75 | 0.79 | 0.82 | 0.83 | 0.85 | 0.87 | 0.89 | 0.93 |

Table F.14: Confusion matrix for (METRE-S, $L^1 + L^2$, $k$ = 3) over 10x10 cross validation. Rows refer to labelled genre and columns to predicted genre.

| | Bolero | Cha-cha | Pasodoble | Rumba | Swing | Waltz |
|---|---|---|---|---|---|---|
| Bolero | **240** | 50 | 9 | 25 | 26 | 0 |
| Cha-cha | 22 | **319** | 13 | 11 | 5 | 0 |
| Pasodoble | 1 | 37 | **339** | 3 | 10 | 10 |
| Rumba | 10 | 29 | 50 | **311** | 0 | 0 |
| Swing | 3 | 22 | 1 | 20 | **284** | 0 |
| Waltz | 0 | 0 | 0 | 0 | 19 | **81** |

Table F.15: Accuracy for (METRE-S, $L^2$, $k$ = 3) over sorted 10x10 cross validation.

| | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 | Fold 6 | Fold 7 | Fold 8 | Fold 9 | Fold 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Run 1 | 0.67 | 0.72 | 0.76 | 0.79 | 0.80 | 0.81 | 0.85 | 0.86 | 0.89 | 0.89 |
| Run 2 | 0.60 | 0.62 | 0.74 | 0.76 | 0.79 | 0.81 | 0.83 | 0.89 | 0.90 | 0.94 |
| Run 3 | 0.70 | 0.71 | 0.80 | 0.83 | 0.83 | 0.83 | 0.84 | 0.86 | 0.86 | 0.89 |
| Run 4 | 0.60 | 0.61 | 0.71 | 0.71 | 0.78 | 0.81 | 0.84 | 0.94 | 0.95 | 0.95 |
| Run 5 | 0.67 | 0.68 | 0.76 | 0.80 | 0.83 | 0.83 | 0.83 | 0.84 | 0.86 | 0.90 |
| Run 6 | 0.61 | 0.67 | 0.75 | 0.76 | 0.76 | 0.78 | 0.84 | 0.84 | 0.90 | 0.95 |
| Run 7 | 0.63 | 0.67 | 0.68 | 0.75 | 0.78 | 0.85 | 0.86 | 0.89 | 0.90 | 0.94 |
| Run 8 | 0.52 | 0.70 | 0.72 | 0.76 | 0.81 | 0.83 | 0.84 | 0.89 | 0.89 | 0.95 |
| Run 9 | 0.67 | 0.68 | 0.70 | 0.81 | 0.83 | 0.84 | 0.86 | 0.86 | 0.89 | 0.90 |
| Run 10 | 0.65 | 0.67 | 0.68 | 0.70 | 0.72 | 0.74 | 0.81 | 0.83 | 0.89 | 0.90 |
| Mean | 0.63 | 0.67 | 0.73 | 0.77 | 0.79 | 0.81 | 0.84 | 0.87 | 0.89 | 0.92 |

Table F.16: Confusion matrix for (METRE-S, $L^2$, $k$ = 3) over 10x10 cross validation. Rows refer to labelled genre and columns to predicted genre.

| | Bolero | Cha-cha | Pasodoble | Rumba | Swing | Waltz |
|---|---|---|---|---|---|---|
| Bolero | **239** | 42 | 15 | 28 | 26 | 0 |
| Cha-cha | 33 | **301** | 10 | 17 | 9 | 0 |
| Pasodoble | 6 | 35 | **335** | 3 | 11 | 10 |
| Rumba | 14 | 23 | 48 | **315** | 0 | 0 |
| Swing | 12 | 21 | 0 | 20 | **277** | 0 |
| Waltz | 0 | 0 | 0 | 0 | 20 | **80** |

# Appendix G

# Low-dimensional projections of distances in conceptual space

The figures below are an attempt to visualise the semantic distances between points in the conceptual spaces of METRE-P and METRE-S. Two small datasets were created, consisting of a variety of metres designed to illustrate different aspects of higher-level conceptual similarity. Pairwise distance matrices were calculated for each dataset, and then projected into 2- and 3-dimensional spaces using *multidimensional scaling* (MDS).[32] Goodness of fit (GOF) is a measure of how well the distances between objects in the lower dimensional projection reflect the original data, where the closer the value to one the better. Stress is a similar measure, except that the closer to zero the better the fit.

Two examples of distances in each space are provided here, which indicate that a plausible notion of conceptual similarity is maintained when comparing various metres at both the same and across a range of tempi. The tactus was held constant for the first dataset visualised in figure G.1 and figure G.3, with the aim of visualising the distances between a range of equal tempo metres common to Western music. The second dataset, visualised in figure G.2 and figure G.4 show how the distance between three simple metres changes over a range of different tempi. Regions corresponding to the three metres are clearly evident in the projection.

[32] 2-dimensional non-metric MDS was carried out using the `isoMDS` fuction from the R MASS package (Venables and Ripley 2002), and 3-dimensional metric MDS using the `cmdscale` function from the R statistical package (R Development Core Team 2010)

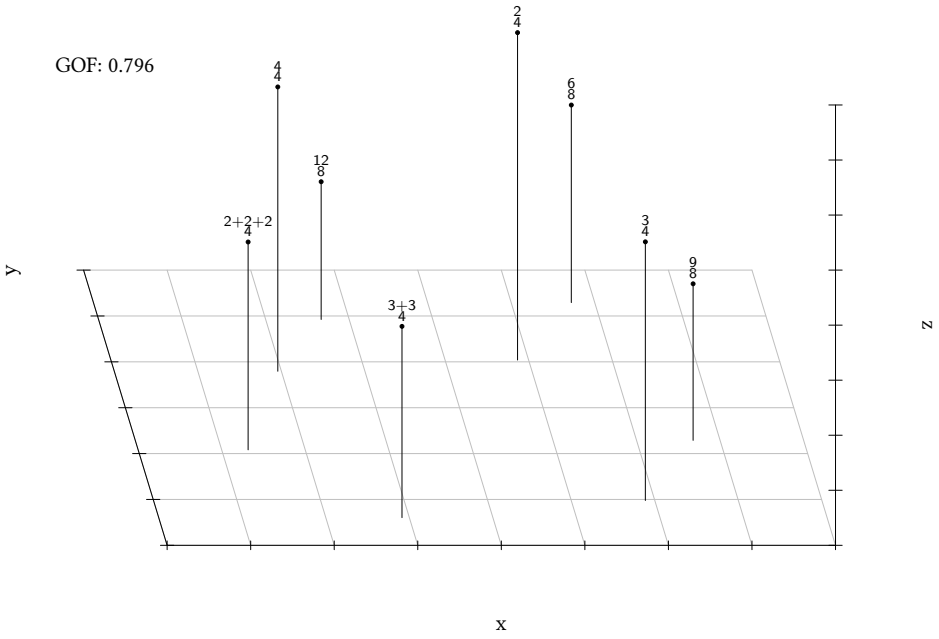



Figure G.1: MDS projection of the distances between prototypical common metres in METRE-P space. All metres are at tactus = 600 ms (100 bpm), and include two levels of tactus subdivision.

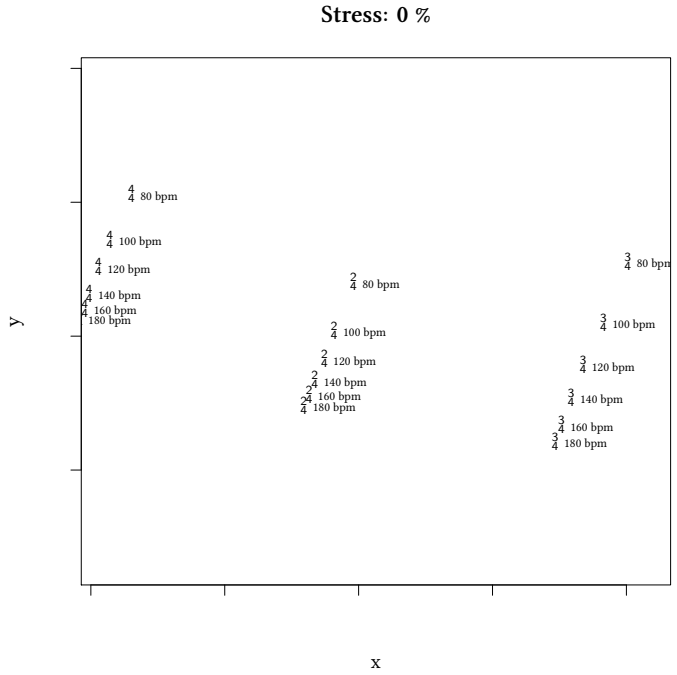



Figure G.2: MDS projection of the distances between prototypical $\frac{2}{4}$, $\frac{3}{4}$ and $\frac{4}{4}$ metres across the tempo range 80–180 bpm in METRE-P space. Each metre has two levels of tactus subdivision.

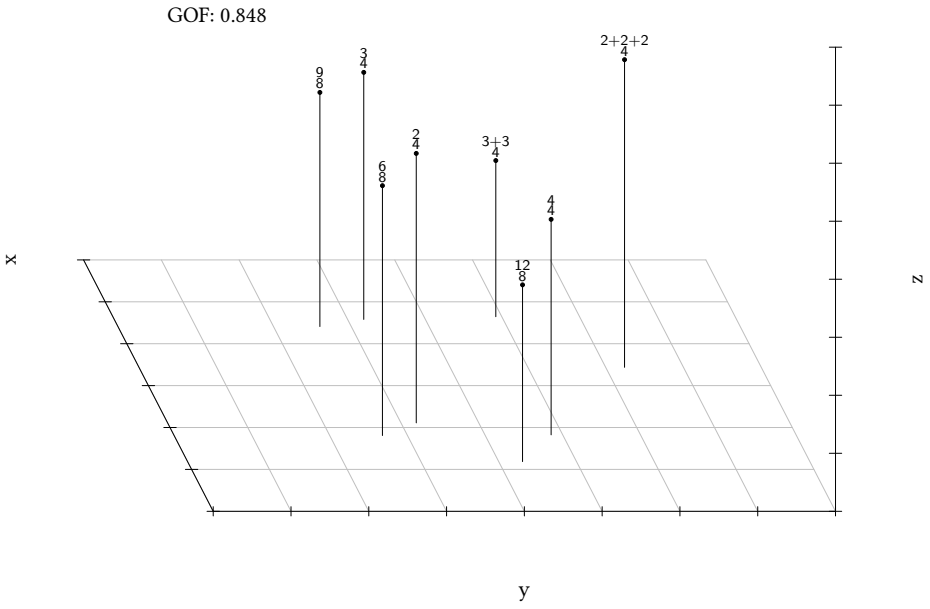



Figure G.3: MDS projection of the distances between prototypical common metres in METRE-S space. All metres are at tactus = 600 ms (100 bpm), and include two levels of tactus subdivision.

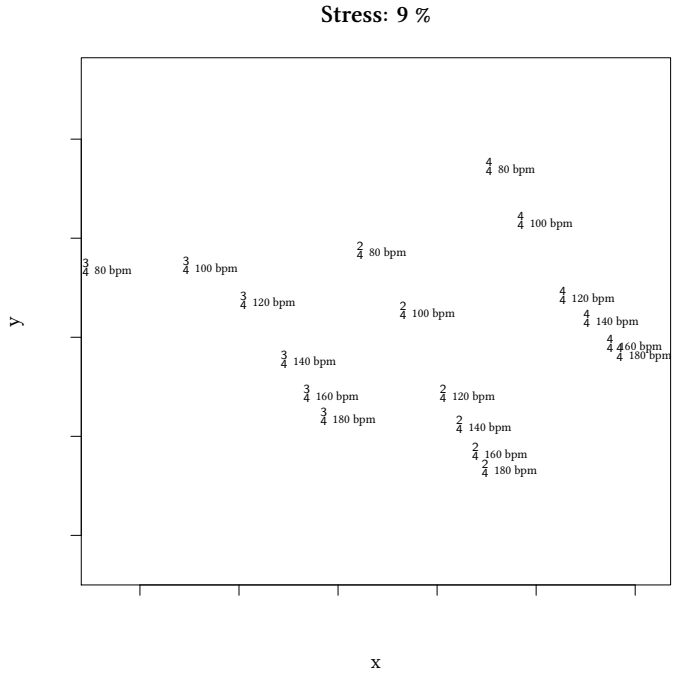



Figure G.4: MDS projection of the distances between prototypical $\frac{2}{4}$, $\frac{3}{4}$ and $\frac{4}{4}$ metres across the tempo range 80–180 bpm in METRE-S space. Each metre has two levels of tactus subdivision.

# Bibliography


Aisbett, J. and G. Gibbon (2001). "A general formulation of conceptual spaces as a meso level representation". In: *Artificial Intelligence* 133.1–2, pp. 189–232. DOI: 10.1016/S0004-3702(01)00144-8.

Andrews, M., G. Vigliocco, and D. Vinson (2009). "Integrating experiential and distributional data to learn semantic representations." In: *Psychological Review* 116.3, pp. 463–498. DOI: 10.1037/a0016261.

Babbitt, M. (1965). "The use of computers in musicological research". In: *Perspectives of New Music* 3.2, pp. 74–83. URL: http://www.jstor.org/stable/832505.

Benjamin, W. E. (1984). "A theory of musical meter". In: *Music Perception* 1.4, pp. 355–413. URL: www.jstor.org/stable/10.2307/40285269.

Bouckaert, R. R. (2004). "Estimating replicability of classifier learning experiments". In: *Proceedings of the 21st international conference on Machine learning (ICML 2004)*. New York, NY, US: ACM. DOI: 10.1145/1015330.1015338.

Bown, O. and G. A. Wiggins (2009). "From maladaptation to competition to cooperation in the evolution of musical behaviour". In: *Musicæ Scientiæ* 13.2 (supplement): *Special Issue: Evolution of Music*, pp. 387–411.

Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA, US: MIT Press.

Brochard, R., D. Abecasis, D. Potter, R. Ragot, and C. Drake (2003). "The 'ticktock' of our internal clock: Direct brain evidence of subjective accents in isochronous sequences". In: *Psychological Science* 14, pp. 362–366.

Caprara, A., M. Fischetti, and P. Toth (1998). *Algorithms for the Set Covering Problem*. Tech. rep. OR-98-3. Bologna, IT: DEIS-Operations Research Group, University of Bolognia.

Chew, E. (2004). "Measuring musical dissimilarity: First and second order center of effect (CE) differences in the spiral array". In: *Proceedings of the 1001th Meeting of the American Mathematical Society: Special Sessions on Mathematical Techniques in Musical Analysis*.

Chvatal, V. (1979). "A greedy heuristic for the set-covering problem". In: *Mathematics of Operations Research* 4.3, pp. 233–235. URL: http://www.jstor.org/stable/3689577.

Clarke, E. F. (1987). "Levels of structure in the organization of musical time". In: *Contemporary Music Review* 2.1, pp. 211–238. DOI: 10.1080/07494468708567059.

— (1999). "Rhythm and timing in music". In: *Psychology of Music*. Ed. by D. Deutsch. 2nd ed. San Diego, CA, US: University of California, pp. 473–500.

Clayton, M., R. Sager, and U. Will (2005). "In time with the music: the concept of entrainment and its significance for ethnomusicology". In: *European meetings in ethnomusicology* 11: *ESEM Counterpoint 1*, pp. 3–75.

Conklin, D. and I. H. Witten (1995). "Multiple viewpoint systems for music prediction". In: *Journal of New Music Research* 24.1, pp. 51–73.

Cook, N. (1998). *Music: A Very Short Introduction*. Oxford, UK: Oxford University Press.

Cooper, G. W. and L. B. Meyer (1960). *The Rhythmic Structure of Music*. Chicago, IL, US: University of Chicago Press.

Cormen, T. H., C. E. Leiserson, R. L. Rivest, and C. Stein (2001). *Introduction to Algorithms*. 2nd ed. Cambridge, MA, US: MIT Press.

Desain, P. and H. Honing (1993). "Tempo curves considered harmful". In: *Contemporary Music Review* 7, pp. 123–138. DOI: 10.1080/07494469300640081.

— (2003). "The formation of rhythmic categories and metric priming". In: *Perception* 32.3, pp. 341–365. DOI: 10.1068/p3370.

Dietterich, T. G. (1998). "Approximate statistical tests for comparing supervised classification learning algorithms". In: *Neural Computation* 10, pp. 1895–1923.

Drake, C. (1993). "Reproduction of musical rhythms by children, adult musicians, and adult nonmusicians". In: *Attention, Perception, & Psychophysics* 53.1, pp. 25–33. DOI: 10.3758/BF03211712.

Dreyfus, L. (1996). *Bach and the Patterns of Invention*. Cambridge, MA, US: Harvard University Press.

Feige, U. (July 1998). "A threshold of $\ln n$ for approximating set cover". In: *Journal of the ACM* 45.4, pp. 634–652.

Forth, J., A. McLean, and G. A. Wiggins (2008). "Musical creativity on the conceptual level". In: *Proceedings of the 5th International Joint Workshop on Computational Creativity*. Ed. by P. Gervás, R. Pérez y Pérez, and T. Veale, pp. 21–30.

Forth, J. and G. A. Wiggins (2009). "An approache for identifying salient repetition in multidimensional representations of polyphonic music". In: *London Algorithmics 2008: Theory and Practice*. Ed. by J. Chan, J. W. Daykin, and M. S. Rahman. Texts in Algorithmics. London, UK: College Publications, pp. 44–58.

Forth, J., G. A. Wiggins, and A. McLean (2010). "Unifying conceptual spaces: Concept formation in musical creative systems". In: *Minds and Machines*, pp. 11–30. DOI: 10.1007/s11023-010-9207-x.

Fraisse, P. (1978). "Time and rhythm perception". In: *Handbook of Perception*. Ed. by E. Carterette and M. Friedmans. Vol. 2. New York, NY, US: Academic Press, pp. 203–254.

Fremerey, C., F. Kurth, M. Müller, and M. Clausen (2007). "A demonstration of the SyncPlayer system". In: *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007)*, pp. 131–132.

Gabrielsson, A. (1993). "The complexities of rhythm". In: *Psychology and music: The understanding of melody and rhythm*. Ed. by T. J. Tighe and W. J. Dowling. Hillside, NJ, US: Lawrence Erlbaum Associates, Inc., pp. 93–120.

Gärdenfors, P. (2000). *Conceptual Spaces: The geometry of thought*. Cambridge, MA, US: MIT Press.

Gärdenfors, P. (2004). "How to make the semantic web more semantic". In: *Proceedings of the 3rd International Conference on Formal Ontology in Information Systems (FOIS 2003)*. Ed. by A. Variz and L. Vieu. Frontiers in Artificial Intelligence and Applications. Amsterdam, NL: IOS Press, pp. 17–34.

Gibson, J. J. (1966). *The Senses Considered as Perceptual Systems*. Boston, MA, US: Hougton Mifflin.

Godøy, R. I. (2003). "Motor-mimetic music cognition". English. In: *Leonardo* 36.4, pp. 317–319. URL: http://www.jstor.org/stable/1577332.

Gouyon, F., S. Dixon, E. Pampalk, and G. Widmer (2004). "Evaluating rhythmic descriptors for musical genre classification". In: *Proceedings of the AES 25th International Conference*, pp. 196–204.

Grachten, M., J. L. Arcos, and R. L. de Mántaras (2005). "Melody retrieval using the Implication/Realization Model". In: *The First Annual Music Information Retrieval Evaluation eXchange (MIREX 2005)*.

Grube, M. and T. D. Griffiths (2009). "Metricality-enhanced temporal encoding and the subjective perception of rhythmic sequences". In: *Cortex* 45.1: *The Rhythmic Brain*, pp. 72–79. DOI: 10.1016/j.cortex.2008.01.006.

Harris, M., A. Smaill, and G. A. Wiggins (1991). "Representing music symbolically". In: *IX Colloquio di Informatica Musicale, Genoa, Italy*. Ed. by A. Camurri and C. Canepa. Genova, IT: Universita di Genova.

Hasty, C. F. (1997). *Metre as Rhythm*. Oxford, UK: Oxford University Press.

Hewlett, W. B. (1992). "A base-40 number-line representation of musical pitch notation". In: *Musikometrika* 4, pp. 1–14.

Hillier, F. and G. Lieberman (1990). *Introduction to Mathematical Programming*. New York, NY, US: McGraw-Hill.

Hirsh, I. J. (1959). "Auditory perception of temporal order". In: *Journal of the Acoustical Society of America* 31.6, pp. 759–767.

Hitchcock, F. (1941). "The distribution of a product from several sources to numerous localities". In: *Journal of Mathematics and Physics* 20, pp. 224–230.

Honing, H. (1993). "Issues on the representation of time and structure in music". In: *Contemporary Music Review* 9.1, pp. 221–238. DOI: 10.1080/07494469300640461.

Huron, D. (1992). "Design principles in computer-based music representation". In: *Computer Representations and Models in Music*. Ed. by A. Marsden and A. Pople. London, UK: Academic Press, pp. 5–39.

— (2006). *Sweet Anticipation: Music and the psychology of expectation*. Cambridge, MA, US: MIT Press.

Iversen, J. R., B. H. Repp, and A. D. Patel (2009). "Top-down control of rhythm perception modulates early auditory responses". In: *Annals of the New York Academy of Sciences* 1169.1, pp. 58–73. DOI: 10.1111/j.1749-6632.2009.04579.x.

Jain, A., K. Nandakumar, and A. Ross (2005). "Score normalization in multimodal biometric systems". In: *Pattern Recognition* 38.12, pp. 2270–2285. DOI: 10.1016/j.patcog.2005.01.012.

Johnson, D. S. (1973). "Approximation algorithms for combinatorial problems". In: *Proceedings of the 5th Annual ACM Symposium on Theory of Computing (STOC 1973)*. New York, NY, US: ACM, pp. 38–49. DOI: 10.1145/800125.804034.

Jones, M. R. (1981). "Only time can tell: On the topology of mental space and time". In: *Critical Inquiry* 7, pp. 557–576. URL: http://www.jstor.org/stable/1343118.

Kahneman, D. (1973). *Attention and Effort*. Englewood Cliffs, NJ, US: Prentice Hall.

Karp, R. M. (1972). "Reducibility among combinatorial problems". In: *Complexity of computer computations*. Ed. by R. Miller and J. Thatcher. New York, NY, US: Plenum Press, pp. 85–103.

Kirkpatrick, S., C. D. Gelatt, and M. P. Vecchi (1983). "Optimization by simulated annealing". In: *Science* 220.4598, pp. 671–680. DOI: 10.1126/science.220.4598.671.

Krumhansl, C. L. (1997). "Effects of perceptual organization and musical form on melodic expectancies". In: *Music, Gestalt, and Computing. Studies in Cognitive and Systematic Musicology*. Ed. by M. Leman. Vol. 1317. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 294–320. DOI: 10.1007/BFb0034122.

— (2005). "The geometry of musical structure: A brief introduction and history". In: *Computers in Entertainment* 3.4, pp. 1–14. DOI: 10.1145/1095534.1095542.

Kvifte, T. (2004). "Description of grooves and syntax/process dialectics". In: *Studia Musicologica Norvegica* 30, pp. 54–77.

— (2007). "Categories and Timing: On the Perception of Meter". In: *Ethnomusicology* 51.1, pp. 64–84. URL: http://www.jstor.org/stable/20174502.

Lemström, K. and A. Pienimäki (2007). "On comparing edit distance and geometric frameworks in content-based retrieval of symbolically encoded polyphonic music". In: *Musicæ Scientiæ* 11.1, pp. 135–152.

Lerdahl, F. and R. Jackendoff (1983). *A Generative Theory of Tonal Music*. Cambridge, MA, US: MIT Press.

Lewis, D., R. Woodley, T. Crawford, J. Forth, C. Rhodes, and G. A. Wiggins (2011). "Tools for music scholarship and their interactions: a case study". In: *Proceedings of the Supporting Digital Humanities Conference (SDH 2011)*. Ed. by B. Maegaard. URL: http://sldr.org/SLDRdata/doc/show/copenhagen/SDH-2011/proceedings.html.

London, J. *Rhythm §II: Historical studies of rhythm*. URL: http://www.oxfordmusiconline.com/subscriber/article/grove/music/45963pg2 (visited on Aug. 9, 2009).

— (2004). *Hearing in time: psychological aspects of musical meter*. Oxford, UK: Oxford University Press.

— (2012). *Hearing in time: psychological aspects of musical meter*. 2nd ed. Oxford, UK: Oxford University Press.

Marchiori, E. and A. Steenbeek (1998). "An iterated heuristic algorithm for the set covering problem". In: *Proceedings of the Workshop on Algorithm Engineering*. Ed. by K. Mehlhorn, pp. 155–166.

McKinney, M. F. and D. Moelants (2006). "Ambiguity in Tempo Perception: What Draws Listeners to Different Metrical Levels?" In: *Music Perception* 24.2, pp. 155–166. DOI: 10.1525/mp.2006.24.2.155.

Meredith, D. (2006). "Point-set algorithms for pattern discovery and pattern matching in music". In: *Proceedings of the Dagstuhl Seminar on Content-Based Retrieval*. Ed. by T. Crawford and R. C. Veltkamp. Dagstuhl Seminar Proceedings 06171. Dagstuhl, DE: Internationales Begegnungs- und Forschungszentrum fuer Informatik (IBFI), Schloss Dagstuhl.

Meredith, D., K. Lemström, and G. A. Wiggins (2002). "Algorithms for discovering repeated patterns in multidimensional representations of polyphonic music". In: *Journal of New Music Research* 31.4, pp. 321–345.

— (2003). "Algorithms for discovering repeated patterns in multidimensional representations of polyphonic music". In: *Cambridge Music Processing Colloquium 2003*. Department of Engineering, University of Cambridge. Cambridge, UK.

Moelants, D. (2002). "Preferred tempo reconsidered". In: *Proceedings of the 7th International Conference on Music Perception and Cognition*. Ed. by C. Stevens, D. Burnham, G. McPherson, E. Schubert, and J. Renwick. Adelaide, AU: Causal Productions, pp. 580–583.

Müllensiefen, D. (2004). "Variabilität and konstanz von melodien in der erinnerung: ein beitrag zur musikpsychologischen gedächtnisforschung". PhD thesis. Hamburg, DE: Institute of Musicology, University of Hamburg.

— (2009). *FANTASTIC: Feature ANalysis Technology Accessing STatistics (In a Corpus)*. Tech. rep. Version 1.5. London, UK: Centre for Cognition, Computation and Culture, Goldsmiths, University of London.

Müllensiefen, D. and K. Frieler (2004). "Cognitive adequacy in the measurement of melodic similarity: Algorithmic vs. human judgments". In: *Computing in Musicology* 13, pp. 147–176.

Nadeau, C. and Y. Bengio (2003). "Inference for the generalization error". In: *Machine Learning* 52.3, pp. 239–281. DOI: 10.1023/A:1024068626366.

Nattiez, J.-J. (1990). *Musicologie générale et sémiologue*. Music and Discourse: Towards a Semiology of Music, trans. by Carolyn Abbate. Princeton NJ, US: Princeton University Press.

Orio, N. (2005). "Combining multilevel and multifeature representation to compute melodic similarity". In: *The First Annual Music Information Retrieval Evaluation eXchange (MIREX 2005)*.

O'Toole, L. M. (1974). *The Gateway Russian Song Book*. London, UK: Collets.

Parncutt, R. (1994). "A perceptual model of pulse salience and metrical accent in musical rhythms". In: *Music Perception* 11, pp. 409–464.

Patel, A. D. (2008). *Music, Language, and the Brain*. Oxford, UK: Oxford University Press.

Pearce, M. T. and G. A. Wiggins (2004). "Improved methods for statistical modelling of monophonic music". In: *Journal of New Music Research* 33.4, pp. 367–385.

— (2006). "Expectation in melody: The influence of context and learning". In: *Music Perception* 23.5, pp. 377–405.

Peleg, S., M. Werman, and H. Rom (1989). "A unified approach to the change of resolution: Space and gray-level". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 11, pp. 739–742.

Pele, O. and M. Werman (2008). "A linear time histogram metric for improved SIFT matching". In: *Proceedings of the 10th European Conference on Computer Vision: Part III (ECCV 2008)*. Berlin, DE: Springer-Verlag, pp. 495–508. DOI: 10.1007/978-3-540-88690-7_37.

— (2009). "Fast and robust earth mover's distances". In: *IEEE 12th International Conference on Computer Vision (ICCV 2009)*, pp. 460–467. DOI: 10.1109/ICCV.2009.5459199.

Peña, E. A. and E. H. Slate (2006). "Global validation of linear model assumptions". In: *Journal of the American Statistical Association* 101.473, pp. 341–354. DOI: 10.1198/016214505000000637.

Plotkin, R. J. (2010). "Transforming transformational analysis: Applications of filtered point-symmetry". PhD thesis. Chicago, IL, US: The University of Chicago.

Polak, R. (2010). "Rhythmic feel as meter: Non-isochronous beat subdivision in jembe music from Mali". In: *Music Theory Online* 16.4. URL: http://www.mtosmt.org/issues/mto.10.16.4/mto.10.16.4.polak.html (visited on Nov. 5, 2011).

Pryer, A. *Notation*. Ed. by A. Latham. URL: http://www.oxfordmusiconline.com/subscriber/article/opr/t114/e4761 (visited on Apr. 26, 2010).

Raubal, M. (2004). "Formalizing conceptual spaces". In: *Proceedings of the 3rd International Conference on Formal Ontology in Information Systems (FOIS 2004)*. Ed. by A. Varzi and L. Vieu. Vol. 114. Frontiers in Artificial Intelligence and Applications. Amsterdam, NL: IOS Press, pp. 153–164.

— (2008a). "Cognitive modeling with conceptual spaces". In: *Workshop on Cognitive Models of Human Spatial Reasoning, Freiburg, Germany*. Ed. by M. Ragni, H. Schultheis, and T. Barkowsky, pp. 7–11.

— (2008b). "Representing concepts in time". In: *Spatial Cognition VI. Learning, Reasoning, and Talking about Space*. Ed. by C. Freksa, N. Newcombe, P. Gärdenfors, and S. Wölfl. Vol. 5248. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 328–343. DOI: 10.1007/978-3-540-87601-4_24.

R Development Core Team (2010). *R: A language and environment for statistical computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, AT. URL: http://www.R-project.org.

Repp, B. H., J. London, and P. E. Keller (in press). "Distortions in reproduction of two-interval rhythms: When the 'attractor ratio' is not exactly 1:2". In: *Music Perception*.

Revelle, W. (2011). *psych: Procedures for psychological, psychometric, and personality research*. R package version 1.0-97. Northwestern University. Evanston, Illinois, US. URL: http://personality-project.org/r/.

Rickard, J. T. (1, 2006). "A concept geometry for conceptual spaces". In: *Fuzzy Optimization and Decision Making* 5.4, pp. 311–329. DOI: 10.1007/s10700-006-0020-1.

Rickard, J. T., J. Aisbett, and G. Gibbon (2007a). "Knowledge representation and reasoning in conceptual spaces". In: *IEEE Symposium on Foundations of Computational Intelligence (FOCI 2007)*, pp. 583–590. DOI: 10.1109/FOCI.2007.371531.

— (2007b). "Reformulation of the theory of conceptual spaces". In: *Information Sciences* 177.21, pp. 4539–4565. DOI: 10.1016/j.ins.2007.05.023.

Rubner, Y., C. Tomasi, and L. J. Guibas (2000). "The earth mover's distance as a metric for image retrieval". In: *International Journal of Computer Vision* 40.2, pp. 99–121.

Salzberg, S. (1997). "On comparing classifiers: Pitfalls to avoid and a recommended approach". In: *Data Mining and Knowledge Discovery* 1.3, pp. 317–328. DOI: 10.1023/A:1009752403260.

Schaefer, R., R. Vlek, and P. Desain (2011). "Decomposing rhythm processing: electroencephalography of perceived and self-imposed rhythmic patterns". In: *Psychological Research* 75.2, pp. 95–106. DOI: 10.1007/s00426-010-0293-4.

Schwering, A. and M. Raubal (2005a). "Measuring semantic similarity between geospatial conceptual regions". In: *GeoSpatial Semantics*. Ed. by M. Rodríguez, I. Cruz, S. Levashkin, and M. Egenhofer. Vol. 3799. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 90–106. DOI: 10.1007/11586180_7.

— (2005b). "Spatial relations for semantic similarity measurement". In: *Perspectives in Conceptual Modeling*. Ed. by J. Akoka, S. W. Liddle, Y. Song, M. Bertolotto, I. Comyn-Wattiau, W.-J. Heuvel, M. Kolp, J. Trujillo, C. Kop, and H. C. Mayr. Vol. 3770. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 259–269. DOI: 10.1007/11568346_28.

Selfridge-Field, E. (1997). "Introduction: Describing musical information". In: *Beyond MIDI: The handbook of musical codes*. Ed. by E. Selfridge-Field. Cambridge, MA, US: MIT Press.

Shepard, R. N. (1962a). "The analysis of proximities: Multidimensional scaling with an unknown distance function. I." In: *Psychometrika* 27.2, pp. 125–140. DOI: 10.1007/BF02289630.

— (1962b). "The analysis of proximities: Multidimensional scaling with an unknown distance function. II". In: *Psychometrika* 27.3, pp. 219–246. DOI: 10.1007/BF02289621.

— (1982). "Structural representations of musical pitch". In: *Psychology of Music*. Ed. by D. Deutsch. New York, NY, US: Academic Press, pp. 343–390.

— (1987). "Toward a Universal Law of Generalization for Psychological Science". In: *Science*. New Series 237.4820, pp. 1317–1323. URL: http://www.jstor.org/stable/1700004.

Sloboda, J. A. and D. H. H. Parker (1985). "Immediate recall of melodies". In: *Musical structure and cognition*. Ed. by R. West, P. Howell, and I. Cross. London, UK: Academic Press, pp. 143–167.

Smaill, A., G. A. Wiggins, and M. Harris (1993). "Hierarchical music representation for composition and analysis". In: *Computers and the Humanities* 27.1, pp. 7–17. DOI: 10.1007/BF01830712.

Smaill, A., G. A. Wiggins, and E. R. Miranda (1993). "Music representation – between the musician and the computer". In: *World Conference on AI and Education workshop on Music Education*. Edinburgh, UK.

Snyder, J. S. and E. W. Large (2005). "Gamma-band activity reflects the metric structure of rhythmic tone sequences". In: *Cognitive Brain Research* 24.1, pp. 117–126. DOI: 10.1016/j.cogbrainres.2004.12.014.

Steedman, M. J. (1977). "The perception of musical rhythm and metre". In: *Perception* 6.5, pp. 555–569. DOI: 10.1068/p060555.

Suyoto, I. S. H. and A. L. Uitdenbogerd (2005). "Simple efficient n-gram indexing for effective melody retrieval". In: *The First Annual Music Information Retrieval Evaluation eXchange (MIREX 2005)*.

Trevarthen, C. (1999-2000). "Musicality and the intrinsic motive pulse: Evidence from human psychology and infant communication". In: *Musicæ Scientiæ: Special Issue: Rhythms, Musical Narrative, and the Origins of Human Communication*, pp. 155–215.

Typke, R. (2007). "Music retrieval based on melodic similarity". PhD thesis. Utrecht, NL: Department of Computer Science, Utrecht University.

Typke, R., M. den Hoed, J. de Nooijer, F. Wiering, and R. C. Veltkamp (2005). "A ground truth for half a million musical incipits". In: *Journal of Digital Information Management* 3.1, pp. 34–39.

Ukkonen, E., K. Lemström, and V. Mäkinen (2003). "Sweepline the music!" In: *Computer Science in Perspective*. Ed. by R. Klein, H.-W. Six, and L. Wegner. Vol. 2598. Lecture Notes in Computer Science. Berlin, DE: Springer-Verlag, pp. 330–342.

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S*. Fourth. ISBN 0-387-95457-0. New York, NY, US: Springer. URL: http://www.stats.ox.ac.uk/pub/MASS4.

Wiggins, G. A. (2008). "Computer-representation of music in the research environment". In: *Modern Methods for Musicology: Prospects, Proposals and Realities*. Ed. by T. Crawford and G. Lorna. Digital Research in the Arts and Humanities. Farnham, UK: Ashgate, pp. 7–22.

Wiggins, G. A., M. Harris, and A. Smaill (1989). "Representing music for analysis and composition". In: *Proceedings of the Second IJCAI Workshop on Artificial Intelligence and Music*. Ed. by M. Balaban, K. Ebcioglu, O. Laske, C. Lischka, and L. Sorisio. Detroit, MI, US, pp. 63–71.

Wiggins, G. A., K. Lemström, and D. Meredith (2002). "SIA(M)ESE: An algorithm for transposition invariant, polyphonic content-based music retrieval". In: *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR 2002)*. Ed. by M. Fingerhut, pp. 283–284.

Wiggins, G. A., E. Miranda, A. Smaill, and M. Harris (1993). "A framework for the evaluation of music representation systems". In: *Computer Music Journal* 17.3, pp. 31–42. URL: http://www.jstor.org/stable/3680941.

Wishart, T. (1996). *On Sonic Art*. Edited by Simon Emmerson. London, UK: Routledge.

Yang, J. and J. Y.-T. Leung (2005). "A generalization of the weighted set covering problem". In: *Naval Research Logistics* 52.2, pp. 142–149. DOI: 10.1002/nav.1009.