Rhythmic Information as a Relevance Criterion for Music Information Retrieval

David M. Weigl

Doctor of Philosophy

School of Information Studies

McGill University

Montreal,Quebec

2016-07-17

A thesis submitted to McGill University in partial fulfilment of the requirements of the degree of Doctor of Philosophy

DEDICATION

To Lewis Arran.

ACKNOWLEDGEMENTS

First and foremost, I thank my supervisor, Catherine Guastavino, for her continued guidance, support, and patience over the years. Catherine's contributions, from discussions on the intricacies of experimental protocols and statistical analyses, to funding arrangements, to the provision of baby clothes and toys, have been instrumental in making the outcomes of this work, and my ability to conduct it, possible. I also thank the faculty at the School of Information Studies, in particular Profs. Jamshid Beheshti and Charles-Antoine Julien; as well as faculty in other departments at McGill, particularly Profs. Daniel Levitin, Ichiro Fujinaga, and Stephen McAdams, for valuable advice and discussions.

Thanks to the members of my (extended) PhD cohort at SIS, and to friends and colleagues at CIRMMT, including (in no particular order) Jonathan Dorey, Prasun Lala, Guillaume Boutard, Daniel Steele, Rhiannon Gainor, Dhary Abuhimed, Isabelle Lamoureux, Nouf Khashman, David Romblom, David Sears, Jason Hockman, Andrew Hankinson, Trevor Knight, Alastair Porter, Cory McKay, and many others who I am clumsily neglecting here, for contributing to my intellectual perspective as the ideas represented in this dissertation took shape; and to my colleagues at the Oxford e-Research Centre, particularly Terhi Nurmikko-Fuller, Kevin Page, John Pybus, David De Roure, Alfie Abdul-Rahman, Ramon Granell, Stef Salvini, Graham Klyne, Andrew Hankinson (again!), Mat Wilcoxson, Will Yeomans, Fred Dulwich, Ian Emsley, and Raz Lakhoo, for their support in my continued struggle to finish the damned thing!

I thank my parents, Ruth and Willi Weigl. Without their loving support throughout my life, this work would not have been possible, on a metaphysical, financial, emotional, and intellectual level.

Finally and most importantly, I thank Christine Elaine Morrison Weigl, my wonderful wife and partner, who has rearranged her life to accompany me through time and space, to whom I give my love and my gratitude forever; and Lewis, my little boy and the centre of my universe, to whom this dissertation is dedicated. At long last, they no longer need to share daddy with the thesis monster.

David M. Weigl (Paris, February 16th, 2016)

Contribution of Authors

This document is formatted as a manuscript dissertation and incorporates the following publications:

Chapter 2 is based on a version of the article *Relevance in Music Information Retrieval* by David M. Weigl, Joan Bartlett, and Catherine Guastavino (in preparation). My two co-authors, and my fellow PhD candidate Daniel Steele acting as research assistant, contributed to the article coding effort in this project, covering just over half of the articles between them. The conception of this project, the creation of the various software tools to assist the coding effort, the coding of the remaining articles, the analysis and synthesis of results, and the preparation of the manuscript were undertaken by myself, with guidance from my two co-authors.

Chapter 3 is based on a version of the article *The Role of Rhythmic Information In Melody Identification* by David M. Weigl, Daniel J. Levitin, and Catherine Guastavino (under revisions, Memory & Cognition). We gratefully acknowledge the assistance of student volunteers Lian Francis and Allison Numerow in coding participant responses; Nadine Kroher for her help in determining dynamic time warping (DTW) values for our stimuli; Emilia Gómez and Francisco Gómez for valuable discussions on measuring rhythmic distances; Gregory Pellechi for proof-reading and providing useful comments on a draft version of the article; and Klaus Frieler and Daniel Müllensiefen for granting access to their implementation of the SIMILE software toolbox. All remaining work on this project was undertaken by myself, with guidance from my two co-authors.

Chapter 4 is based on a version of the article *A Convergent-Methods Investigation of Beat Salience in Sensorimotor Synchronization* by David M. Weigl, David Sears, Stephen McAdams, and Catherine Guastavino (in preparation). All experimental work detailed in this chapter took place at the Music Perception and Cognition Laboratory, Music Technology Area, Schulich School of Music, McGill University. The initial project conception, experimental design, and stimulus generation was undertaken in equal contributions by David Sears, Jason A. Hockman, and myself. We gratefully acknowledge the assistance of Bennett Smith, who designed the interfaces used in the experiments; Emily Wilson, for scheduling participants and running the experiments; and Jose R. Zapata, for kindly running the committee-based beat tracker on our stimuli and providing the resulting Mean Mutual Agreement (MMA) values. David Sears conducted the analysis of the preliminary study. All remaining work, including the analysis of the results of the beat induction, sensorimotor synchronization, and beat salience judgement experiments, the correlation with the MMA values, and the preparation of the manuscript, was undertaken by myself with guidance from my co-authors.

All remaining work presented in this dissertation was prepared by myself under the supervision of Catherine Guastavino. Experiments were undertaken with ethics clearance under `REB# 642-0306` (C. Guastavino).

This work is funded in part by: School of Information Studies, McGill University; Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT); Vivi Martin Fellowship; and WYNG Trust Fellowship in Information Studies.

ABSTRACT

Relevance, a notion at the heart of information retrieval (IR), has received prolific attention in the textual IR domain. While the creation of rigorous and practicable theories concerning the nature of relevance has long been identified as a key priority for the field of Music Information Retrieval (MIR), relevance-related research has remained scarce in comparison. This dissertation employs a large-scale systematic analysis of the user-focussed MIR literature to identify different conceptualizations of relevance in a musical context. We establish a broad account of the present state of knowledge in the field by triangulating convergent findings of disparate studies in order to identify areas of commonality, and outline several under-explored areas, pointing the way for future research. We build on this foundation by investigating rhythmic information as a relevance criterion, employing methodologies from music perception and cognition. The role of rhythmic information as a cue for melody identification is examined in a series of experiments employing distorted versions of familiar melodies. Further, we investigate beat salience, a measure of the perceptual prominence of the beat in the context of finding music to move to, employing a convergent-methods approach investigating perceptual beat induction, sensorimotor synchronization, and beat salience judgement. Primary contributions include the reassessment of the role of rhythm in melody identification, underlining its importance in the definition of an experiential criterion of topical relevance; and the establishment of the validity and reliability of beat salience as a situational relevance criterion for use cases involving synchronized movement to music.

ABRÉGÉ

La pertinence, une notion au cœur de la recherche d'information, a été étudiée en profondeur dans le contexte d'information textuelle. Si le développement de théories rigoureuses et applicables sur la pertinence dans un contexte d'information musicale a été identifié depuis longtemps comme une priorité dans le domaine de recherche d'information musicale, les études sur le sujet restent peu nombreuses. Cette thèse vise tout d'abord à identifier les différentes conceptualisations de la notion de pertinence dans un contexte musical au moyen d'une revue systématique à grande échelle des études centrées sur l'utilisateur dans la littérature sur la recherche d'information musicale. La triangulation des résultats issus de différentes études nous permet d'identifier les domaines de convergence, de mettre en évidence les domaines peu étudiés et d'établir ainsi un état des lieux des connaissances et de proposer des perspectives de recherches. À un niveau plus spécifique, nous étudions le rôle de l'information rythmique comme critère de pertinence en utilisant des méthodes dérivées du domaine de la psychologie de la musique. Dans une série d'expériences utilisant des distorsions rythmiques, nous examinons l'importance de l'information rythmique comme indice pour l'identification de mélodies connues. Dans une autre série d'expériences, nous examinons la saillance du beat, une mesure de l'importance perceptive du beat, dans un contexte de synchronisation sur la musique au moyen de mesures perceptives, de mesures de synchronisation sensorimotrice et de jugements sur échelles. Nos contributions nous amènent à réévaluer le rôle du rythme dans l'identification de mélodies, soulignent l'importance du rythme comme critère de

pertinence expérientielle et thématique et permettent d'établir la validité et fiabilité de la saillance du beat comme critère de pertinence situationnelle dans un contexte de synchronisation du mouvement sur la musique.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
Introduction

Relevance, as a concept or set of concepts, has received prolific research attention in the field of information science. Discussions on the nature of relevance and its place at the heart of information retrieval (IR) have been ongoing for decades, and have been traced to the very beginnings of organized academic research into IR (Saracevic, 1975). A large majority of IR systems organize information around textual entities, such as words and phrases (Saracevic, 2007, p. 1931), and so it is not surprising that relevance research has focused overwhelmingly on textual information domains; indeed, so ingrained is the notion that the relevance concepts under discussion operate on textual data, that this assumption is rarely stated explicitly. In comparison, the notion of relevance has received scant attention in music information retrieval (MIR) research. Yet, operationalizing relevance criteria to meaningfully address the music information needs of (potential) users of MIR systems, with sufficient detail to guide the design and implementation of such systems, is a formidable challenge – as we shall see.

This dissertation presents an attempt at understanding the current state of knowledge on relevance in music information research. It is motivated by the *multi-experiential challenge* outlined by J. S. Downie in his early, comprehensive overview of the MIR field:

> "Music ultimately exists in the mind of its perceiver. Therefore, the perception, appreciation, and experience of music will vary not only across the multitudes of minds that apprehend it, but will also vary within each mind as the individual's mood, situation, and circumstances change ... How do we adjust our relevance judgments under this scenario of ever-shifting moods and perceptions? ... The creation of rigorous and practicable theories concerning the nature of experiential similarity and relevance is the single most important challenge facing MIR researchers today." (Downie, 2003, p. 304-306)

The textual domain affords the luxury of lexical meaning, of representational semantics; specific concepts expressed in the words of a query bear an "unquestioned correspondence" with *information*, the explicit goal of conventional IR (Byrd & Crawford, 2002, p. 21). Whilst textual searches are possible in the domain of music, operating on lyric fragments or on bibliographic facets, searches employing or operating on purely *musical* information must forgo such luxuries.

> "Where the poet or playwright can evoke sadness by narrating a recognizably sad story, musicians must create sadness through non-representational sounds. Where a comedian might evoke laughter through parody, wordplay, or absurd tales, musicians must find more abstract forms of parody and absurdity." (Huron, 2006, p. 1)

Acknowledging the experiential, and thus psychological, nature of the problems posed by Downie's multiexperiential challenge, this dissertation draws on approaches informed by the field of music perception and cognition. The experience of music

has been subject to psychology research since at least the time of the Gestalt theorists (e.g., Wertheimer, 1923). The potential benefits of integrating insights from such research have been under discussion in the MIR field since its early days (e.g., Huron, 2000), but rarely was such work actually adopted into MIR studies (Futrelle & Downie, 2002). The dialogue between the two fields remains difficult (Aucouturier & Bigand, 2012), although it is ongoing—evidenced by the annual seminar on cognitively based music informatics research (CogMIR)[1] , first held in 2011.

The experimental work presented in this dissertation bridges these fields by exploring melody identification, beat perception, and sensorimotor synchronization (moving to music) from the perspective of music perception and cognition, and relating outcomes and implications to guide the implementation of MIR systems operationalizing measures of *topical* and *situational* relevance. Here, these terms are defined as per Jansen and Rieh (2010, p. 1525): topical relevance is a measure of "the connection between information objects retrieved and a query submitted" to an IR system; situational relevance is a measure of the user's judgement of "the relationship between information and information need situations in a certain time".

This experimental work is constrained to the investigation of musical information facets relating to the temporal dimension of the music, which will be collectively referred to as *rhythmic information*. Included in these facets are musical parameters such as note event durations, onset times, and metric regularity, and perceptual

[1] http://www.cogmir.org/

parameters such as beat salience and movement affordance (sensorimotor synchronization). Rhythmic information is fundamental to our experience of the music, being a primary component of melody (Whittall, 2011) along with pitch (i.e., the sequence of note heights), and accordingly, providing determinant sensory cues in melody identification (Hébert & Peretz, 1997). Further, rhythmic information drives core MIR processes such as beat tracking, onset detection, tempo estimation, and melodic similarity estimation. However, little research has yet examined algorithmic outcomes of such MIR processes in light of experimental data on the human perception of rhythmic information.

1.1 Research questions

In order to adequately address questions at the intersection of (algorithmic) music information processing, and listener experience, a conceptual relevance framework must be defined for the music information domain; such a framework may then form the grounding for the experimental work outlined above. This requirement motivates the work presented in this dissertation.

Concretely, we examine the following research questions:

1. How may the notion of relevance be conceptualized for music information research? And, building on this conceptualization:
2. What is the role of rhythmic information in melody identification, and what are the implications in formulating an experiential criterion of topical relevance in MIR? And,
3. Can beat salience inform a valid and reliable criterion of situational relevance in MIR?

1.2 Chapter overview

Along with this introduction and the conclusion, this dissertation addresses the research questions posed above in a collection of three manuscripts in Chapters 2–4. Versions of each of these manuscripts are currently under preparation for journal submission; an article based on Chapter 3 is currently under revisions for Memory & Cognition.

The remainder of this introductory chapter presents a brief history of MIR; a reflection on the IR challenges posed by musical in contrast to textual information; and a discussion of the notion of melodies as music information objects in the field of music perception and cognition.

Chapter 2 presents a systematic analysis and synthesis of 159 user studies in the MIR literature, based on an application of Saracevic's stratified model of relevance interactions (Saracevic, 1997; 2007b) to the music information domain. This chapter serves to provide a broad overview of the notion of relevance in MIR, including discussions on the topics studied, and the insights obtained, as well as on the current gaps in our knowledge.

Chapter 3 investigates the role of rhythmic information in melody identification, finding this role to have been underestimated in previous research. The chapter draws implications for the formulation of topical relevance criteria for MIR systems focussing on tasks relating to melody identification (e.g., query-by-humming systems, where the user hums a melody into a microphone, and the system returns matching documents from a musical database).

Chapter 4 presents a series of experiments investigating beat salience, a measure of the perceptual prominence of the beat in music, and evaluates the validity and reliability of this measure as a relevance criterion in the context of finding music to move to.

Chapter 5 concludes this dissertation, drawing together the main findings of the preceeding chapters, summarizing theoretical, methodological, and practical contributions, as well as limitations, and finally, proposing directions for future research.

1.3 A brief history of music information retrieval

The field of music information retrieval (MIR) is highly interdisciplinary in nature, uniting researchers from a diverse range of disciplines in pursuit of the common goal of providing robust, comprehensive access to musical information (Downie, 2004a). The field is rooted in traditional textual information retrieval research, but is also informed by research areas such as music perception and cognition, music theory, signal processing, audio engineering, and computer science.

MIR is a young field; while pioneering academic work can be traced back to the 1960's (Kassler, 1966), literature in the area remains sparse until the late 1990's. As new technologies, such as the MP3 format, plummeting costs of digital storage, and the widespread adoption of the Internet by the public made digital music ubiquitously available, the increased necessity for music information storage and retrieval engaged the interest of researchers from a variety of disciplines. The International Society for Music Information Retrieval (ISMIR) has been the focal point of this renewed research activity; Downie, Byrd, & Crawford (2009) offer an overview of ISMIR's first decade, from 1999–2009.

The origin and growth of ISMIR has been motivated by the textual Information Retrieval (IR) world. Plans for an evaluation platform based on that used by the Text REtrieval Conference (TREC) were under discussion from the beginning (Downie et al., 2009), and eventually led to the creation of MIREX, the Music Information Retrieval Evaluation eXchange (Downie, West, Ehmann, & Vincent, 2005).

The primary emphasis of research in MIR has been placed on the development of MIR systems. Much valuable work has gone into the creation and continued improvement of algorithms to perform tasks integral to MIR, such as onset and key detection, tempo extraction, beat tracking, genre classification, and many others (Downie, 2008). Evaluation metrics are generally applied to system performance parameters such as precision and recall. Formal consideration of user information needs and information behaviour has been sparse in comparison. The situation reflects the early state of research in the field of textual IR, where similar early emphasis on information systems gradually gave way to a more user-centric paradigm (Dervin & Nilan, 1986; Wilson, 1981).

1.4 What challenges does retrieval of music pose in contrast to text retrieval?

Parallels and key difference between the worlds of textual IR and MIR exist on a number of levels. We now consider the relationships between the two fields in order of decreasing abstraction, starting at the broadest level of intellectual property rights and research infrastructure, proceeding via the examination of paradigmatic approaches, to relevance measures, and finally arriving at the level of information representation, querying, and retrieval.

1.4.1 Infrastructure and intellectual property challenges

For practical reasons, collaborative research on information retrieval requires common access to shared collections of information. Without such shared data sets, and without validated and standardized evaluation methodologies, generalizability is threatened; experiments cannot be accurately replicated between research locations, and algorithm performance cannot be fairly evaluated (Voorhees & Harman, 2005). Researchers in the textual retrieval world have access to corpora in numerous languages, extracted from a wide variety of sources (Harman, 2005). The complexities of intellectual property law with regards to musical works (Levering, 2000) prevent the creation and dissemination of analogous corpora for MIR; in the "post-Napster era," holders of musical rights are "notoriously litigious" (Downie, 2004b) (p. 18), and even public domain works cannot be incorporated into shared data sets without considerable legal obstacles (Downie, 2003).

The absence of community-wide music collections poses considerable challenges to the rigorous cross-evaluation of different approaches employed by MIR researchers (Futrelle & Downie, 2002). This need to provide a common ground for evaluation eventually resulted in the creation of MIREX, the Music Information Retrieval Evaluation eXchange (Downie, 2008). MIREX is an evaluation platform inspired by the Text REtrieval Conference (TREC) testing and evaluation paradigm. As with TREC, MIREX functions along a yearly cycle: research tracks focussing on information retrieval tasks that warrant specific investigation are decided by community discussion; independent teams of researchers develop algorithms to address the tasks; the performance of the developed algorithms is evaluated on a shared data set; and

the results are disseminated to the community at large, informing future research and development. In TREC's case, the data set is distributed to each research team; the teams then perform their evaluations independently using identical steps (via evaluation scripts distributed alongside the data), and submit their results to a central contact at TREC. This approach generates only a fairly low administrative overhead, as the evaluation work is distributed among the community. Unfortunately, the threat of litigation prevents such an approach in the world of MIR.

In order to overcome these legal obstacles, MIREX employs a "centralized algorithm-to-data model" (Downie, 2008, p. 249). Rather than distributing evaluation data sets, and the work of running evaluations, to the community, the individual research teams submit their algorithms to a central testing platform located at IMIRSEL, the International Music Information Retrieval Systems Evaluation Laboratory at the University of Illinois at Urbana-Champaign (Downie, Futrelle, & Tcheng, 2004). IMIRSEL houses evaluation data sets (i.e., collections of music and metadata) for the various MIREX tasks. The submitted algorithms are then evaluated locally, and the results are disseminated to the community. This model avoids intellectual property concerns, as the music never leaves the building; however, the burdens of infrastructure requirements and administrative overhead are much larger:

> "the largest amount of effort expended by IMIRSEL on behalf of MIREX is in compiling, debugging, and verifying the output format and validity of submitted algorithms. Collectively, managing and monitoring the algorithms submitted to MIREX consumes nearly a 1,000 person-hours each year." (Downie, 2008, p. 250-251)

The algorithm-to-data model is applicable in an evaluation context; however, local copies of data sets available to individual researchers have their uses, for instance in classifier development and corpus analysis. Furthermore, the size of these data sets matters, as developing with small amounts of data can lead to classifier overfitting, and may not expose relatively rare but interesting phenomena to analysis. The relatively high cost of musical works makes the acquisition of vast quantities for research purposes prohibitively expensive for many researchers, and the sharing of music between labs runs afoul of the "well-known antagonistic stance of the recording industry to the digital sharing of their data" (Bertin-Mahieux, Ellis, Whitman, & Lamere, 2011, p. 592). In recent years, several large corpora have been published that bypass these problems, by distributing a wealth of algorithmically derived or hand annotated metadata, while withholding the music itself. Examples include the McGill Billboard 1,000 corpus (Burgoyne, Wild, & Fujinaga, 2011) and the Million Song Dataset (Bertin-Mahieux, Ellis, et al., 2011).

1.4.2 Paradigmatic challenges

Given the MIR community's emulation of TREC with MIREX, and the strong influence of IR traditions on MIR research, it is perhaps unsurprising that the primary emphasis of research in MIR has been on the development of MIR systems. There have been repeated calls for a greater emphasis on user-centric research in articles discussing the state of the field. Without empirical data regarding users' needs and requirements, music seeking behaviours, and on the usability of MIR interfaces, design decisions are based on "intuitive feelings for user information seeking behaviour," (Cunningham, Reeves, & Britland, 2003) and on "anecdotal evidence

and a priori assumptions of typical usage scenarios" (Lee, 2010). Downie, Byrd, and Crawford (2009) acknowledge that "the knowledge acquired by interacting with users ... can only improve the quality of the community's research output"; such work will "also go a long way to helping ISMIR researchers create truly useful music-IR systems" (p. 17). Some of the earliest papers in the ISMIR proceedings highlight the "need to draw extensively on research in music perception and cognition" (Huron, 2000), acknowledge that "much research has been done on music perception in psychology, music psychology, and cognitive science", but note that "[s]ignificantly, however, MIR researchers have so far rarely adopted work in these areas as a basis of MIR studies" (Futrelle & Downie, 2002, p. 218).

Ingwersen and Järvelin (2005) have noted the dichotomy of IR research, outlining the distinction between systems-oriented IR, and user-oriented and cognitive IR research. They note that research adopting the cognitive viewpoint is "not limited to user-centered approaches to information. Essentially, it is *human*-oriented," involving "humanistic aspects with respect to contents of messages, technological insights of tools for processing, and social scientific dimensions due to the information activities taking place in a social contextual space" (p. 25).

Results from the small pool of existing user studies in MIR (Weigl & Guastavino, 2011) have underlined the importance of such human-oriented research regarding music information needs and behaviours. Several studies have shown the primacy of social factors in the development of musical tastes and the acquisition of new music (Cunningham & Nichols, 2009; Laplante, 2010a; Laplante, 2011); furthermore, studies of peoples' encounters with new music (Cunningham, Bainbridge, & McKay,

2007; Laplante & Downie, 2006; Laplante, 2010b) suggest that music information seeking most commonly occurs as a non-goal oriented activity: "it is mostly the pleasure [participants] take in the activity itself that motivates them to seek for music rather than an actual information need ... most [participants] admitted that they are sometimes so absorbed when they browse for music that they have a problem stopping" (Laplante & Downie, 2006, p. 382).

The sparsity of human-oriented research in the MIR field remains a challenge to the understanding of music information needs, and to the creation of MIR tools that are usable and useful both within and beyond academia. While the "cognitive turn that took place in [textual] IR" (Ingwersen & Järvelin, 2005, p. 3) has not been emulated in the predominantly systems-focussed MIR field, there has been a growing trend toward greater publication of user studies in recent years, although the overall number remains small in terms of overall MIR research output (Lee & Cunningham, 2012).

1.4.3 Experiential similarity and relevance

Among the open research problems of MIR, the definition and operationalisation of experiential similarity and relevance measures—a key research priority for the field (Downie, 2003)—may stand to benefit most from user-focussed research efforts. The concept of relevance is central to both (human) information seeking and (systems-centric) information retrieval research (Jansen & Rieh, 2010). From the systems perspective, IR tools make use of relevance-predicting algorithms to match available information objects to a query in order to generate a set of results. When operationalised as a quantitative metric of the match between a submitted query

and retrieved information objects, the concept is more precisely referred to as *topical* relevance. From the human-oriented perspective of information searching, relevance is a relation between information and contexts (e.g., the user's information needs at a given point in time), based on "some property reflecting a relevance manifestation (e.g., topicality, utility, cognitive match)." Operationalised as the outcome of human judgement of the relation between information and information needs, the concept may be termed *situational* relevance.

Topical relevance is generated deterministically by a given algorithm mapping a given query to set of information objects; thus, it is a static measure. Situational relevance is dynamic; defined by reference to a user's judgement at a given time, it may vary both between individuals, and within individuals, given different contexts. In the domain of music, there are major implementational challenges for both of these relevance concepts.

Topical relevance. Performance measures such as precision and recall, and metrics such as TF·IDF (Zobel & Moffat, 1998) provide implementational approaches for topical relevance in textual IR. Applying such concepts to abstract musical information is problematic. Textual information offers relatively straight-forward units of analysis (e.g., words, or phrase-level chunks); while the words and statements of natural language may be ambiguous, various techniques exist to address such ambiguities (Meadow, Boyce, & Kraft, 2000). Identifying musical correlates is non-trivial. Melodies retain their identity under transposition; thus, any isolated tone could conceivably match any given location in any suitably transposed melody. In other words, "isolated notes have no inherent meaning in a melody identification

task" (Schulkind, Posner, & Rubin, 2003). While a particular musical phrase may acquire well-defined representative meaning through association with a certain character, location, or idea—the concept referred to as 'leitmotif' (Copland, 1957)—such usage is rare within a generally non-representative domain. Accordingly, determining the concrete shapes and boundaries of music information objects is a non-trivial task. A naïve approach defining songs as information objects swiftly runs into problems: musical identity is highly resilient, and a song may exist as a multitude of divergent instances (e.g., alternate takes, live performances, covers by different artists) yet remain fundamentally the same song.

Furthermore, musical experience is inherently multifaceted; when listening to music, various perceptual qualities (pitch, timbre, contour, tempo, rhythm, harmony, loudness) arise through the complex interactions of various physical properties (frequency, amplitude, spectrum, temporal evolution). Musical identity can survive distortions along any of these facets. Melodies reproduced by amateur singers may remain recognizable despite inaccuracies such as altered pitch intervals and distorted rhythms. Trained singers and musicians make controlled use of such distortions, e.g., by applying rubato (variations in tempo) or portamento (sounding of between-note pitches), to affect expressive qualities of a piece. This resilience of recognisability is not limited to popular music. The systematic and creative manipulation of the musical parameters of melodies is at the core of traditional jazz music. Similarly, a number of studies have investigated expressive timing in classical music; a study of the same classical piece by Chopin played by 108 pianists (Repp, 1998) revealed different timing profiles in each take.

Downie refers to this complexity as the *multifaceted challenge* to MIR (2003). Many strands of MIR research, and many MIREX tasks, target specific musical facets in isolation, and much progress has been made toward building algorithmic classifiers in areas such as key finding, beat tracking, tempo extraction and onset detection. However, when it comes to matching the features gleaned through these classifiers to a catalogue of music, the challenges of resilient identity and experiential measures of topical relevance come into play.

Situational relevance. Individuals may interact with music and MIR systems in many different ways, and in many different contexts. Listeners modify their listening habits based on contextual factors, characterizing music by intended use cases—e.g., 'driving music' vs. 'working music' (Cunningham, Jones, & Jones, 2004)—and according to their affective state, to accommodate or to modify their mood (Laplante, 2010b). Furthermore, social factors enormously influence listening behaviour (Taheri-Panah & MacFarlane, 2004; Laplante, 2010b). These are the complexities of the multiexperiential challenge to MIR (Downie, 2003).

Results from studies by Laplante and Downie (2006), and Cunningham, Bainbridge and McKay (2007) suggest that the information behaviour implicit in the most common encounters with new music take a form akin to the "ongoing" mode of searching outlined by Wilson (1997); the information seeker possesses an already-established framework of musical tastes and knowledge, but updates this continuously in order to inform future listening decisions, and to derive pleasure from the process. Opportunistic browsing as outlined by de Bruijn and Spence (2001) also plays a role: passive, serendipitous events during everyday life constitute the majority of people's

encounters with new music (Cunningham et al., 2007). Such results demonstrate the importance of user studies in the creation of workable definitions of situational relevance for music information research.

Clearly, situational aspects of relevance also play a role in textual IR; however, the sparsity of representation and semantics in the abstract realm of music imbues additional importance on information gleaned from contextual and situational factors.

1.4.4 Challenges of information representation, querying, and retrieval

Musical information exists in a large number of different formats. These may be subdivided into symbolic representations, and audio recordings. Symbolic representations of music include printed sheet music, text, specialised notation formats such as guitar tablature, and a number of widespread computer formats such as MIDI, MusicXML, and the Music Encoding Initiative (MEI) framework. Audio representations are encoded in a variety of analogue and digital formats, including phonograph records (LPs), CDs, WAV, and MP3 files. Certain formats arguably involve both classes; for instance, MEI events may include timestamps to align them with a given audio representation of the encoded music, and MIDI files may be turned into audible music using sequencing software.

While textual information may be similarly represented in both symbolic (textual) and audio (recorded speech) formats, querying and indexing is almost always done symbolically. Textual storage for retrieval access is less resource intensive by several orders of magnitude, and users of textual IR systems are generally literate and able to form symbolic textual queries with little difficulty. Conversely, the potential

users of MIR systems cannot be assumed to be musically literate, and may exhibit great difficulty in attempting to formulate symbolic musical queries (Uitdenbogerd & Yap, 2003). Audio queries and representations based on features extracted from audio signals thus play a much larger role in MIR systems.

The difficulty of implementing topical relevance measures complicates retrieval in this context. Audio fingerprinting systems such as the popular commercial music identification software Shazam (Wang, 2006) provide a useful example: while such systems are particularly good at identifying specific recordings (Typke, Wiering, & Veltkamp, 2005), they fail when encountering different recordings of the same song, even when they are very similar (e.g., alternate takes in the studio, or live performances by the same artist). The measure of topical relevance employed by the software ("musical objects are relevant for retrieval if their *audio fingerprint* matches that of the input") does not necessarily match that employed by the user ("musical objects are relevant for retrieval if their *musical identity* matches that of the input").

1.4.5 Summary: Challenges of music information retrieval

The fields of textual IR and MIR share much common ground; the history of MIR can be understood as an evolution of textual IR traditions, acknowledging strong influences from a number of external fields. Nevertheless, many challenges apply uniquely or differently to MIR. There are pragmatic difficulties, such as the challenges of designing and maintaining the tools and infrastructure for international research collaboration in the face of a litigious music industry antagonistic to the sharing of digital music. There are representational complications, such as the much greater role of non-symbolic formats (i.e., audio recordings), both in information

storage and in query formulation. There are challenges of semantics, as the comparatively straight-forward mappings between language and meaning are much more tenuous in the abstract domain of music. There are differences in information seeking behaviour, as opportunistic, non-goal oriented activities are behind the most common encounters with new music. Finally, experiential measures of topical relevance are more difficult to define, as musical identity exhibits a curious resilience to various sorts of distortions; and situational relevance takes on a pronounced importance with the influence of affective, cognitive, and situational factors.

A greater emphasis on the perceptual and cognitive processes of listeners, and on the music information needs and behaviours of potential users of MIR systems, offers one potential avenue towards addressing many of these challenges. However, as in the early days of textual IR, the dominant paradigm of MIR remains firmly entrenched in systems-focussed research.

1.5 Perspectives from music perception and cognition

Of the various perceptual facets of musical experience, those related to pitch and time appear to be most fundamental to the identity of musical objects. In its simplest definition, melody is "the result of the interaction of rhythm and pitch" (Whittall, 2011). Accordingly, "the cues that arise from sequential variations along the pitch and the temporal dimension are apparently the most determinant" sensory cues in the identification of familiar melodies (Hébert & Peretz, 1997, p. 518). Cues provided by other facets, such as *timbre*—the tonal quality of the sound that differentiates a note played on a piano from the same note played on a harpsichord—are insufficient for the task of melody identification; although above-chance identification of musical

genre has been demonstrated with stimuli as short as 250ms, less than the duration of half a beat at a moderate tempo of 110 beats per minute, and far too little time for pitch and rhythm variations to unfold significantly (Gjerdingen & Perrott, 2008).

1.5.1 Topical relevance and musical objecthood

A number of methodologies have been applied in music perception research to determine the contributions of individual musical facets to the identity and well-formedness of musical objects in the mind of the listener. Such concerns are of interest in an MIR context; given the definition of topical relevance as a measure of the mapping between a query and a set of information objects, a notion of perceptual objecthood is a requirement if topical relevance is to be operationalized in the music domain (see section 1.4.3). Approaches can be broadly categorised into two camps: investigations of the conditions under which perceptual objects arise from individual stimulus dimensions, and investigations of the degree to which these individual dimensions may be distorted without affecting the holistic properties of the perceptual object.

1.5.2 Gestalt theory, stream segregation, and auditory scene analysis

In the perception literature, Gestalt theory, a movement in psychology seeking to explain perceptions in terms of high-level holistic forms ("Gestalts") rather than by analysing finer-grained constituents, offers an approach towards explaining the nature of perceptual objecthood. While research in this direction has primarily been directed at visual perception (Bregman, 1990), other modalities, including auditory sensation, have been studied within this framework from the very beginning; addressing an aspect of the musical facet of harmonicity, Wertheimer (1923) exclaims: "I hear a

melody (17 notes!) with its accompaniment (32 notes!). I hear a melody and an accompaniment, not simply '49' and certainly not a straight line, nor '20' and '29' at my leisure" (p. 301). Addressing the Gestalt principle of figure-ground segregation, he discusses the "emergence of a Motif from the cacophony, the elevation from the background of an accompaniment, the phenomenal 'breaking of the silence' " (p. 350).

Figure-ground segregation and other Gestalt principles, such as grouping by proximity, masking, belongingness, and perceptual closure, offer a theoretical framework upon which the notion of perceptual objecthood may be conceptualised. Bregman (1990) outlines analogous applications of these principles in the visual and auditory domains. Bregman refers to the cognitive process by which raw sensory evidence is combined into perceptual structures as 'scene analysis'. 'Auditory streams' are defined as "perceptual grouping[s] of the parts of the neural spectrogram that go together" (p. 6); as "perceptual units" that represent distinct events happening in the physical world; and as a "computational stage on the way to the full description of an auditory event" (p. 8).

Bregman details a study that elaborates on Wertheimer's insight about the perceptual nature of melody and accompaniment. In the study (Bregman & Campbell, 1971), a loop of six tones with distinct pitches is presented to participants; three of the tones are 'low' and three are 'high' pitched, with a gap of at least one and a half octaves between the two groups. The tones are arranged in alternating order between the low and high group; that is, if the six tones are numbered from 1 to 6 in ascending pitch order, the tones are presented in the sequence 142536. When

the sequence is presented very slowly, participants hear the tones in the correct presentation order. However, as presentation speed is increased, participants begin to perceive two distinct streams of tones, "one containing a repeating cycle of the three low pitched tones, 1-2-3- (where dashes indicate silences) and the other containing the three high tones (-4-5-6)" (p. 12). Bregman refers to this phenomenon as 'stream segregation'.

The perceptual framework outlined by Bregman is complementary, but not identical, to Gestalt theory. In particular,

> "[Gestalt theory] sees the principles of grouping as phenomena in themselves, a self-sufficient system whose business it is to organize things. The scene-analysis approach relates the process more to the environment, or, more particularly, to the problem that the environment poses to the perceiver as he or she (or it) tries to build descriptions of environmental situations" (Bregman, 1990, p. 21)

1.5.3 Theory of Indispensable Attributes

Another complementary approach is offered by Kubovy and Van Valkenburg's theory of auditory objecthood (2001). In their approach, the term 'perceptual object' is defined as "that which is susceptible to figure-ground segregation" (p. 102). By their account, grouping, as discussed by Bregman (1990) and Wertheimer (1923) for visual and auditory modalities,

> "produces Gestalts, or perceptual organizations, which are also putative perceptual objects. Attention selects one putative object (or a small set of them) ... and relegates all other information to the ground ... the putative

> objects that become figure are perceptual objects, whereas the ground remains undifferentiated information" (Kubovy & Van Valkenburg, 2001) (p. 102).

Perceptual objects are emergent properties formed by aggregations of individual elements. There are two types of emergent properties: 'eliminative' and 'preservative', depending on whether the properties of the individual elements remain accessible after aggregation. A melody, from this perspective, is identified as an emergent property of the set of notes defining it; the individual notes remain accessible, and hence a melody is a preservative property.

Perceptual grouping occurs over a set of elements distributed over one or more stimulus dimensions. In the case of the melody, the elements (notes) are distributed over two dimensions, frequency and time (Bregman, 1990; Kubovy & Van Valkenburg, 2001). Since the melody is a preservative property, the elements retain their numerosity—the notes remain countable even as the emergent melody is perceived. Kubovy and Van Valkenburg (2003) refer to this situation in defining their criterion for indispensable attributes: "*If* you distribute elements over a medium and perceptual numerosity is perceived, *then* the attribute is indispensable" (p. 227).

Their Theory of Indispensable Attributes (TIA) provides a conceptual framework for the definition of edges; edges are important as they define the boundaries of perceptual objects. In visual perception, edges are defined over the indispensable attributes of the visual domain, space and time. In auditory perception, the analogous indispensable attributes are frequency and time; Kubovy and Van Valkenburg illustrate this with a set of thought experiments. Their proposed plenacoustic

function—the process that characterizes the edges of perceived auditory objects—identifies fundamental frequencies as potential lower edges of "auditory objects as harmonic complexes (such as a voice)", but acknowledge that "more complex combinations of sound" may have upper edges in frequency/time as well. (p. 229)

TIA, as outlined by Kubovy and Van Valkenburg, thus proposes workable definitions regarding the boundaries of auditory objects along the domains of frequency and time. Their plenacoustic function assumes a local snapshot of time during which contributions of different frequencies may be measured for edge detection. The situation suggested by their plenacoustic function is of theoretical value for auditory perception, and may explain certain aspects of musical experience. However, there are limitations in applying a notion of perceptual objects based on time snapshots to more longitudinal aspects of music, such as melody identification.

1.5.4 Melodic objects: Analytical and holistic processing

Melodies retain their identity under transposition; thus, an isolated tone could conceivably match a given location in any suitably transposed melody. The plenacoustic function proposed by Kubovy and Van Valkenburg performs the process of edge detection in auditory perception, based on the individual elements—frequencies at a given time instance—that are aggregated by perceptual grouping into auditory objects. By analogy, a plenmelodic function to perform edge detection in music perception would be concerned with identifying perceptual groupings among the individual elements—notes—that aggregate into melodic objects.

As noted above, melodies are preservative properties, and thus, perceptual access to the individual notes of a melody is retained. This suggests two distinct types

of cognitive processes involved in melody perception: those that operate on the melody as a unit, and those that operate on the set of notes underlying the melody. Schulkind, Posner, and Rubin (2003) make this distinction, referring to the former as holistic processing, and to the latter as analytic processing. They discuss a number of studies from the music cognition literature to support such a distinction. A more precise functional definition of the two processes is given as follows:

> "[C]haracteristics of individual notes or intervals [are] considered to be analytical properties"; "characteristics of longer sequences of notes [are] considered to be holistic properties." (Schulkind et al., 2003, p. 221)

Schulkind, Posner, and Rubin employ a melody identification task in which familiar melodies are presented iteratively to participants on a note-by-note basis, one note at a time. If the melody is not identified based on the notes played in a given trial, the next iteration presents the same notes, plus the next note in the melodic sequence. This approach, inspired by the cohort theory of spoken word identification (Marslen-Wilson, 1987), tests the hypothesis that melody identification proceeds by note sequences activating a cohort of melodies in memory that share these initial elements; further notes provide additional information to progressively prune the size of the cohort, until the target is sufficiently distinguished from all other melodies, and identification occurs.

Schulkind, Posner, and Rubin apply musicological analysis to identify different characteristics present in the melodies that are associated with analytical processing (e.g., interval size and direction, interval consonance, pitch height, and duration of individual notes), and with holistic processing (e.g., tonal function, contour patterns,

metrical accents, rhythmic factors, and phrase boundaries). Their results implicate note placement within phrase boundaries as the most consistent predictor. More generally, four of five characteristics identified as significant code information "about the overall temporal and pitch shapes of the melody rather than characteristics of individual notes or intervals," and would thus be associated with holistic processing. This implies that melody identification is a holistic task, largely taking place at a level of abstraction above that of individual notes. The relative predictiveness of note placement within a musical phrase suggest that "musical phrases may be processed as unified, coherent entities or gestalts" (Schulkind et al., 2003, p. 241). The results suggest melodic phrase boundaries as potential candidates to define the edges of perceptual musical objects in a plenmelodic function.

1.5.5 Melodic facet distortion

The studies on auditory objecthood discussed thus far have explored the merging of perceptual qualities of acoustic or melodic stimuli into perceptual entities—acoustic or musical objects. A number of further studies have investigated the malleability of perceptual objects; that is, the degree to which the elements distributed along the stimulus dimensions of the perceived object may be altered without fundamentally altering the object's identity. White's pioneering study on melody identification (1960) explores such questions in a musical context. In this study, participants attempt to recognize distorted versions of familiar melodies; distortions are performed on pitch intervals (e.g., doubling the size of all intervals) and on rhythm (e.g., setting all note durations to the same value).

Hébert and Peretz (1997), conducting research inspired by White's study, address certain methodological flaws; White's subjects are pre-informed about the limited list of 10 melodies from which the stimuli were drawn, whereas Hébert and Peretz's participants do not have access to such information, requiring them to draw upon their long-term memory for identification. Kuusi (2009) further builds on Hébert and Peretz's work, adding harmony as a parameter alongside pitch and rhythm.

The results of these studies have indicated a strongly diminished role in the importance of rhythm, compared to pitch. The rhythmic distortions are typically achieved by imposing isochrony (setting all note durations in the melody to the same value) on the assumption that this nullifies rhythmic information. However, the stimuli used in these studies – highly familiar melodies – often feature very simple rhythms to begin with; for instance, "Twinkle Twinkle Little Star", a commonly used stimulus, is entirely composed of quarter notes. Melodies undergoing isochronic transformations may thus retain large parts of their original rhythmic structure (Schulkind et al., 2003).

Furthermore, relatively little focus is given to the separation of holistic and analytical processing; Prince's study (2011) exploring the contributions of both types of processing in the integration of stimulus dimensions in music perception is an exception.

In Prince's study, participants are presented with melodic stimuli exhibiting four levels of structural conformity in both pitch and time: an unaltered version; a random reordering of the original elements, preserving conformity to existing tonal

or metrical structure; a random reordering perturbing the original structure; and a completely randomized sequence, including pitch classes and durations not present in the original sequence. Participants are tasked with providing either a rating of the 'well-formedness' of a melody ("if it sounded like a normal, typical melody, conversely if it sounded like there was something wrong with the melody", p. 2134); or a classification of structural conformity ("is this melody metric or random?","is this melody tonal or atonal?", p. 2143). In either condition, participants are presented with three experimental blocks during which they are told to focus exclusively on the pitch or temporal characteristics of the melodies, or on both. One interesting finding among all conditions is that the differences in response to levels 2 and 3 (random reorderings either preserving or perturbing conformity to the pitch and temporal structure of the original) generally do not reach strongly significant levels. This may indicate "all-or-none" effects of tonality and syncopation on pitch and rhythmic conformity in musical contexts (p. 2137).

The latter condition, involving the classification task, presents a situation in which "information along one dimension is purely irrelevant to the other ... provid[ing] a context that promotes the ability to process pitch and time more independently and therefore an opportunity to test how involuntary this interference is" (p. 2142). This offers empirical access to the question of holistic and analytical processing in the melodic context; the absence of interference of e.g. temporal factors in a pitch-judgement task would indicate strictly analytical contributions of the two attributes, whereas the presence of such interactions implies holistic processing of melodies.

In fact, Prince finds that "pitch and time always affected responses in every condition of all experiments", but that there was "variation in how these dimensions combined—sometimes showing independent and additive properties, and other times demonstrating more complex interactions". (p. 2147) In particular,

> "interactions were more likely as the salience of pitch and time approached parity. Yet pitch remained the more salient dimension except when listeners attempted to consciously ignore it, and even in these conditions the advantage of time over pitch never approached the corresponding advantage of pitch over time when rating or classifying pitch". (Prince, 2011, p. 2147)

The various behavioural studies discussed here all make use of stimuli representing distorted versions of 'real' music. Similar distortions are encountered readily in everyday life, for instance when a novice pianist hesitantly stumbles through a piano performance, or when one band produces a radically altered cover version of another band's song. Furthermore, the ability to recognise melodies is fairly universal, and thus results are likely to generalise well. Although the findings produced through such studies may lack the deep, explanatory power of fundamental auditory perception studies as covered in section 1.5.2, they apply much more readily to the everyday music tasks addressed by MIR systems. In Chapter 3, we employ a related paradigm to re-evaluate the role of rhythmic information in melody identification, in order to provide approaches towards the creation of practicable experiential measures of topical relevance in MIR.

In the preceding discussion, MIR has been characterised as a young field of research with a predominantly systems-focussed outlook. The user-centric problems of the multiexperiential challenge, and the perceptual complexities inherent in the multifaceted challenge (Downie, 2003) have been discussed to characterise the establishment of robust criteria for musical relevance as a research priority for the field. However, we have shown that the definition of relevance criteria in the abstract domain of musical information is not a trivial task. To address these challenges, we will need to consider insights from studies of the information needs and behaviours of MIR users, and of listeners' music perception and cognition more generally.

Such studies are generated by a wide range of research fields. In order to maximise the benefits of this interdisciplinary collaboration, while addressing as far as possible the communication difficulties inherent in what Downie terms the "multidisciplinary challenge" (Downie, 2003, p. 306), it is worth considering some details in terminology, pertaining in particular to the term *cognitive* and its use within the fields of information science, and music perception and cognition. Information scientists differentiate between *user-centric* and *cognitive* approaches to information retrieval research—the former concerned, broadly, with understanding user motivations, intentions, information seeking strategies and behaviours (e.g. Cunningham, 2002), whereas the latter is concerned in particular with the perception and interpretation of document contents, casting the concept of information as the result of this interpretation process (Ingwersen & Järvelin, 2005, Chapter 5). In contrast, research in music perception and cognition is directly concerned with the processes taking place in the mind of the listener, a "chain of transformations affecting a stimulus (e.g. a

piece of music reaching the ears)" (Aucouturier & Bigand, 2012, p. 397), eventually producing a behaviour, e.g., an emotional reaction, recognition, learning, etc. In this context, cognition is to be understood as a set of explicable processes, triggered by external and internal signals, and resulting in behaviours that can be measured, e.g. via psychometric laboratory experiments.

In order to allow the outcomes of studies from such diverse backgrounds to inform the task of operationalizing relevance criteria, we require a conceptual framework that is flexible enough to incorporate these diverse approaches, enabling findings to be compared and triangulated. We now proceed to Chapter 2, in which we employ Saracevic's stratified model of relevance interactions (Saracevic, 1997), adapted to the music information domain (Weigl & Guastavino, 2013), to establish such a framework, and to synthesise results from its application to a large subsection of the MIR user literature.

CHAPTER 2
Relevance in Music Information Retrieval

The field of music information retrieval (MIR) is highly interdisciplinary in nature, drawing on research from a diverse range of disciplines in joint pursuit of providing robust, comprehensive access to musical information (Downie, 2004). The field is rooted in traditional textual information retrieval (IR) research, but is heavily informed by areas including digital signal processing, audio engineering, computer science, musicology, and music perception and cognition.

Research dedicated to music as information to be stored and retrieved is a relatively recent phenomenon. While a few pioneering studies relating to MIR can be identified in the "distant past" (e.g., Kassler, 1966), research interest remained sparse until the use of computerized databases became more prevalent in humanities scholarship; an early example is Huron's paper on detecting and handling errors in music databases (1988). By the late 1990's, the arrival of new technologies such as the MP3 file format, file-sharing platforms such as Napster, plummeting costs of digital storage, and the widespread adoption of the Internet created unprecedented requirements for efficient music storage and retrieval. This new urgency to handle vast quantities of digital music correspondingly mobilized research attention. The annual conference of the International Society for Music Information Retrieval (ISMIR), first held in 1999 as the International Symposium for Music Information Retrieval, provides a platform for collaborative research on this topic.

The origin and growth of ISMIR has been motivated by the textual IR world. Plans for an evaluation platform based on that of the Text REtrieval Conference (TREC) were under discussion from the beginning (Downie et al., 2009), and eventually led to the creation of MIREX, the Music Information Retrieval Evaluation eXchange (Downie et al., 2005).

Given this emulation of developments in the field of textual IR, it is perhaps unsurprising that the primary emphasis of research in MIR has been placed on the development of MIR systems. An enormous amount of concentrated research activity has gone into the creation and continued improvement of algorithms to perform tasks integral to MIR, such as onset and key detection, tempo extraction, beat tracking, genre classification, and many others (Downie, 2008). Evaluation metrics are generally applied to parameters of system performance (e.g., precision and recall of automatic classification outcomes), against baseline datasets generated by pre-existing reference algorithms, or against human "ground truth" typically generated either by musicologists' expert annotation, or by crowd-sourcing from non-expert listeners (e.g., Aljanaki, Wiering, & Veltkamp, 2014; Bertin-Mahieux, Hoffman, & Ellis, 2011; Burgoyne et al., 2011).

Although there have been repeated calls in the literature for a greater emphasis on the (potential) users of music information systems—complementing this valuable and thriving research on MIR algorithms—formal consideration of user information needs and information behaviour has been relatively sparse in comparison (Lee & Cunningham, 2012; 2013; Schedl, Flexer, & Urbano, 2013; Weigl & Guastavino, 2011). The situation reflects the early state of research in the field of textual

IR, where similar early emphasis on information systems gradually gave way to a more user-centric paradigm (e.g., Dervin & Nilan, 1986; Wilson, 1981). Ingwersen and Järvelin (2005) outline this distinction between systems-oriented IR and user-oriented and cognitive IR research, noting that the latter viewpoint involves "humanistic aspects with respect to contents of messages, technological insights of tools for processing, and social scientific dimensions due to the information activities taking place in a social contextual space" (p. 25).

The distinction is formalised by Jansen and Rieh (2010) in their framework of human information behaviour and information systems. Jansen and Rieh perceive (human) information searching and (systems-oriented) information retrieval as two fields that share common ground, yet pursue "distinct research agendas, with limited exchange of research" (p. 1517). They identify 17 theoretical constructs of information searching and information retrieval, categorised by "intellectual perspective"—the utilisation of a given construct in either or both of the fields—and by "theoretical orientation"—the "focus of the construct in terms of the three core elements in both fields: people, technology, and information"(p. 1521). Central to the information-theoretical orientation of both fields in their framework is the concept of *relevance*. From the systems-centric perspective, information retrieval systems make use of relevance-predicting algorithms to match available information objects to a query in order to generate a set of results. Operationalised as a quantitative metric of the match between a submitted query and retrieved information objects, the concept is more precisely referred to as *topical relevance*. From the user perspective, relevance is a relation between information and contexts (e.g., the user's information needs at

a given point in time), based on "some property reflecting a relevance manifestation (e.g., topicality, utility, cognitive match)" (p. 1525). Operationalised as the outcome of human judgement of the relation between information and information needs, the concept may be termed *situational relevance*.

In his series of comprehensive reviews on the subject, Saracevic (1975; 2007a; 2007b) similarly positions relevance as a central concept: "a, if not ***the***, key notion in information science in general and information retrieval in particular" (2007b, p. 1915; emphasis his). He expands the dichotomy of systems- and user-orientation into a model with more refined analytical granularity (1997; 2007b) by casting the act or process of information retrieval as a set of interactions between users and systems, through an interface at the surface level. Both user and system are represented by a "set of interdependent, interacting layers" (Saracevic, 2007b, p. p.1927) that characterize this dialogue: the system by content, processing, and engineering layers; and the user by cognitive, affective, and situational layers. There is an implicit assumption that this process is "connected with cognition and then situational application" of the retrieved information (Saracevic, 1997, p. 315). A contextual component characterizes the influence of social and cultural factors that may influence or trigger adaptations in various layers. This stratified model of relevance interactions is capable of representing the influence of factors beyond topicality and situation, including those relating to the user's personal views, tastes, history of information consumption, and emotional state. Such factors carry particular significance in non-problem-solving or hedonic information seeking behaviour (Xu, 2007).

2.1 Approaches to musical relevance

The definition and operationalisation of relevance measures corresponding to the user's listening experience has been identified as a key research objective for music information retrieval (Downie, 2003). Although research in this direction remains sparse, several authors have explicitly addressed the topic in recent years.

Laplante (2010b) investigated the relevance criteria of potential users of MIR systems in everyday life situations in a series of in-depth interviews. 15 young adult francophone participants were selected from the Montréal metropolitan community. The interviews were conducted in order to identify clues used by the participants to make relevance inferences about music items, to investigate the influence of individual characteristics (e.g., knowledge, experience) on participants' relevance judgements, and to determine the influence of contextual factors. Her findings suggest that while there is significant overlap in terms of relevance criteria between the textual and music information domains—"e.g., quality, authority, familiarity, situation, user's knowledge and experience" (p. 605)—there are also criteria of unique importance to MIR; for instance, musical genre "displaced topicality as the most commonly used criterion to start a search" (p. 606). Criteria pertaining to the listener, such as individual tastes and beliefs, "have a greater impact on selection than in other contexts." (p. 605). Further, affective considerations and novelty prominently influence participants' relevance judgements, mirroring findings on hedonic information seeking behaviour in the textual information domain (Xu, 2007).

Inskip, MacFarlane, and Rafferty (2010) conducted a user study investigating the musical relevance criteria of creative professionals. Seven expert ad-agency and

independent music supervisors evaluated the relevance of result sets obtained during real-world music searches made by creative professionals in the course of unknown-item searches for music to accompany moving images (TV and cinema commercials, and TV programmes). The evaluators were found to make use of a range of content-based and contextual criteria. The relevance judgements of these creative music professionals are "situated in a socio-cognitive paradigm" (p.526), taking into account not only the appropriateness of the result sets to their corresponding search queries but also the evaluators' estimation of each song's utility to the end user. As in Laplante (2010b) a key finding is the significant overlap of relevance criteria with those discussed in relation to textual retrieval: "Overall relevance judgement categories in music ... appear to relate strongly to earlier findings in those relating to text, despite the many differences between music and text in their actual content" (p. 527).

Knees and Widmer (2008) presented a user evaluation of an MIR system providing natural language music search by combining textual IR techniques (a tf-idf variant) operating on web pages resulting from Google queries, with acoustic similarity calculations based on the music's audio content. Their system further provides an explicit relevance feedback mechanism that adjusts the weighted term vectors representing the query and result sets in order to adapt subsequent search results to the user's preferences. A small user study with 11 participants was conducted to investigate the impact of relevance feedback on user's relevance judgements of result sets. Results when comparing a version of the system employing relevance feedback in the generation of results, compared to a control system that did not

take relevance feedback into account, were not entirely consistent, with participants apparently differing in their relevance judgements between evaluation trials. Nevertheless, a general trend toward improved results was identified with the system incorporating relevance feedback.

2.2 Modelling relevance in the music information domain

Acknowledging the work exploring relevance and music presented above, to our knowledge, no research has as yet attempted to outline a broad conceptual model of relevance for the domain of musical information. Reflecting on the diversity of musical information needs and MIR use cases, it is clear that relevance is a highly complex notion in music—as it is in text—information retrieval. Even seemingly clear-cut requests to musical information systems ("please retrieve musical work X by composer Y"), where a binary relevance model might seem sufficient at first consideration ("the system's response is relevant if it includes the musical work I seek; otherwise it is irrelevant"), quickly become complicated by the multifaceted nature of musical information: musical identity is highly resilient, and can be retained even when a work is transposed to a different key, orchestrated using atypical instrumentation, performed in the stylistic manner of another genre; melodies may remain recognisable even when undergoing severe distortions of pitch and/or rhythm, as revealed in the music perception literature (e.g Hébert & Peretz, 1997; White, 1960; Chapter 3 of this dissertation), or more viscerally, by an evening's entertainment at a Karaoke bar. Live performances, cover versions, remixes, sampling, and melodic

quotation[1] all further complicate relevance considerations even in this very simple scenario. As in textual retrieval, the user's individual tastes and preferences, current mood, their situational, social, and cultural context, and their particular information needs may complicate things further; far less specific queries are formulated as users seek music for a range of utilitarian and hedonic reasons (Laplante & Downie, 2011).

In adopting a conceptual relevance framework that is broad enough to reasonably span the domain of musical information, given these complexities and key correspondences to and differences from textual IR, it is thus desirable to aim for an established model that is of sufficient analytical granularity to differentiate between concepts of interest regarding relevance, while being abstract and flexible enough to facilitate application from a textual to a musical information context. It is for these reasons that we have decided to adapt the stratified model of relevance interactions (Saracevic, 1997; 2007b) to the music information domain.

2.3 Systematic analysis

We now present a systematic analysis of user studies in the context of music information, focusing on the notion of relevance. Our review is motivated by three overarching aims: i.) to establish the present state of knowledge and to generate hypotheses for future MIR user research; ii.) to draw together existing insights in

[1] Melodic quotations incorporate a melodic passage into a foreign musical context, e.g., the echoing of the introductory notes of Stravinsky's "The Rite of Spring" in Frank Zappa's "Fountain of Love", a satirical tribute to the Doo Wop genre.

order to inform MIR system design; and iii.) to test the suitability of Saracevic's stratified model of relevance interactions within the music information domain.

As per Lee and Cunningham (2012), we are conscious of the challenges in the systematic synthesis of research results in this area, resulting from the highly interdisciplinary nature of the MIR field, and thus the scattering of research articles of potential interest across journals in multiple domains. Thus, we gratefully avail ourselves of the list of articles subject to analysis by Lee and Cunningham, which they have made available on the web[2] . The list was assembled using a selection strategy that focussed on "1) empirical investigation of needs, behaviors, perceptions, and opinions of humans, 2) experiments and usability testing involving humans, 3) analysis of user-generated data, or 4) review of [such] studies" (p. 391). This list represents a set of research output that has already been subject to bibliometric analysis; by making this set the subject of our present review, we hope to avoid "reinventing the wheel" when it comes to defining the articles of interest to MIR researchers interested in music user studies, while contribute to the richness of descriptive detail available about this collection of works.

Various studies included in this collection investigate music information behaviour under very specific task specifications (e.g., in user evaluations of specific MIR systems). In order to capture subtleties relating to the *modes of searching* employed, we further augment our application of the stratified model of relevance interactions to employ Wilson's formulation of the modes of information searching and

[2] Available at http://www.jinhalee.com/miruserstudies

acquisition (1997, p. 562) as an additional category for the classification of findings. These modes include *active search*, in which an individual actively seeks out new information (e.g., searching for a specific song); *passive search*, in which one type of information behaviour incidentally leads to the acquisition of unrelated information that happens to be relevant to the searcher (e.g., finding out about a new release by a favourite artist while seeking out an older track); *passive attention*, in which information acquisition may take place without intentional seeking (e.g., overhearing the top-40 on a radio playing in the background); and *ongoing search*, in which a pre-existing knowledge framework is updated and expanded through occasional, exploratory continuing search (e.g., periodically visiting a favourite record shop to keep abreast of new releases). We extended Wilson's categories with two additional modes that were expected to occur frequently in the MIR user studies under review: *playlist generation*, a special instance of *active search* that involves the assembly of an ordered collection of music, typically reflecting a planned listening session; and *browsing*, in which a collection is explored in order to obtain an overview of what is available, without a particular further search goal in mind. Of course, users may be involved in many or none of these modes of searching at any point during everyday listening, and only some of the studies under discussion are explicitly concerned with such topics.

2.3.1 Methodology

In order to address the workload inherent in the analysis of this large collection of articles at a low level of granularity (i.e., with the article content, rather than bibliographic metadata, as our primary focus), we distributed the task of coding

the collection of articles among four researchers: the author of this dissertation, two professors at McGill's School of Information Studies, as well as one post-graduate research assistant (a PhD candidate at the school). To promote coding consistency between the four researchers, we initially conducted an iterative process of picking two to four articles from the list, coding them in isolation according to a preliminary application of Saracevic's model of stratified relevance interactions to MIR (Weigl & Guastavino, 2013), and then meeting for group discussion of the coding process and comparison of individual coding results. A mutually agreed coding was determined for each article, and the outcomes of the discussion informed the next iterative coding cycle. This processes was repeated until a concensus in terms of the coding approach applied by each researcher was reached after 20 mutually coded articles. The remaining 139 articles were then distributed among the four researchers, one researcher for each article.

At this point, the agreed-upon coding procedure was formalized. Our basic unit of analysis was the individual relevance-related finding, as generated by empirical research reported in the respective article being coded; i.e., findings based on citations from the previous literature, or grounded in speculation not supported by the reported data, were not coded in conjunction with a particular study. Where an article reported on two or more studies applying differing methodologies, investigating different research questions, or sampling participants from a different population, each study was coded separately; where two studies in the same article replicated the same finding, the finding was accordingly coded multiple times, once for each study.

In order to assist the researchers in their coding activity, a simple web application was developed to track article assignments and coding progress. For each article, the researchers were able to encode one or more studies presented by providing short descriptions of the study, the employed methodology, the sample frame and size.

The researchers assigned findings to each study by providing a short textual description, and by choosing descriptors from a set of drop-down menus corresponding to each stratum of our adaptation of Saracevic's model (Figure 2–1). Any number of the available descriptors could be selected for each stratum (including zero). The set of descriptors was originally determined based on the preliminary application of the model to the music domain, and on our subsequent iterative discussions during the initial phase of the coding process. Each descriptor is associated with a specific stratum of the model, and comprises the label that appears in the drop-down menu, and a short textual definition that is visible in the coding interface when new findings are configured for assignment to a study. For the purposes of the coding of findings, descriptors from the *modes of searching* category were handled using the same mechanism employed for the various strata of the model of relevance interactions; the *modes* were treated separately during analysis.

In cases where none of the available descriptors suitably conveyed the meaning intended by the coding researcher, new descriptors could be minted by specifying a stratum of the model and providing a corresponding new label and definition. Provenance information describing the particular coding researcher responsible for the addition, as well as the specific article, study, and finding that prompted the addition, was recorded automatically by the system. New descriptors added to the

system in this way became immediately available to all researchers in their subsequent coding activities in order to promote consistency and to prevent overlapping concepts with different descriptors (according to the respective researchers) from introducing unnecessary ambiguities in subsequent analysis.

The researchers agreed to flag instances where none of the available strata suitably accommodated aspects of a particular finding; however, after the conclusion of the iterative discussion phase, no such instances were identified.

2.3.2 Analysis and results

Quantitative analysis of the distribution of findings across the stratified model of relevance interactions. Our application of the stratified model of relevance interactions to the music domain outlines nine stratum-level coding classes: three systems-related classes (content, processing, engineering); three user-related classes (cognitive, affective, situational); the interface class; and two contextual classes, relating to social and to cultural context. Note that Saracevic's original model implicitly combines the latter two classes into one combined contextual component, although the ambiguity and multiplicity of context in information science is acknowledged: "Context is a plural" (Saracevic, 2007b, p. 1927). Here, we have elected to explicitly separate social and cultural context into their own strata, as several of the user studies in our collection look individually at the influence of social or of cultural context. In our conception, *social context* refers to influences from or on specific individuals, including friends, relatives, colleagues, or other users involved in interactions with the user of an online system. *Cultural context*, in turn, refers to influences from or on members of a particular culture or subculture, sampled

Finding: Majority of participants attach importance to recommendations, reviews, and ratings. Recommendations from friends or colleagues perceived to have discriminating tastes carry the greatest influence.

Content	Processing	Engineering	Interface	Cognitive	Affective	Situational	Social Context	Cultural Context	Mode of Searching
Extra-musical information				Taste profile			Influence of friends		

Delete

Figure 2–1: Coding interface.

Left: Form to add findings. Researchers chose appropriate descriptors from drop-down menus corresponding to the model's relevance strata, specifying new descriptors (with short explicatory hints, e.g., see text in yellow highlight) as required.

Right: Example of an encoded finding. All completed findings for a particular study are listed on the same page in order to track coding progress.

to represent behaviours and attitudes of that group as a whole; examples include broad-brush demographic factors (e.g., comparing music information behaviours of participants based in New York with those of participants based in Hong Kong), as well as finer distinctions (e.g., relating to typical listeners of "classical" or of "heavy metal" music). The *cultural* context category further includes vectors of enculturation including the influence of television and radio programming.

While coding the articles, we chose descriptors from any combination of the nine relevance strata, according to what we felt most suitably corresponded to the finding under discussion. Where appropriate, we also included descriptors of the *mode(s) of searching* relating to the finding. We now present an analysis of the distribution of the stratum interactions identified during this coding process in terms of the co-occurrence of relevance strata within this nine-dimensional relevance space.

Of the 866 identified findings, none combined more than five interacting relevance strata in practice, with only two findings reaching this maximum. 48% of the findings identified were coded according to the interactions of two strata, with the remainder largely distributed, almost evenly, among single-stratum findings, and interactions of three strata; a small portion of the identified findings (4.8%) encoded the interactions of four strata (Figure 2–2).

The co-occurrence heat map in Figure 2–3 displays the distribution of specific stratum interactions within our corpus of findings. This figure displays a projection of the (maximally) five-dimensional space of relevance interactions into a two-dimensional plane, in order to simplify interpretation; thus, a single finding encoding an interaction of strata A, B, and C is represented as three distinct interactions,

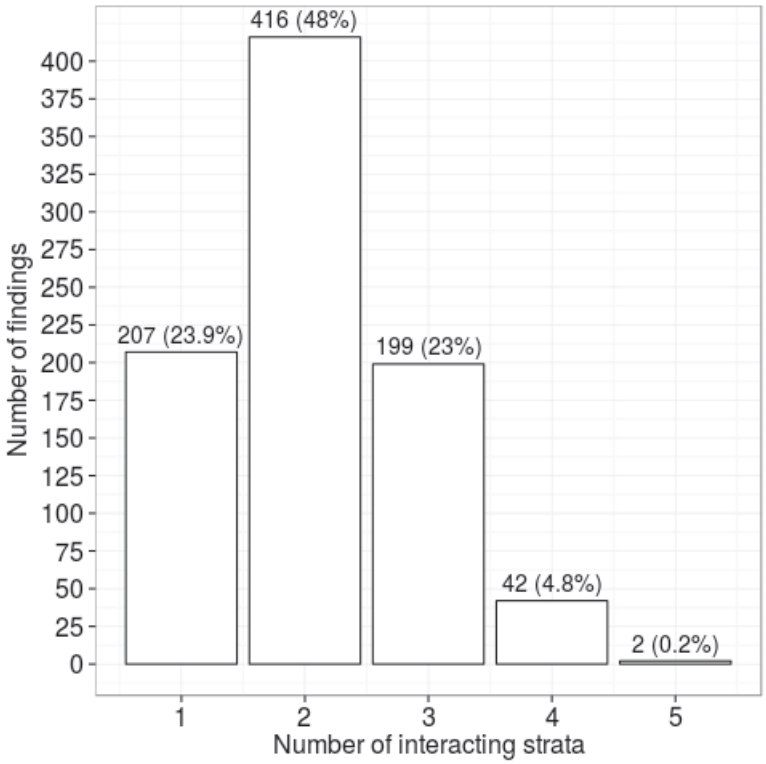


Figure 2–2: Distribution of the number of stratum interactions in the identified findings.

of A & B, A & C, and B & C; the numbers along the diagonal represent the total number of coding occurrences of each individual stratum class, regardless of interaction context — thus the finding encoding an interaction of A, B, and C is included three times in the diagonal, once for each of the three strata. The percentage values thus do not add up to 100; instead, the values express the number and percentage of

findings that describe interactions involving at least those relevance strata described by the axis labels corresponding to the respective cell.

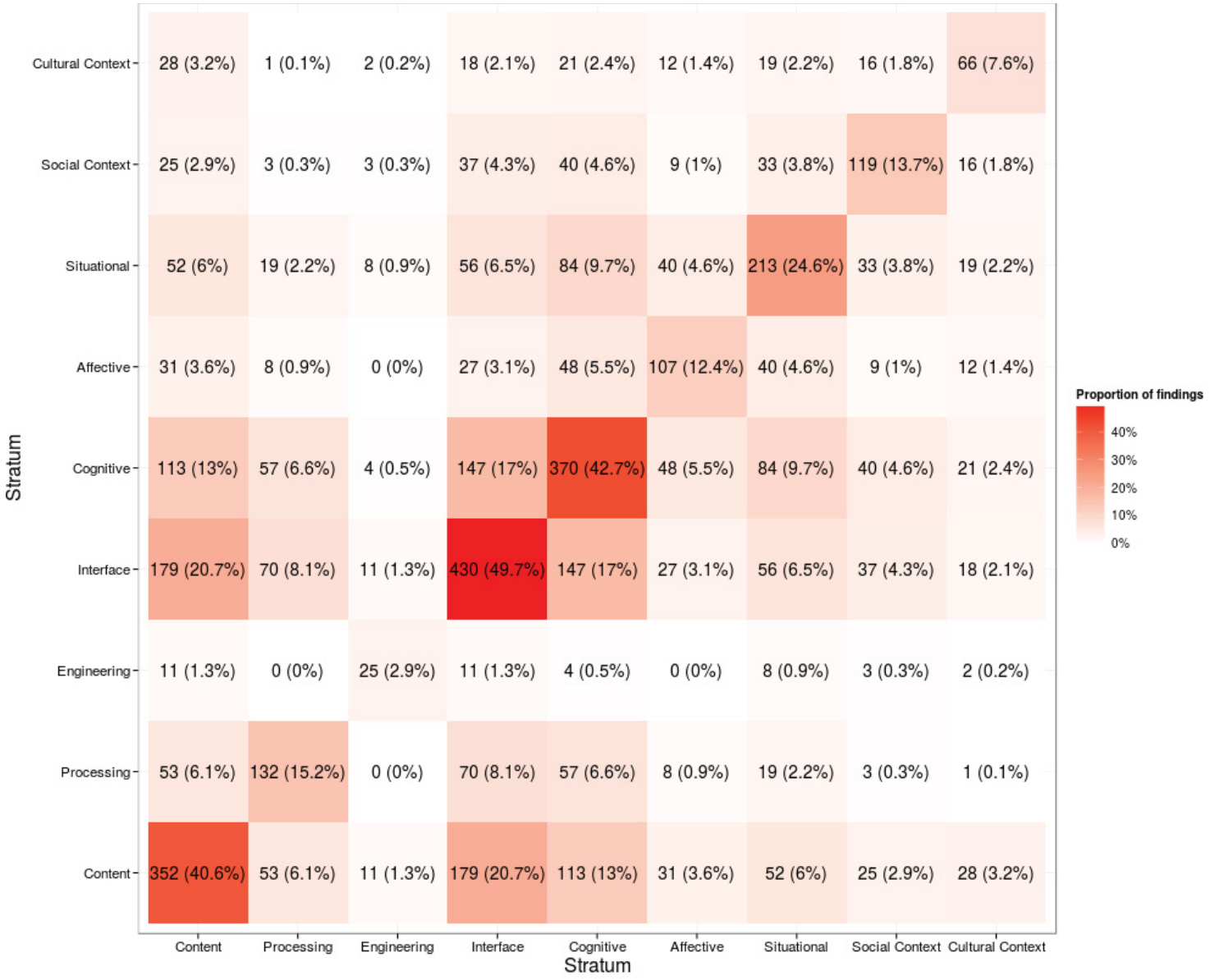


Figure 2–3: Co-occurrence of stratum-level classes within the corpus of coded findings (N=866). Values indicate the number and proportion of findings that incorporate at least the respective strata; thus percentages do not add up to 100%.

Inspection of the number of occurrences of each individual stratum (i.e., the diagonal of Figure 2–3) reveals that almost half of the coded findings involve aspects of the *interface* between user and system (49.7%). Also well represented are the *cognitive* stratum on the user side, encoding aspects of the user's perceptions, behaviours, and preferences (42.7%); and the *content* stratum on the system side, encoding aspects of the information resources available to the system (40.6%). The *engineering* stratum with a focus on the system's hardware characteristics is notable for its virtual absence, appearing in only 2.9% of the findings. The *cultural context* is also relatively under-represented, appearing in 7.6% of the findings (note that the separation of *cultural* and *social context* is specific to our application of the stratified model, differing from Saracevic's conception).

In terms of interactions between individual strata, prevalent pairings include those of the most prevalent individual strata, as would be expected — with *interface* and *content* co-occurring in more than a fifth of the coded findings (20.7%); *interface* and *cognitive* in 17% of the findings; and *content* and *cognitive* in 13% of the findings. Other relatively common co-occurrences include *cognitive* and *situational* (9.7%), *interface* and *processing* (8.1%), *cognitive* and *processing* (6.6%), *interface* and *situational* (6.5%), and *affective* and *cognitive* (5.5%), with all other pairings occurring in less than one twentieth of findings, respectively.

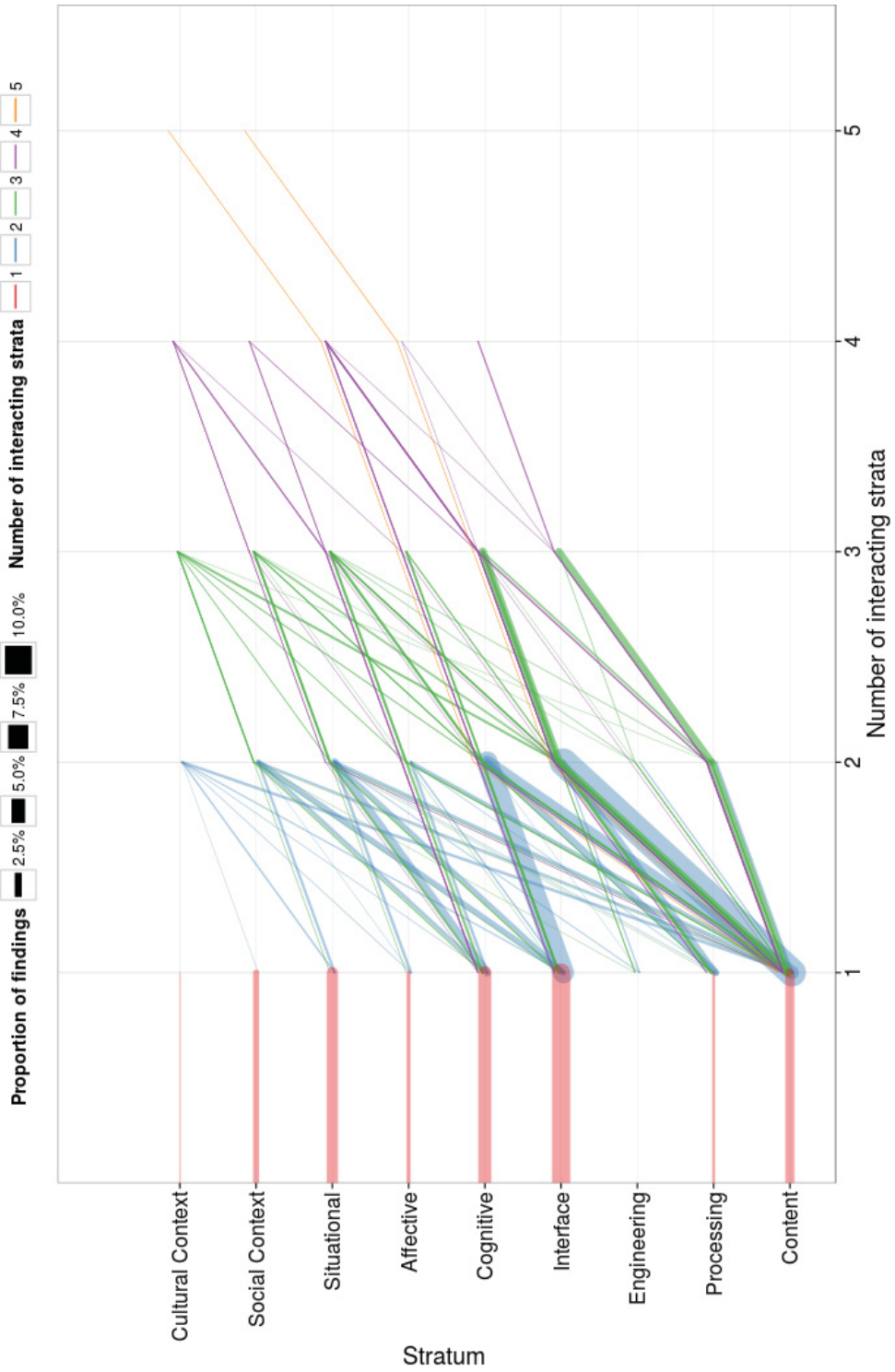


Figure 2–4: Distribution of multi-dimensional stratum co-occurrences within the corpus of coded findings (N=866). In practice, no finding combined more than 5 interacting strata. Line colour indicates the number of strata involved in a given interaction (i.e., the dimensionality of the interaction). Line width corresponds to the number of findings encoding a particular interaction.

In order to obtain a complete overview of the distribution of relevance interactions within our corpus of findings, it is necessary to look beyond two-stratum pairings, by accounting for the full dimensionality of our findings. Figure 2–4 provides a visualisation of these findings within our corpus. The figure is generated by drawing line segments connecting the strata constituting each particular stratum combination; the width of the segments corresponds to the number of findings exhibiting this interaction, and the colour corresponds to the number of strata involved. To aid visibility, strata are always connected in a particular order of precedence, starting at the bottom of the y-axis and working upwards. The numerical data visualized in Figure 2–4 is provided in Appendix A.

Inspection of the red line segments in Figure 2–4 reveals the distribution of the roughly one quarter of our findings that only encode relevance descriptors along a single stratum (see also Figure 2–1). As expected from inspection of the co-occurrence heatmap (Figure 2–3), we find the largest proportion of these singletons in the *interface* stratum (59 findings, corresponding to 6.8% of all findings), with *cognitive* (42 findings; 4.9%) and *content* (29 findings; 3.35%) also relatively well represented. Contrary to the expectations from Figure 2–3, we find a similar proportion of the singleton findings encoded as *situational* (36 findings; 4.2%), suggesting that findings on situational cues to relevance skew toward describing the user's situation in isolation of other relevance criteria.

The largest proportion of findings fall within the two-stratum interactions (blue line segments in Figure 2–4) of *content* and *interface* (90 findings; 10.4%), *interface* and *cognitive* (69 findings; 8%), and *content* and *cognitive* (43 findings; 5%). Of

the three-stratum interactions (green line segments), the combinations of *content*, *interface*, and *cognitive* (27 findings; 3.1%) and *content*, *processing*, and *cognitive* (24 findings; 2.8%) occur most commonly. Findings combining four or more interacting strata are more dispersed, with no combination present in more than 1% of findings.

Distribution of findings by associated *modes of searching*. Table 2–1 displays the identified findings in terms of their associated *modes of searching*. The large majority of findings (70%) were not associated with or descriptive of any particular mode of searching, and hence no mode was specified during the coding process. Of the four modes of information searching and acquisition proposed by Wilson (1997), only *active search* was significantly represented, with 15.9% of findings encoded correspondingly. The remaining modes of Wilson's formulation were less prevalent, with *ongoing search* involved in 2.7% of findings; *passive attention* in 0.9%; and *passive search* associated with only a single finding: serendipitous discovery of new music was frequently mentioned as a highly desirable outcome of MIR system use by participants in a large-scale online survey investigating user requirements for music information services (Lee & Waterman, 2012).

Of the two additional descriptors we included as *modes of searching* after discussions following our initial rounds of collaborative coding, *playlist creation* was relatively prevalent, occurring in 9.2% of findings; with *browsing* represented in merely 3.5% of findings. Further descriptors added to the system during the coding process — *general/comparative* (where differences in modes of searching were the subject of the finding), and *personal catalogue maintenance*, were only sparsely represented.

Table 2–1: Distribution of modes of searching associated with the identified findings (N=866).

Mode of searching	Number of findings
None specified	605
Active search	138
Playlist creation	80
Browsing	30
Ongoing search	23
Passive attention	8
General/comparative	7
Personal catalogue maintenance	4
Passive search	1

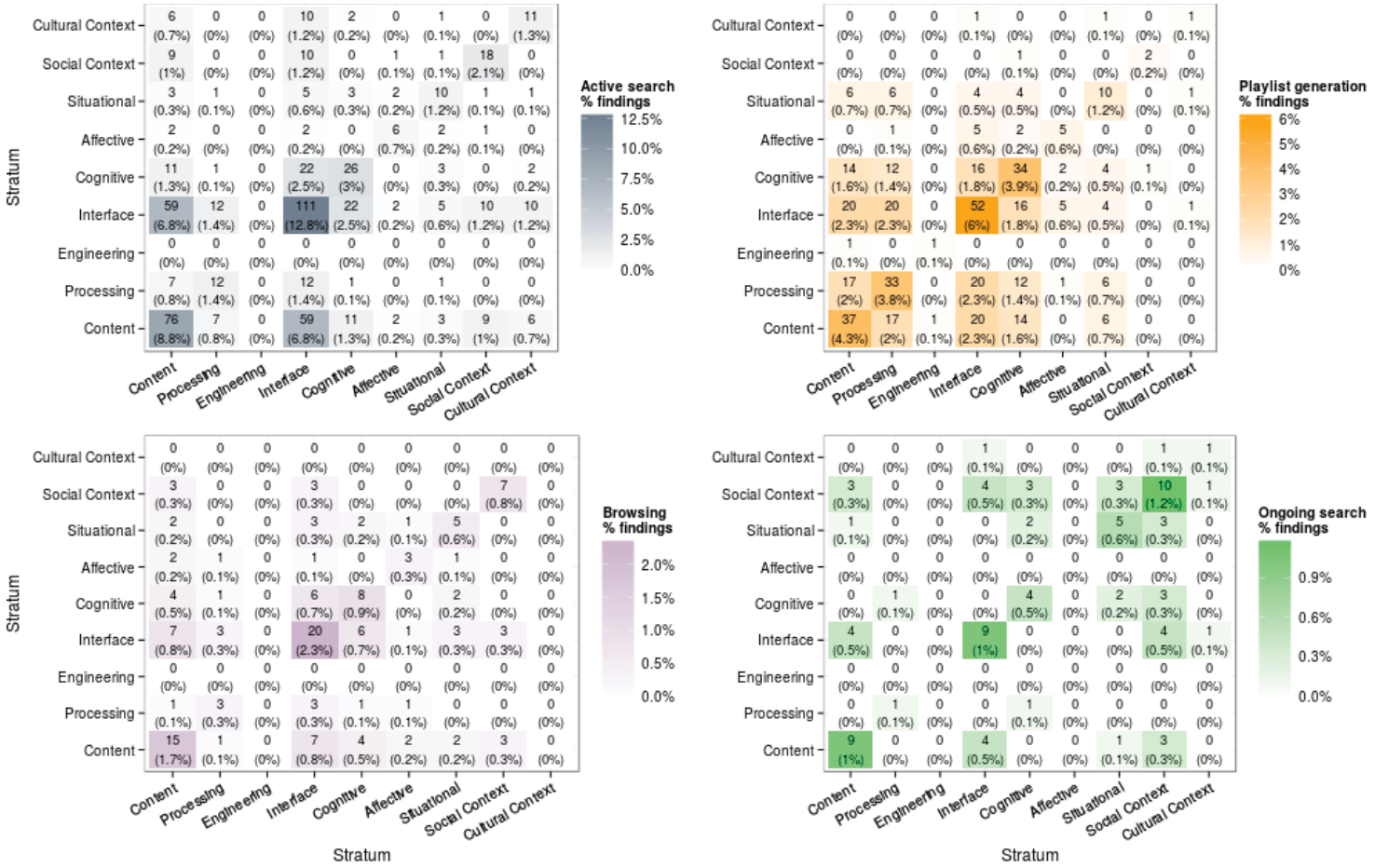


Figure 2–5: Distribution of findings representing the four most common *modes of searching* in terms of stratum-level co-occurrence within the corpus of coded findings (N=866). Note differences in the scale of proportions: the number of findings encoding particular modes differ widely (Table 2–1).

Figure 2–5 illustrates the distribution of stratum-level co-occurrences for the four most frequently encoded *modes of searching*. Although the limited number of data points for all modes but *active search* prohibit conclusive interpretation, some suggestive trends may be inferred. Findings on *active search* predominantly involve interactions of the *content* and *interface* strata, reflecting known-item search situations where a specific piece of content (say, a particular song) is retrieved using the system interface, frequently in the context of a system evaluation task. Findings on *playlist generation* further involve the *cognitive* and *processing* strata, reflecting a number of studies exploring users' decision-making processes when assembling playlists, and evaluating the performance of automatic playlist generation algorithms. Findings on *browsing* most frequently focus on the *interface* or on the *content* stratum in isolation. These findings typically concern relevance cues inherent in or associated with the content (e.g., album covers; genre markers), or particular interface mechanisms that assist browsing activities (e.g., faceted navigation), but do not typically involve specific system evaluation tasks that might produce findings on the interactions of these strata. Finally, findings on *ongoing search* tend to involve social context: many of these findings involve users perusing physical music stores or music libraries, creating opportunities for social influence; further, a primary motivation of ongoing search activities that update one's framework of musical knowledge appears to be to drive social interactions, where music discussion can be used to build personal relationships or project the listener's self-image, one's knowledge of music acting as a "social badge" (as per Laplante & Downie, 2011, p. 208).

Distribution of studies. The quantitative analysis of findings presented thus far provides an overview of the degree of insight on the notion of relevance, understood through the lens of the stratified model, represented within the subsection of the literature under discussion. A related but distinct approach is to focus on the number of individual studies that concern themselves with each relevance interaction, i.e., that include at least one finding encoding a given stratum combination. This approach provides an overview of the degree to which the various possible relevance interactions have been studied in the present literature. Figure 2–6 provides such an overview.

While the two distributions visualised in the co-occurrence heatmaps in Figures 2–3 and 2–6 are broadly similar, with interactions of the *content*, *interface*, and *cognitive* strata being best represented, there are significant differences (Wilcoxon paired signed rank test on the proportion of studies vs. findings according to stratum co-occurrence: V=946, p <.0001); see corresponding data in Appendix B.

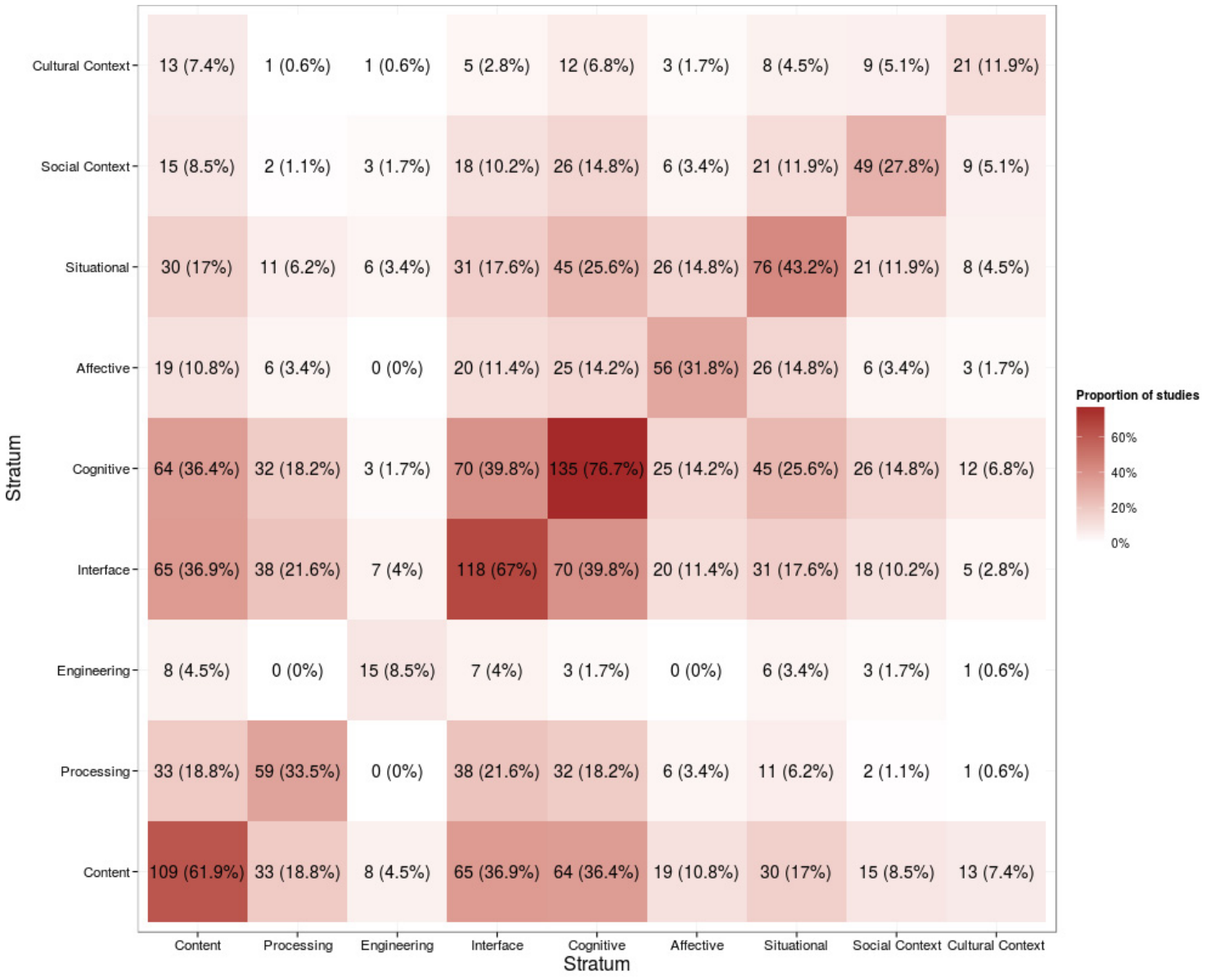


Figure 2–6: Proportion of studies (N=176) mapped onto stratum-level class co-occurrence according to the presence of at least one finding encoding the given interaction within the corpus.

Particularly notable in terms of differences in ranking between studies and findings is the cultural context stratum, implicated in all differences with a magnitude greater than 4 (see Appendix B, Table B2). These differences demonstrate that findings encoding interactions with the cultural context strata are represented in disproportionately greater numbers than studies encoding at least one such finding, suggesting that there are relatively few MIR user studies investigating the role of cultural context, and that those in existence tend to focus predominantly on this topic, in that many corresponding findings are generated.

Discussion. This quantitative description of the findings identified in our survey according to their distribution within the multidimensional space framed by our application of Saracevic's stratified model of relevance interactions outlines the state of insight on the notion of relevance in a music information context within the sizeable collection of user studies under discussion. As might be expected given its central role in the interaction of user and system, we find the *interface* stratum to be most prevalent; in part, this reflects the predominant focus, noted by Lee and Cunningham (2012) in describing the research design of the studies in the collection under discussion, to "[evaluate] what is out there rather than focusing on deeper problems or questions" (p. 394).

Next most prevalent are the *content* stratum on the system side, and the *cognitive* stratum on the user side. This suggests that, although the focus of MIR research activity in general remains concerned primarily with systems aspects (Lee & Cunningham, 2012; Weigl & Guastavino, 2011), MIR user research has expanded beyond the systems-centricity inherent in early textual IR user research (Dervin &

Nilan, 1986; Wilson, 1981): with an eye towards contemporary understanding in textual IR, the focus on the content available within music information systems can be related to the subject literature view of relevance outlined by Hjorland (2010); whereas the focus on the user's cognition relates to the paradigm shift toward outlined by Dervin and Nilan (1986) and the cognitive turn described by Ingwersen and Järvelin (2005), conceiving of the user as an agent actively involved in the subjective, contextual construction of relevance, taking internal cognition into account, and seeking to understand the user experience as a holistic phenomenon incorporating history and consequences, beyond an atomistic focus on the moment of information system interaction.

Notable by its near-absence is the *engineering* stratum, involved in only 25 findings (2.9%). This scarcity of attention suggests there may be few engineering-related concerns facing users of MIR systems, given the raw computing power, high-speed connectivity, and capacious digital storage inherent in the mobile devices now carried in the typical user's pocket; not to mention the computing infrastructure driving the recommendation algorithms of music content-streaming providers such as Spotify or Pandora. Modern music playback systems may simply have become "powerful enough" for engineering concerns not to play much of a role from a relevance perspective. Another explanation for this absence of findings incorporating *engineering* suggests that legitimate, relevance-related engineering concerns do exist for MIR, but that these are under-represented in the subsection of the literature under review. The selection strategy employed by Lee and Cunningham (2012) to assemble

the collection of articles under discussion may have missed articles concerning the *engineering* stratum that only address user needs indirectly but nevertheless determine important relevance considerations. Alternatively, such concerns may also represent a legitimate gap in the literature, with potentially informative research questions remaining to be explored. Do relevance interactions arise when considering the user's choice of listening device in context of the auditory characteristics of the listening environment? Are there certain types of music more suited to, say, noise-isolating over-ear headphones versus noise-permitting on-ear headphones when the listener is exposed to certain types of background noise, and could relevance determination be improved by taking these into account? Are there implications in terms of user requirements of mobile versus stationary listening on a hardware level that impact on music relevance? It is conceivable that the investigation of such topics may yield informative results. In failing to reveal existing findings impacting on engineering, we hope that the stratified model may yet serve to guide the generation of hypotheses for future research.

One further important but relatively under-represented interaction is that of the system's *processing* and the user's *cognitive* strata; this co-occurrence, as a pair or in conjunction with other strata, is represented in 57 findings (6.8%). The outputs of MIR algorithms addressing a particular task — such as beat tracking, key finding, melody extraction, or mood detection — are typically evaluated against human-generated reference classifications, generally provided by a dedicated but relatively small number of expert musicologists, or crowd-sourced from larger numbers of anonymous users. Collections of such classifications are termed "ground-truth",

the veracity of which is often taken at face value. While algorithm performance is meticulously evaluated against such datasets (Downie, 2008), the perceptual and cognitive processes driving the human classification decisions often remain opaque in MIR research. This is clearly an issue where the "correct" choice of classification is demonstrably subjective; even given the undisputed sincerity, dedication, and expertise of the human volunteers, the usefulness of the resulting data as "truth", aiming to reflect the perceptions of all (or at least, a majority of) the potential users of an MIR system, is questionable. For instance, the ground-truth dataset for the MIREX Audio Music Similarity and Retrieval (AMS) tasks run over four years from 2010 — 2013 incorporated similarity judgements provided by trusted volunteers from the MIR community. Organisciak and Downie's subsequent analyses of inter-rater agreement (2015), regarding the broad similarity of pairs of songs on a categorical ranking of "very similar", "somewhat similar", and "not similar", revealed category agreement in only 35% of cases — nearly half of which were on "somewhat similar" items.

Even in MIR tasks not focussing on user judgements, the scarcity of consideration of the listener's music perception and cognition alongside the algorithm's task performance presents potential concerns, especially given the multifaceted complexity of musical information, and thus the diversity of cognitive processes, the performance of which the MIR algorithms are ultimately aiming to emulate. For instance, temporal facets relating to such fundamental aspects of music listening as beat and tempo perception, melody recognition, and sensorimotor synchronization (moving in time with the music) are present in interactions of the *processing* and

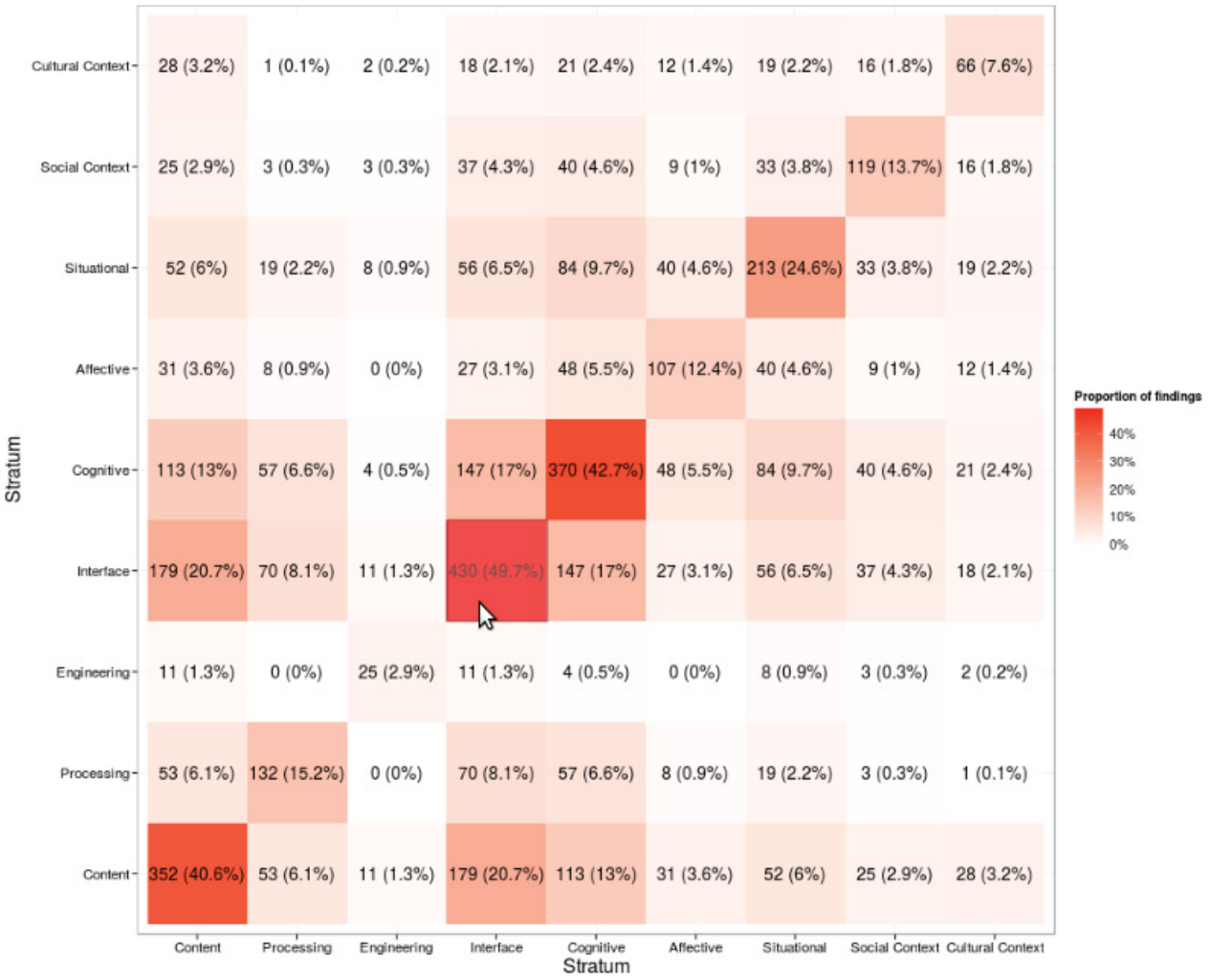
cognition strata in a mere 4 findings in our corpus (0.46%); yet many core MIR tasks, such as beat tracking, onset detection, tempo estimation, downbeat detection, query-by-humming, and melodic similarity, plainly revolve around such facets. Music perception and cognition research has suggested a wide range of inter-individual differences in beat perception and performance in the general population (e.g., Iversen & Patel, 2008); what effects do such differences have on relevance criteria incorporating rhythmic information? Such questions bridging the fields of MIR and music perception and cognition remain largely unaddressed in the present subsection of the literature.

2.3.3 Providing community access

Access to the entire corpus of findings, queryable on the level of stratum interactions and of sub-stratum descriptors, is provided by an online resource[3] . Users of this resource are provided with the stratum-level co-occurrence heatmap displayed in Figure 2–3 and can click on particular cells of interest to load a list of findings on interactions incorporating the corresponding strata, numbered and ordered according to the frequency of sub-stratum descriptor co-occurrence (Figure 2–7). Further, users can select a descriptor co-occurrence of interest to retrieve full details of the corresponding findings, linking to the corresponding research articles and including descriptions of the study that generated each finding. An alternative view presents users with a set of stratum-level drop-down menus presenting all sub-stratum descriptors (as in Figure 2–1, left), enabling the retrieval of all findings incorporating

[3] Available at http://relevance.linkedmusic.org

narrowly defined topics (e.g., all findings relating to the *situational* descriptor of *music practice*). We hope that this tool be useful to both user researchers, and MIR system designers aiming to address particular music information needs or use cases.



Selected strata:
Interface : Interface

8 studies, 9 findings: Interface: General human information interaction +
5 studies, 5 findings: Interface: Specific interface Cognitive: Interface satisfaction +
4 studies, 5 findings: Mode of Searching: Active search Interface: General human information interaction +
4 studies, 4 findings: Mode of Searching: Playlist creation Interface: General human information interaction +
3 studies, 4 findings: Mode of Searching: Active search Interface: Specific interface Cognitive: Interface satisfaction +
3 studies, 3 findings: Mode of Searching: Active search Interface: Specific interface Interface: Query mechanisms +
3 studies, 3 findings: Interface: Specific interface +
2 studies, 4 findings: Content: Bibliographic information Interface: General human information interaction Interface: Query access points +
2 studies, 3 findings: Content: Collections Interface: General human information interaction +

Figure 2–7: Web application providing an interactive analytical interface with views on the corpus of findings.

2.4 Synthesis of findings

While the distribution of findings analysed above provides insight into the multi-faceted notions of relevance in MIR, we must go beyond this distribution and inspect the qualitative content of these findings if they are to guide further user research and MIR system design. We now present a synthesis of the findings represented in the relevance interactions most commonly encountered in our corpus as determined in the quantitative analysis, in order to provide insight into the state of knowledge on the topic across the studies included within the subsection of the literature under discussion. In order to restrict this synthesis to a manageable size, and to exclude "solitary" findings that have only appeared in a single study with no further elaboration or replication, we limit this synthesis to combinations of sub-stratum descriptors that are shared by at least 2 findings in the corpus. The full corpus, including "solitary" findings, is available through the web resource described above.

Note: *For brevity, the synthesis presented in this section refers to articles from the collection by number; a full listing is provided in Appendix C.*

2.4.1 Music use cases

Music information needs often relate to certain uses of music, beyond sheer listening enjoyment, for instance to regulate mood (e.g., relaxation, energizing), to contribute to the background atmosphere at social gatherings and parties, or to fulfil specific functions, e.g., in a ceremonial context [67, 154]. The upkeep of interpersonal relationships, where music is used to promote and maintain social interaction, emerges as an important reason for listening to music; this may take the shape of listening to music in order to dance together, attending live concerts together, or

listening in order to have a topic for shared discussion with others [109]. Users may categorize music according to very specific situational descriptors corresponding to application scenarios (e.g., "programming music" [115], "gym music" [30]). Songs are selected from different styles and genres to serve in different situations [82]. However, users tend to assign multiple use cases to individual songs [115], and musical choices often fulfil a number of different functions simultaneously [43].

One particular use case that has been well described involves the use of music in synchronization, that is, to accompany video (e.g., films, advertisements). This scenario has been studied from the perspectives of the information needs, behaviours, and relevance criteria of creative professionals seeking music [57, 58], and of the record publishers providing catalogues for this purpose [59]. It has also been studied from the pesrpectives of the media consumers who listen to and watch the synchronized media: background music can significantly influence the remembering of filmed events [17]; the congruence and incongruence in terms of the mood expressed by the music compared to the events on screen plays a role, with mood-*incongruent* music resulting in better recall of events in the episode when a scene is foreshadowed, but mood-*congruent* music resulting in better recall when the scene is accompanied by the music [17]. Synchronization with video also has effects on the listener's perception of the music: the product name featured in a TV commercial accompanied with a certain song may be mistaken for the song's title; and conversely, corporate names are typically associated with songs heard during TV commercials very accurately, when these are used as a search cue [86].

2.4.2 Effects of listening

When participants choose to listen to music, this has several effects, linked to style of music. Whether or not the music was chosen by the listener affects the functions the music may fulfil. The reasons for listening to music, and the effect of hearing the music on the listener, vary depending on the location that the music is heard [126].

Mood management has been identified as a key reason for listening to music [67, 154], enabling the listener to calm down, relieve tension, or alleviate boredom. Music listening may be used to provide a means of rest at intervals between work periods; conversely, stimulating rhythms may be used to energize the worker and speed progress [41]. Aside from selecting music to manage one's own mood, proprietors of commercial premises may select background music to significantly affect listener's perceptions of the atmosphere of the commercial environment, extending to some degree to affecting purchasing decisions [118].

2.4.3 Musical engagement

Social interactions are a common source of musical information [109], with users who listen for greater periods disproportionately more likely to encounter new songs and artists by peer influence ("information diffusion") [40]. Conversely, users exposed to more new content tend to see greater diffusion from peers. Highly involved listeners tend to be more positive about their use of music, based on music importance, mood enhancement, coping and identity construction [143]. The extent of users' music collections follow an exponential distribution curve, both in terms of

number of songs and collection size in gigabytes, with collection sizes ranging from dozens of songs to tens of thousands [20].

An analysis of user behaviour across five very large datasets (Netflix ratings, Yahoo! Music ratings, Yahoo! search queries, clicked Yahoo! Results, and Nielsen web browsing data) [42] emphasises the long-tail effect in terms of web resources and music collection sizes; with 10 web sites accounting for 15% of page views, and 10,000 sites for up to 80% of views; and only 5 - 10% of users satisfied with collections the size of large physical retailers, suggesting most have at least some eclectic tastes. Accordingly, the availability of "long tail" inventory brings second-order benefits like customer satisfaction, and small increases in popular inventory lead to high increases in satisfaction. Niche products that are generally unknown also tend to be generally disliked, with the most and least popular songs tending to receive the highest ratings, and a dip in the middle. Listeners' engagement (amount of usage) was not correlated with their eccentricity (median rank of items consumed), although the number of unique items consumed is higher for more engaged listeners.

2.4.4 Specific interface evaluations

Many of the articles under discussion present user evaluations of specific, novel interfaces, including a location-based audio recommendation service [5], haptic audio-tactile playlist and music collection navigation devices [2, 64], visualizations of musical content [46] or of affective parameters associated with the music [157], a tempo sensitive music search engine [156], a bimanual "hands-on" interface to create mappings for a tactile mobile music playback interface [15], and a zoomable user interface widget to enable music collection management on large and small screens [32], among

others. Such studies are typically presented in conference or workshop papers, generally employing lab-based evaluation experiments with relatively small groups of participants. These tend to be students and other members of university communities, sometimes predominantly involved in disciplines related to the researchers' field (e.g., computer science, HCI). As might be expected, most such studies find novelty effects, with users typically rating the presented systems as more satisfying or fun to use than corresponding baseline systems; and learning effects, where participants' task performance improves as they are exposed training or practice with the presented system.

Additional findings respective to the individual interaction paradigms are identified. For instance, users of the location-based audio service find that listening to sounds, and to a less clear extent, music, enrich their walking experience; further, that sounds are easier to associate with specific locations than music, but that music-listeners were more engaged with the experience than sound-listeners [5]. The studies of tactile interfaces produced findings on differences in terms of usability of different interaction mechanisms depending on the user's physical activity [64]; and on the suitability of specific haptic-to-audio mappings [2]. A study investigating the use of emotion-based visual icons to represent music found that the presence of such icons improved playlist generation speed (task performance in the context of this particular evaluation), and further that evaluating valence from the visual representation is more difficult than evaluating arousal [157]. The evaluation of a zoomable user interface widget for music collection browsing [32] found significant differences in

completion times between users of large (PC) and small (portable) displays; however, comparing the number of data items was more difficult on smaller screens. The evaluation of a music information retrieval system based on user-driven similarity resulted in better task performance, and a greater feeling of control, compared to baseline comparison systems [147]. An evaluation of a music recommendation system based on dynamic weighting of content features and user access patterns suggests that a content-based approach is best for generating recommendations exhibiting content similarity; access-pattern and artist-based approaches give better diversity; with hybrid approaches capable of providing balance between the two [134].

2.4.5 Cultural differences

A few studies examine cultural differences in music information behaviour, e.g., between American and Chinese survey respondents [54], or between Western and Korean music searchers in a search log analysis [92]. Perception of the music may be culturally affected, with significant differences detected along a number of dimensions, e.g., between American and Chinese responses in terms of mood clusters chosen to correspond to stimulus songs, even taking into account demographic differences such as gender or age; listeners tended to agree more with others from their cultural background than those from another cultural background [54]. Information needs may be different; e.g., 36% of music searches queries conducted using Naver, a prominent Korean search portal, were looking for music recommendations, compared to only 5% of Google users. Correspondingly, the usefulness of specific query access points may be culture-dependent: Google searchers provided date, genre, lyrics, and

region information significantly more often, but "phonetic sound of lyrics" less often than Naver searchers. Other access points may be universal; e.g., there were no significant differences between the frequency of searches for associated use (i.e., involving a movie or advertisement featuring a song), by audio or video example, by name or gender of the artist, by the title of the work, mood/affect, and others [92]. The most common information need expressed across cultures was the identification of bibliographic information pertaining to artist and work. Issues of transliteration can cause issues for searches across language, where the correspondence between e.g., the English and Korean version of an artist name is not clear.

2.4.6 Interface interaction behaviour

A large number of findings concern user interactions with different query mechanisms and interfaces. The focus in these cases is not the evaluation of a particular system, but rather, the understanding of user behaviour when using various types of interface.

An investigation of user interactions with a hypothetical "query by voice" mechanism demonstrates that users may conduct musical known item queries by reproducing melodies using sung lyrics, or nonsense syllables when lyrics aren't known; by humming or whistling; by hand percussion, and by spoken comment. Given the choice, a small majority of participants in this investigation adopted one of these methods, while the rest combined multiple mechanisms (e.g., singing lyrics and syllables); query lengths varied with the chosen mechanism, with lyrical queries significantly shorter than queries presenting nonsense syllables. Query mechanisms

producing pitch intervals were chosen predominantly over rhythmic or spoken approaches [105].

In a playlist creation scenario, users describe songs according to lyric-based features [140]. Tempo and rhythm are among the most important audio-content based features in (manual) playlist creation. Participants tend to use longer strings when providing keyword descriptions of music, compared to query string length when searching for music, and tend to use more concrete symbolic or metaphorical categories referring to nature or particular objects [67]. Tools to support playlist creation are desired by certain users, as they lower the barriers to playlist creation, making the task easier and shortening the time required [132]. Playlist variety is desired, with song similarity merely one of many different factors involved in playlist ratings [89].

In a search context, participants instead tend to describe music as "for, about, or on" certain occasions, events, or specific activities [67]. Participants conducting searches for musical information seek bibliographical information in a large majority of cases [7, 30, 82]. Such searches tend to involve performer name, song title, notions of date, lyrics, and more rarely, orchestration info, collection title (e.g., album name), or label name [7, 30, 82, 91]. Related but less straightforwardly bibliographical search vectors may be useful, for instance when links between artists are implicit by a known, liked artist citing another as an influence [82]. In specialist circumstances, non-bibliographic information gains significance. For instance, most participants in an investigation of the users of a music library required historical information, for example collector's manuscripts or notes [55].

Queries for music information tend to be longer than average web queries [23], likely because such queries include terms from song / album titles and artist names, which tend to be longer. Music search engines must be resilient to errors and confusions; inaccuracies abound, with spelling errors in song titles and artist names, and confusions between the chorus lyrics and the song title, or between similar-sounding artists, occurring commonly. Lyric queries are particularly prone to error, with inaccuracies including missing word(s), additions of word(s), misspellings, errors in contraction, pronoun use, preposition use, tense, and (most commonly), confusion of similar sounding or semantically similar words [86].

Searches are often conducted for reasons beyond single-item identification, such as to assist in the building of collections [91]. In the maintenance of physical collections (e.g., CDs), people employ various different strategies in organizing their main active collection: by date of purchase, by release date, by artist, by genre, by country of origin, by degree of liking, or in order of recency that the CDs have last been listened to. Secondary organization may be applied within top-level categories, e.g., sorting by artist within genre; rarely is the classification scheme more than 2 levels deep [28].

Complex correspondences may form between the listener's mood and their interactions with a particular interface. For instance, serendipitous music encounters through random "shuffle mode" listening may be recognized as a happy coincidence, raising the listener's mood [104]; or be baffling or surprising, capturing the imagination [102]. Mood may influence song selection behaviour; for instance, once users

skip a song during active listening, they are disproportionately likely to skip a few more, due to a lowered aspiration level [13].

2.4.7 Notions of genre / style

Musical genre emerges as a useful, wide-spread, but problematic notion in music information retrieval. The definition of a genre taxonomy is non-trivial [30, 138], and genre classifications are neither consistent nor objective [146]. While strong agreement has been demonstrated between expert-created taxonomies and crowdsources "folksonomies" for certain genres, including "blues" and "hip-hop", other genres (e.g., "rock") exhibit disagreement [138].

Algorithmic approaches to genre classification perform well in certain cases: for instance, genre clusters generated according to probabilistic reasoning over large amounts of listening data generated by Zune users [158] correspond well to certain pre-defined genre notions, such as "Latin" and "Electronic/Dance"; while other clusters comprise mixtures of high-level genres, e.g., the combination of "Electronic/Dance", "R&B", "Pop", and "World".

Music listeners often describe their listening habits in terms of genre affiliation, and indeed user studies commonly prompt for these in their investigations. Listeners typically listen to music from several different genres [39, 106]. One study identifies some demographic links with genre choice, with adolescents preferring "Pop" and "Dance" music over other genres [119]; other genres may be collectively deemed unimportant or redundant by some participants (e.g., "Gospel", "New Age") [48] . In one experiment, queries according to taxonomic style produced highly relevant results, outperforming queries by music example or by music group; however, there

was a drop-off in precision over time, as the number of available inputs was limited [77].

Notions of musical genre extend into extra-musical content. Participants make inferences about musical style based on album cover images, although these can be misleading [30]. There is some evidence suggesting significant associations between certain genres and fonts [47] and even colours [48]; in particular, there is significant agreement between participants on which fonts look "Metal", and that the colour associated with this genre is "black".

2.4.8 Interacting with others

Several studies examine users' online interactions with other individuals in a music information seeking context. In a study of music-related queries to Google Answers, shorter queries were significantly more likely to be answered by other users, suggesting that conciseness is valued in such exchanges; by comparison, the mean monetary value offered in exchange for an accepted response did not significantly affect this likelihood [95]. An analysis of postings to a music message board reveals social and contextual elements relating to the associative or environmental context of the desired musical information appearing in almost a fifth of all postings analysed. Such queries offer qualitative descriptions that users use to contextualize their queries - "Heard it at a couple of jam sessions"; "used to sing while bowling"; "last week at the Fraley festival" - that provide vivid detail of the particular information need, but may be difficult to quantify [35].

In terms of physical interactions with information professionals, several studies have investigated user's search behaviours in locations including music libraries [24,

30, 55, 79, 141] and record shops [24, 30, 82, 141]. Although such locations typically benefit from the availability of information professionals or shop assistants capable of answering questions expertly, some participants demonstrated a reluctance or an embarrassment in engaging in such interactions [30], especially in any kind of "query by voice" (i.e., singing a musical query) [24]. That said, the expertise of such individuals is valued, e.g., with recommendations by staff in small, specialist record stores being trusted above expert reviews in music journals [82].

2.4.9 Music perception

A few studies within this subset of the literature have shed findings on interactions of music perception and cognition, and music information retrieval. One preliminary study suggests that participants shared similar perceptions regarding fundamental aspects of music, including tempo and brightness (timbre), with only small differences between individuals [159]. Tempo appears to have an effect on users' choice of mood tags, with faster versions of songs associated with positive tags (happy / energetic), while slower versions of the same songs were associated with more negative tags. Instrumentation also had an effect, with distorted guitars associated with tags such as "aggressive", while e.g., cover versions featuring banjos were associated with "soothing" and "relaxing" [94]. User-supplied mood tags appear to follow a long-tail (Bradford) distribution, the majority appearing only once, with only a smaller set of core mood terms used heavily.

2.5 Conclusion: Relevance in MIR

The creation of rigorous and practicable theories concerning the nature of relevance has been identified as a key challenge to the field of music information retrieval

(Downie, 2003). In this chapter, we have sought to address this challenge by exploring the state of knowledge on music relevance available in the literature. Cross-applying and extending an established conceptual model of relevance from textual information retrieval—Saracevic's stratified model of relevance interactions (Saracevic, 2007b) — to the domain of music, we have provided a highly detailed description of the aspects of relevance studied, and the findings generated, in the user-centric music information research literature. As our central object of study, we have adopted a collection of research articles subject to previous bibliometric analysis and shared in the spirit of collaborative and transparent research synthesis in previous research (Lee & Cunningham, 2012). We reciprocate by making our corpus of coded studies and findings available, accessible and queryable through a simple user interface. We hope that the outcomes of this process will be useful to future MIR user research, providing a reference of current knowledge, a tool for hypothesis generation, and a lens to promote the "synergic impact" (as per Lee & Cunningham, 2012, p. 396) of these studies in terms of implications for MIR system design. Further research on the present corpus might usefully extend Lee and Cunningham's analysis of the research study designs and methodologies employed in the articles under review, in terms of the findings generated. This could in turn suggest opportunities for the diversification of investigations of particular relevance interactions that may thus far have been predominantly studied from only a limited set of perspectives.

Some limitations of our work remain to be addressed. As a consequence of building our review on a previously established collection of articles, we are bound to the consequences of the selection strategy employed in the previous research. As

Lee and Cunningham note, the articles are widely dispersed in various publication venues represented in multiple databases tied to different fields, as a consequence of the interdisciplinarity of MIR research: "despite our best efforts, we would not be surprised if there were studies we were not able to find" (Lee & Cunningham, 2012, p. 396).

A further consequence of our building on this pre-existing collection is that the latest research under review is already some three years old at time of writing. Future research could address this by periodically extending the collection of articles using a compatible search strategy. Alternatively, authors of MIR user research studies might be interested in contributing new findings from their own work to this collection; perhaps this could form part of a sustainable archive of citation information for user studies related to music, managed by multiple stakeholders, as Lee and Cunningham have proposed.

Another limitation of our work is the inherent degree of subjectivity necessitated by our approach. We have worked to promote consensus in our coding activities, by beginning with an iterative series of parallel coding of the same articles by all four coding researchers in isolation, followed by collaborative discussion and review, until a consensus coding approach was established. Further, we have employed a coding system that automatically propagated new classification terms among the four researchers as and when they were necessitated in the coding activity of any individual coder. Nevertheless, the final decision of which parts of an article should be identified as a relevance-related finding, and the precise configuration of descriptors by which the finding should be incorporated into the conceptual framework, remains subject

to the judgement of the individual. We thus acknowledge that four different individuals would likely not have produced an identical outcome; however, we maintain that the overall shape of the distribution of studies and findings would likely remain comparable, given the systematic methods we have employed. Further, we hope and expect our detailed corpus of findings to be useful to future research, regardless of the degree of individual judgement involved in its creation.

The remainder of this dissertation shifts focus from the broad conceptualization of relevance in the MIR literature provided in the preceding discussion, to the investigation of very specific aspects of music and relevance. In particular, we investigate concerns at the intersection of MIR and music perception and cognition focussing on temporal facets of music, an area that has remained under-explored in the literature (section 2.3.2, p. 61). We now turn to Chapter 3, where we present a series of experiments investigating the role of rhythmic information in melody identification, drawing implications for the formulation of topical relevance criteria for MIR systems along the way.

CHAPTER 3
The Role of Rhythmic Information in Melody Identification

The human ability to recognize familiar melodies is curiously resilient to a wide range of transformations (Dowling & Harwood, 1987; White, 1960). Melodies are recognized under key transposition (Ehrenfels, 1937), alterations of tempo (Andrews, Dowling, Bartlett, & Halpern, 1998), and even complex, non-linear distortions (Hébert & Peretz, 1997; White, 1960). Melody identification tasks have been used to probe the nature of melodies as perceptual objects in the context of cognitive science (e.g., Schulkind et al., 2003). Insights on melody identification also have practical importance in the field of information science, in particular for music information retrieval, in which digital signal processing, feature extraction, and computational classification processes operating on musical audio must aim to closely match human judgements in order to produce results relevant to the user (Byrd & Crawford, 2002; Downie, 2003).

Consideration of the invariance of melodic recognition under transposition is evident in the earliest thought in Gestalt theory (Ehrenfels, 1937). The theory of indispensible attributes outlined by Kubovy and van Valkenburg (2001), a modern elaboration on Gestalt theory, provides a conceptual framework for the definition of edges—or boundaries—of perceptual objects. Consideration of melodies as exemplars for category membership and the degree of family resemblance among various

instances of a melody and its distorted versions is consistent with Rosch's prototype theory (1978).

White's pioneering investigation of distorted melodies (1960) explored perceptual objects in a musical context. In his study, participants attempted to recognize distorted versions of ten familiar melodies from a list revealed at the start of each session. Distortions were systematically performed on individual musical attributes of each melody, including pitch and rhythm. Participants were presented with either the first 24 notes or the first six notes of each stimulus. From the perspective of prototype theory, the task amounts to placing each presented stimulus into one of ten distinct target categories, corresponding to the ten familiar melodies. The prototype for each category was represented by the undistorted, canonical version of the corresponding melody; this was not played back to participants during the experiment, but was assumed to be available as an internal mental representation to each individual (Halpern, 1989; Levitin, 1994). Each distorted melody occupied a particular location within a continuous frequency-temporal space subsuming these categories.

White's distortions included various alterations of pitch information, e.g., doubling the size of all intervals, or imposing *equitonality* by setting all pitch intervals equal to zero, thus assigning each note a constant pitch value; and one alteration of rhythmic information, imposing *isochrony* by setting all note durations to a constant value. The assumption is taken that the equitonal and isochronous conditions respectively negate pitch and rhythmic cues. Only one type of alteration was applied in each experimental trial, so that rhythmic information was preserved when pitch was

altered and vice versa. All stimuli were generated by a computer program, affording the reproduction of each stimulus without expressive variation in tone, loudness, or accentuation.

White reported relatively minor impairment of melody recognition rates in the pitch-retaining isochronous condition (88% correct in the 24 notes condition; 60% correct in the six notes condition), compared with recognition rates of undistorted stimuli (94% correct in both the 24 notes and six notes condition). In contrast, alterations of relative pitch interval sizes resulted in the strongest impairment effects, with the rhythm-retaining equitonal condition producing the lowest recognition rates (33% correct in the 24 notes condition; 32% correct in the six notes condition).

A related study by Hébert and Peretz (1997) investigated the relative contributions of pitch and rhythm to the task of melody identification in a series of experiments. Whereas the participants in White's study recognized distorted versions from an explicit closed set, the list of 10 familiar melodies presented at the start of each session, Hébert and Peretz's participants were required to draw upon their long-term memory for free identification.

In their first experiment, participants were tasked with identifying melodies that were altered using either equitonal or isochronous distortions, as used by White (1960), as well as melodies presented in their unaltered condition. Participants were asked to respond with the name of the melody where possible. They were also required to provide a "feeling of knowing" rating on a Likert scale ranging from 1 ("I did not recognize it at all") to 5 ("I did recognize it very well"). Identification

rates were found to be substantially higher in the isochronous condition (M=49%) compared to the equitonal condition (M=6%).

Two further experiments were conducted employing identification tasks to contrast pitch and rhythmic information: an investigation of responses to "chimeric" stimuli presenting the pitch sequence of one melody with the durational sequence of another melody; and a response time task investigating the contributions of top-down processing featuring the stimuli of the first experiment alongside a visually presented song title that either corresponded or did not correspond to the stimulus melody, with participants instructed to respond as quickly as possible on whether the title matched the melody. Results of both studies strongly emphasized the importance of pitch information over rhythmic information in melody identification: participants instructed to attend to the pitch components of the chimeric stimuli in the second experiment outperformed those attending to the rhythms in terms of identification success (pitch: M=53%; rhythm: M=8%)—indeed, participants instructed to attend to the rhythm and ignore pitch nevertheless tended to name the pitch-contributing melody (M=28%)—and performance on both response time and accuracy were more severely impaired in the pitch-disrupting equitonal condition than in the rhythm-disrupting isochronous condition in the third experiment.

Based on these results, Hébert and Peretz concluded that while the unaltered condition with its combination of rhythmic and pitch cues provides optimal grounds for melody identification from long-term memory, pitch strongly predominates over rhythmic information in terms of enabling identification when isolated in the altered conditions.

Newton (1990) investigated the ability to *convey* musical identity using only rhythmic information. Motivated by problems of cognitive bias in the construal of intention in inter-personal communication, Newton set up a study in which participants were grouped into pairs of "tappers" and "listeners". Pairs were seated with their backs facing each other. Tappers selected three targets from a list of 25 well-known melodies, and tapped out the rhythm of each target for the listener. Listeners were tasked with identifying and writing down the names of the corresponding melodies. After tapping each song, tappers were asked to estimate the likelihood that the listener would identify the song correctly.

Newton hypothesized that tappers, drawing on their memories of the melodies in question and accessing a rich, multifaceted internal representation of the music, would tend to overestimate the listeners' chances of identifying the tapped rhythms. Estimates ranged from 10% to 95%, with a mean of 50%. In actual fact, there were only three instances of identification success in all 120 trials of the experiment—a success rate of 2.5%.

Cast in the terms of prototype theory, these preceding studies investigate the cue validity of the pitch and rhythmic information presented by the experimental stimuli. In White's study, the cue validity of each stimulus can be quantified empirically because the set of possible categories is limited to the ten melodies presented to participants. Each pitch class, pitch interval, or durational value contained within a particular stimulus can thus be exhaustively mapped against all melodies in the set. In the studies of Newton, and of Hébert and Peretz, participants must draw on an open set of potential categories (only Newton's tappers are given the list of songs;

the listeners remain uninformed), and thus such calculations are not practicable; however, the concept of cue validity may still guide the interpretation of results.

The outcomes of these preceding studies led previous investigators to infer a substantially diminished role for rhythm compared to pitch in melody identification. In the studies employing alterations of musical facets, the predominant method of rhythmic manipulation is the imposition of isochrony—setting all note durations in the melody to a constant value while retaining pitch information—resulting in a metronome-like durational structure. In the case of rhythmically heterogeneous melodies, this manipulation affects the durational structure both at the level of individual notes, and on the metric level; as metric boundaries are blurred, previously unaccented note events may be relocated to metrically "strong" positions through the process of subjective rhythmization; likewise previously accented notes may transfer to metrically "weak" positions.

The isochronous manipulation is applied on the assumption that it effectively nullifies rhythmic information, and that pitch thus remains as the only musical facet providing contributions toward identification success. However, highly familiar melodies as used in these studies often feature very simple, homogeneous rhythms; for instance, the rhythmic structure of "Twinkle Twinkle Little Star" is composed predominantly of quarter-note durational elements. Melodies undergoing isochronous transformations may thus retain large parts of their original rhythmic structure after distortion (Schulkind et al., 2003). The assumption that successful identifications in such cases rely on cues from pitch in isolation, without incorporating rhythmic

information, is thus unwarranted. This puts previous claims about the relative contributions of pitch and rhythm into question, and creates a gap of empirical evidence in this area. Further, all results of the previous studies have been reported in aggregate across all melodies in a particular condition, making it difficult to draw conclusions about participants' performance on individual melodies. We thus cannot determine whether, for instance, the degree of homogeneity present in individual rhythms had an effect on identification rates in altered conditions, or whether certain stimuli may have been more readily identifiable by rhythm alone.

The present study aims to reassess the cue validity of rhythmic information in the context of melody identification. We build on the methodological approach of the melody distortion studies outlined above by applying complex, stochastic manipulations of durational elements in order to more thoroughly minimize rhythmic cues. As we are interested in memory identification from long-term memory, we follow Hébert and Peretz' approach of not revealing the set of target melodies to our participants. Empirically quantifying cue validity is thus impracticable, as the set of possible categories effectively includes every melody that a given participant happens to be familiar with. However, it is still possible to quantify the degree of dissimilarity between each rhythmically distorted stimulus and its prototypical target melody using algorithmic and mathematical similarity metrics developed in music information retrieval and related fields. Measures obtained through such means can ensure that a given manipulation does in fact introduce sufficient levels of distortion to justify claims about the nullification of rhythmic cues.

Familiar melodies also carry connotative qualities through associations in episodic memory, aside from denoting a particular sequence of pitches and durations. The precise nature of such associations will vary between individuals. Given that the melodies in these studies were selected for inclusion based on being commonly known among members of the sampled population, certain associations are likely to be consistent across participants through shared processes of enculturation. Although participants are tasked to respond with the name of the given target, categorization is thus likely to occur on levels of abstraction beyond that of the individual melody. For example, a certain sequence of pitches and durations may be identified not merely as the melody to "Rudolph the Red Nosed Reindeer", but also, at a more abstract level of categorization, as being connotative of "Christmas music". Similarly, a distorted version of the "Wedding March" may be recognized not only as corresponding to that particular melody based on pitch and durational similarity, but also as belonging to the higher-level category of "ceremonial music". Examinations of personal music collections (Cunningham et al., 2004) have indeed revealed music organization principles based on context: People organize music on the basis of the situation in which they intend to listen to a particular set of music (e.g., "work music", "driving music").

This distinction between cues from elements of the constituent structure—in our case, durations and pitches—and contextual, connotative cues is discussed in the case of food categorization by Ross and Murphy (1999), who show that food can be cross-classified either into *taxonomic categories* relating to constituent structure (e.g., yoghurt as a dairy product) or into *script categories* relating to the context of

use (e.g., yoghurt as a breakfast item). In the case of everyday listening, Guastavino (2007) and Dubois (2000) provide converging evidence that the categorization of environmental sounds relies on experiential knowledge of the context in which everyday sounds are typically encountered, giving rise to goal-driven *script categories* in the auditory domain, integrating notions of time, location, and activity.

The present study investigates the role of rhythmic information in the context of melody identification by analyzing participants' identification performance under different conditions of rhythmic distortion. We further investigate the applicability of notions of script categorization in musical memory by analyzing instances of partial identification and of misidentification of melodies. Our goals are to contribute insights on the nature of perceived melodic identity. By investigating contributive factors to melody identification success, we also aim to inform notions of experiential relevance of musical information. The operationalization of such notions has been identified as a key challenge to the field of Music Information Retrieval (Downie, 2003).

3.1 Study 1: Randomized and reordered conditions

3.1.1 Participants

Participants (N=31; 18 female; mean age: 25.3 years, SD=7.4) were members of the McGill University community with no reported hearing deficit. As the stimuli used in this study are culturally specific (e.g., nursery rhymes, folk, and pop songs), participants were required to be native English speakers, or to have learned English during early childhood. Participants varied in their degree of musical training (mean years of musical training: 5.9, SD=5.6), assessed using a modified version of the

Queens University Musical Experience questionnaire (Chapados & Levitin, 2008; Cuddy, Balkwill, Peretz, & Holden, 2005). Participants either received course credit or a financial compensation of $10 CAD for their participation.

3.1.2 Norming study

Stimuli were generated from short excerpts of 38 English language songs that had been established as familiar to a separate group of individuals drawn from a similar participant pool in previous research (Kim & Levitin, 2002). In that study, a list of popular songs used in previously published studies on song recognition (Andrews et al., 1998; Halpern, 1988; 1989; Hébert & Peretz, 1997; White, 1960) was administered to 600 students in an undergraduate class at McGill University. Participants were asked to indicate those songs "that you know so well you can hear them playing in your head", and to rank the 20 best known songs from 1 - 20. In addition, they were given the opportunity for free response, to add up to five songs to the list that they felt they knew well. These results were tallied and from them a list of 42 songs was generated. Monophonic piano versions of these 42 songs were presented to a group of 44 pilot participants and four songs were eliminated from the set due to poor recognition (fewer than 30% of the participants), thus leaving a stimulus set of 38 songs (see Appendix A).

3.1.3 Materials

Each melodic excerpt was performed monophonically by a professional musician using a MIDI keyboard. All expressive information was removed from the resulting MIDI files; this was done using the MIDI editor in ProTools (Digidesign, Daly City, CA), to quantize note durations and to set all note velocities—corresponding to the

force with which the note is played—to the same constant value (as per Bhatara, Tirovolas, Duan, Levy, & Levitin, 2011). Distortions were then applied, according to experimental condition, using a Perl script and the MIDI-Perl module (Burke & Conklin, 2010). The resulting files were synthesized to WAV format using the Akai Steinway III piano SoundFont (AKAI Professional, Tokyo, Japan) to create the final stimuli.

3.1.4 Stimulus conditions

Each melody was distorted, or altered, in two different ways, leading to three experimental conditions: the two alterations plus the original, unaltered stimulus. The alterations involved systematic changes to the durations of MIDI events (notes and pauses) according to experimental condition. Only event durations were affected; other information, in particular pitch chroma and height, and the ordering of pitches, remained unchanged.

In the first altered condition ("*reordered*"), durations were randomly re-assigned among the MIDI events in a given melody. In the second altered condition ("*randomized*"), each event's duration was set to a random MIDI-tick value, limited by the shortest and longest note durations present in the unaltered melody. In the *unaltered* condition, event durations remained unchanged.

The generated files are likely to differ with each execution of the script, as both altered conditions introduce stochastic elements. The degree of rhythmic distortion varies accordingly with each random outcome. In order to control for this variation, 100 distorted versions were generated for each combination of melody and condition. A measure of rhythmic similarity—the chronotonic distance—was determined for

each version, and the version with the highest distance for each such combination was selected for inclusion in the study.

The chronotonic distance measure was obtained by representing the event durations of the distorted and undistorted versions of a given melody as histograms, with the vertical axis showing inter-onset intervals and the horizontal axis showing onset times; this is known as "TEDAS representation" (Gustafson, 1988). Each distorted version's TEDAS representation was then superimposed over the undistorted version's histogram. The area of difference between the two representations was then determined to obtain the chronotonic distance (Guastavino, Gómez, Toussaint, Marandola, & Gómez, 2009). The choice of this measure was motivated by Toussaint (2006), who demonstrated the chronotonic distance's suitability in providing insights on the structural interrelationships that exist within groups of related rhythms.

3.1.5 Procedure

Experimental sessions were conducted with small groups of between two and six participants. The participants were seated in a quiet room facing a projector screen. Each participant was given a pen and a lined piece of paper; the lines were numbered sequentially, one line for each experimental trial.

Participants were informed that they would be listening to a number of commonly known melodies that had been distorted in different ways. They were told to write down the name of each recognized melody. If they could not remember the name, they were asked to write down any other identifying information that comes to mind; this might include song lyrics, artist name, or contextual information about where one would expect to hear the melody. If no identifying information came to

mind, but they nevertheless felt that they recognized the melody (i.e., if they experienced a "feeling of knowing"), they were asked to respond with the word "Yes"; if they simply did not recognize the melody, they were asked to respond with the word "No". They were also informed that an undistorted version of each melody would be presented at the end of the experiment, to verify whether they actually knew the individual songs.

In each trial, a stimulus was presented over the loudspeaker in the room (Anchor LN-100 Powered Monitor) at a comfortable listening level of approximately 70 dB(A). The corresponding trial number was projected onto the screen as the stimulus was playing. Participants were tasked to write down their response on the paper at the corresponding line as soon as they were ready. The next trial started playing automatically after a silent period of six seconds following the end of the previous trial. The screen briefly changed color at one second prior to the end of this period to alert the participants to the imminent start of the next trial.

Two pseudo-random orderings of all melodies were generated and respectively assigned as the presentation order of the two altered conditions, in counterbalanced fashion between experimental sessions. The two orderings were then interleaved, so that participants were exposed to the experimental stimuli with conditions in alternating order. Care was taken during the generation of presentation orders to ensure that the two altered versions of a given melody would never be presented consecutively.

Participants listened to the unaltered stimuli after completing trials for each altered stimulus in this sequence. Each experimental session comprised a total of

114 trials, corresponding to two distorted and one undistorted presentation of each of the 38 melodies. Experimental sessions lasted approximately one hour, including time for the questionnaire on musical background and a debriefing.

3.1.6 Classification

All target melodies in the stimulus set, as well all other melodies specified in participants' responses, were classified by mutual agreement of the coauthors into one of five categories: Children; Christmas; Pop; Theme (e.g., from a movie or television show); and Ceremonial (e.g., the "Wedding March", or "Happy Birthday"). A full list of melodies and their classifications is included in Appendix E.

3.1.7 Scoring

Three raters independently scored two thirds of all observations following the guidelines described below. Observations were assigned to raters in such a way that every participant response was scored by two of the raters. Score parity was achieved in 96.9% of cases; the primary author resolved the remaining individual inconsistencies.

Responses were scored by assigning a value of 1 to a correctly identified melody, 0.5 to a partially identified melody, and 0 to a wrongly identified or unidentified melody. The aim was not to assess the participants' recollection of song names, but rather the correct identification of the song itself. As such, responses that included names that differed from the official name but nevertheless unambiguously represented the same song were scored as correct; e.g., "Barney" was an acceptably correct response for "Yankee Doodle", as the theme tune for children's TV show "Barney the Dinosaur" shares "Yankee Doodle's" melody. Responses that included

lyrics narrowly matching those of the target melody were also scored as correct identifications; e.g., "life is but a dream" was deemed an acceptably correct response for "Row, Row, Row Your Boat". Similarly, responses correctly naming composers or performers—e.g., "Beethoven" as a response to "Ode to Joy", or "Beatles" as a response to "Hey Jude"—were deemed to be correct and received full credit. Finally, responses that correctly described a very narrow context for the melody were also scored as correct identifications; e.g., "From [the film] 2001: A Space Odyssey" was deemed an acceptable response to the "Blue Danube Waltz", which is featured during an iconic scene in the film.

Responses that correctly described a higher-level contextual category for the melody—e.g., "Christmas" for "Hark! The Herald Angels Sing"—were deemed as partial identifications, receiving a score of 0.5. Responses that made no attempt at identifying the melody, specified a melody clearly different from the target, or otherwise provided non-matching information (e.g., a mistaken category or unrelated lyrics) were deemed as unidentified, receiving a score of 0.

3.1.8 Analysis

Withheld observations. Two of the 38 melodies—"Coming 'round the Mountain" and "If You're Happy and You Know It"—were mistakenly presented in their reordered versions during both the reordered and randomized condition due to technical problems. Responses for these two melodies were thus withheld from analysis.

Identification. Participants' rate of identification of the unaltered stimuli was variable (proportion of successful identifications: *Median*=.71, *M*=.68, *SD*=.24), suggesting that individuals differed in terms of the number of melodies in our set

that were highly familiar to them. It would be erroneous to penalize an individual's incorrect responses to altered stimuli in cases where the corresponding unaltered melody is unknown to the individual. Taking the assumption that a failure to identify an unaltered stimulus indicates a lack of sufficient familiarity with the corresponding melody, any responses to melodies that were not correctly identified in their unaltered version were excluded from analysis per-participant.

A proportional odds analysis was conducted to analyze the resulting dataset using the R statistical programming language (R Core Team, 2015) and the clm function of the ordinal R module (Christensen, 2013), fitting a cumulative link model with response score (0, .5, or 1) as the dependent variable and condition (unaltered, reordered, or randomized) as a treatment effect. The condition factor was coded to contrast between the unaltered and altered (reordered and randomized) conditions, and between the two altered conditions.

Partial identification and misidentification. Responses indicating a contextual category without naming the target melody were investigated by comparing each target's category as determined in the scoring stage with the categories indicated in participants' responses. Misidentifications—responses specifying a melody that differs from the target—were similarly investigated in terms of category pairings, by comparing each target's category with the categories determined during scoring for the melody specified by the participant. Furthermore, target-response pairings were analyzed directly to determine recurring misidentification patterns.

"Feeling of knowing" responses. Responses that only indicated the presence or absence of a "feeling of knowing" without any further specification were

analyzed by calculating the sensitivity index (d') for all "yes" and "no" responses in the two altered conditions. In order to arrive at a ground truth, we assume that the ability to identify a melody in its unaltered form indicates true familiarity with that melody. Accordingly, a "yes" response to an altered version of a melody correctly identified in the unaltered condition was treated as a hit, and a "yes" response to an altered stimulus that was not correctly identified in the unaltered condition was treated as a false alarm.

Correlation with algorithmic measures. Two alternative measures of the rhythmic dissimilarity between the unaltered and distorted stimuli were determined, in addition to the chronotonic distance measure used in stimulus generation, in order to investigate the degree of correlation between these metrics and participants' identification performance. The first of these, *Dynamic Time Warping* (DTW), was chosen because it makes no assumptions about the two sequences to be compared in terms of their conformity to standard musical structure. The second, *rhytfuzz*, was included based on its performance in related musical contexts.

A MATLAB implementation of the DTW procedure (D. Ellis, 2003) was used to determine the amount of modification (i.e., stretching or shortening of segments of the melody) required to align each distorted stimulus with its unaltered version. As mentioned above, the DTW metric may operate on any temporal sequences with no assumptions about musicality, and has been applied in diverse non-musical contexts such as word recognition (Myers, Rabiner, & Rosenberg, 1980), gait analysis (Boulgouris, Plataniotis, & Hatzinakos, 2004), and the classification of whale vocalizations (Brown, Hodgins-Davis, & Miller, 2006) as well as in musical contexts (e.g.,

Kroher, 2013; Macrae & Dixon, 2010). This motivated its inclusion in the present study, given that the distortion process applied in the randomized condition results in sequences that are unlikely to conform to standard musical durational patterns. Melodies altered to exhibit such randomized durational elements are likely to be deemed metrically ill-formed by participants (Prince, 2011).

Further, the rhytfuzz measure of Müllensiefen and Frieler (Müllensiefen & Frieler, 2004) was determined for each stimulus. This measure "fuzzifies" rhythmic information by placing each durational component of the stimulus into one of five categories, ranging from *very short* through *normal* to *very long*, where the center of the "normal" category is determined by the most commonly occurring duration in the stimulus. An edit distance is then determined to quantify the minimum number of edits (deletions, insertions, or substitutions) required to transform the fuzzified durational sequence of the distorted stimulus into that of the unaltered original. The rhytfuzz measure was determined using the SIMILE software application (Müllensiefen & Frieler, 2006). This measure was included in the current study as it was the most successful purely durational measure in a previous evaluation of melodic similarity metrics against human judgements (Müllensiefen & Frieler, 2004).

3.1.9 Results

Identification. Identification was inhibited in both altered conditions, with the randomized condition inhibiting identification scores to a greater degree than the reordered condition (randomized: M=.30, SD=.46; reordered: M=.57, SD=.49; see Figure 3–1). The results of the proportional odds analysis reveal these differences to be significant: the altered conditions significantly differed from the unaltered

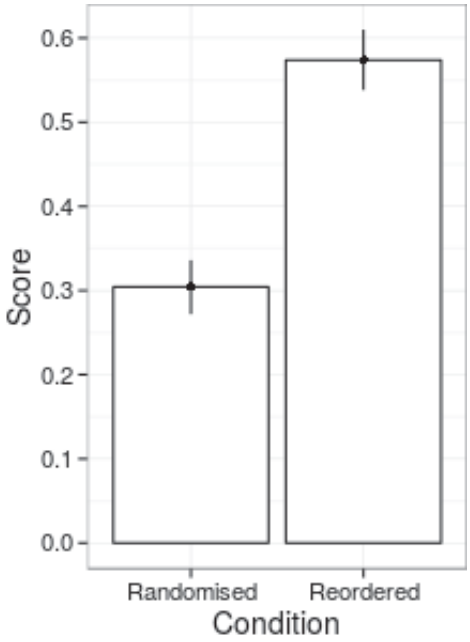


Figure 3–1: Aggregated cross-participant identification scores for stimuli in the randomized and reordered conditions of Study 1. A score of 1 indicates successful melody identification; 0.5 indicates partial identification; and 0 indicates misidentification or no identification attempt. Melodies not correctly identified in their unaltered condition are excluded, per-participant. Point ranges indicate bootstrapped 95% confidence intervals (1,000 resamplings).

condition ($\beta_1 = 3.35, SE = 0.18, z = 19.12, p < .0001$), and from each other ($\beta_2 = 1.15, SE = 0.11, z = 10.81, p < .0001$). Appendix D lists identification scores by melody, aggregated over participants.

Partial identification. There were 174 partial identifications, where participants named a category without specifying a particular song. Of these, 45.4% occurred in response to an unaltered stimulus; 39.7% occurred in the reordered condition; and 14.9% occurred in the randomized condition. Participants tended to remain "in-category", correctly matching the category of the target song in 134 of the 174 instances (77%).

Misidentification. There were 158 instances of misidentification, where a participant explicitly specified a song other than the trial's target. Of these instances, 24.7% occurred as participants attempted to identify an unaltered stimulus;

36.1% occurred under the reordered condition; and 39.2% under the randomized condition. Interestingly, misidentification pairings between the three songs, "Old MacDonald", "Bingo", and "Yankee Doodle", accounted for 24 of the 158 misidentifications (15.2%; well above the conservative chance level of 2.9% that results when limiting the set of possible confusions to those melodies within our stimulus set). As was the case with partial identifications, participants tended to remain "in-category" with their misidentifications, specifying a song within the same category as the target 115 of the 158 times (72.8%).

"Feeling of knowing" responses. Results of the analysis of responses indicating only the presence or absence of a "feeling of knowing" (i.e., "yes" or "no") without any further specification reveal closely matched hit rates (H.R.) and false alarm rates (F.A.R.), and correspondingly very low sensitivity in both the reordered ($H.R. = .29, F.A.R. = .28, d' = 0.03$) and randomized ($H.R. = .12, F.A.R. = .1, d' = 0.12$) conditions. This suggests that in absence of attempted identification, participants' "feeling of knowing" of the distorted stimuli was not a reliable predictor of their familiarity with the unaltered melodies.

Correlation with algorithmic measures. Pearson's product-moment correlations were calculated comparing each of the three distance measures (chronotonic, DTW, and rhytfuzz) determined between each altered stimulus and its corresponding unaltered original with the mean identification score for that stimulus, aggregated across participants. For each distance measure, two correlations were determined, one for each altered condition (i.e., reordered and randomized). No significant correlations were found between the chronotonic distance and aggregated

identification scores in either altered condition (reordered: $r = .08, p = .67$; randomized: $r = -.18, p = .32$). A marginally significant trend was determined between the DTW distance measure and scores in the randomized condition (reordered: $r = -.24, p = .17$; randomized: $r = -.29, p = .09$). Conversely, a significant correlation was determined between the rhytfuzz measure and the reordered condition (reordered: $r = .56, p < .0005$; randomized: $r = .15, p = .4$).

3.1.10 Discussion

The results of Study 1 demonstrate the detrimental effect of rhythmic distortion on melody identification success. There is a significant effect of the type of distortion; while identification rates are adversely affected in the reordered condition that restructures the durational components of the melody, the effect is considerably stronger in the randomized condition that distorts these individual components, disrupting the rhythmic information contained in the melody.

The tendency for participants to remain within the target's contextual category in cases of partial identification and misidentification suggests that participants attended to stimuli at a higher level of abstraction—termed the superordinate level by Rosch—and exhibit a tendency to err within the membership of the target's category, at Rosch's basic level (1978). This implicates cues from contextual category as exerting an influence on melody identification, even in controlled conditions excluding lyrics, variations in orchestration, or other presented sources of contextual information.

The frequent co-occurrence of misidentifications between three songs—"Old MacDonald", "Bingo", and "Yankee Doodle"—is especially interesting given that

these songs are fairly heterogeneous and dissimilar in terms of pitch contour. They all share the contextual category of "children's songs," but this is the most common category for the melodies in our data set (see Appendix E). Further, there is some lyrical similarity with shared references to barnyard animals, but again this is a common theme in children's music. However, their durational structure is highly similar, especially up to the first phrase boundary of each melody; Schulkind et al. (2003) have found melody identification facilitated at phrase boundaries. This misidentification pattern is thus consistent with a pronounced attention to rhythmic cues, even in rhythmically distorted conditions, and suggests that such cues may predominate over cues from pitch information in these cases.

As a consequence of the optimization toward greater chronotonic distance in the generative procedure for the randomized condition, there is a greater overall duration of the resultant stimuli compared to the unaltered condition (mean duration of 548% compared to unaltered; SD: 284%). Previous studies have shown decreased identification performance on uniformly slowed melodic stimuli (Andrews et al., 1998). This suggests alternative hypotheses regarding the low identification rate in the randomized condition: The effect may be explicable by the disruption of rhythmic information, by the slow overall tempo, or by a combination of the two.

A second study was undertaken in order to address this issue. The *stretched* condition of Study 2 is designed to control for this potential confound introduced by stimulus duration. In this condition, each melody is uniformly slowed to match the overall duration of the randomized version of each melody, while maintaining

the durational structure (i.e., the sequence of ratios of note durations) of the undistorted version. Additionally, the *isochronous* condition is introduced to facilitate comparisons with previous studies in the literature, which typically make use of this metronome-like transformation. Whereas the reordered and randomized conditions of the first study were designed to disrupt the rhythmic information contained in the melodic excerpts, the distortions in Study 2 were designed to assess the validity and transferability of the findings in Study 1.

3.2 Study 2: Stretched and isochronous conditions

3.2.1 Participants

A new set of participants in Study 2 (N=29; 22 female; mean age: 20.8 years, SD=1.6) was recruited from the McGill University community. Participants were native English speakers or had learned English during early childhood, and had no known hearing problems. Their degree of musical training (mean years of musical training: 8.9, SD=4.7) was assessed as in Study 1 using the same questionnaire. As before, participants either received course credit or a financial compensation of $10 CAD for their participation.

3.2.2 Materials

The stimuli for the second study were generated from the same set of melody excerpts used in Study 1. The generative process was identical, except that the distortions applied to each excerpt's MIDI event durations differed between Studies 1 and 2.

3.2.3 Stimulus conditions

The second study presented stimuli in two altered conditions—*stretched* and *isochronous*—as well as in the *unaltered* condition also presented in Study 1. In the stretched condition, the tempo of each excerpt was uniformly slowed by multiplying each MIDI event's duration by a value held constant for all events within the individual excerpt. This value was determined so that the final duration of the stimulus matched that of the same melody's randomized condition in Study 1. The rhythmic structure remained undistorted, isolating the potential effects of presentation slowness on identification rate.

Study 2 also included an isochronous condition to investigate the extent to which our findings may be transferable to the contexts of previous studies employing different participant pools and stimulus sets. The alteration for this condition assigns a constant duration to all MIDI events across all melodies, resulting in metronome-like rhythms akin to those used in previous studies in the literature (Hébert & Peretz, 1997; Kuusi, 2009; Schulkind, 1999; White, 1960).

3.2.4 Procedure

The procedure from the first study was repeated identically for Study 2. The same presentation order was applied, interleaving stretched and isochronous conditions and counterbalancing this alternating sequence between experimental sessions. Again, the undistorted stimuli were presented after the completion of all distorted stimuli. As before, each experimental session lasted approximately one hour, including time for the questionnaire on participants' musical backgrounds and a debriefing.

3.2.5 Scoring

Scoring was undertaken by the same three raters and using the same guidelines as in Study 1. Inter-rater scoring parity was achieved in 96.7% percent of cases; the primary author resolved the remaining inconsistencies.

3.2.6 Analysis

The two melodies withheld due to technical problems in the analysis of the previous study were correctly presented in their isochronous and stretched conditions in Study 2, but responses to these melodies were again withheld from analysis to maintain consistency in cross-study comparisons. The responses of one of the 29 participants who had failed to complete the experiment were also withheld.

Participants' rate of identification of the unaltered stimuli, as in the previous study, was variable (proportion of successful identifications: *Median*=.68, *M*=.69, *SD*=.22), suggesting that individuals differed in terms of the number of highly familiar melodies in the set. As in the previous study, a proportional odds analysis was conducted in order to investigate the effect of rhythmic alteration on response score, and to determine whether the stretched and isochronous alterations differed significantly in their contribution to this effect.

Four two-sample z-tests for proportions were performed on mean identification rate to assess the significance of the different distorted conditions across the two studies. These tests corresponded to the four cross-study pairings of the reordered and randomized conditions of Study 1 and the stretched and isochronous conditions of Study 2.

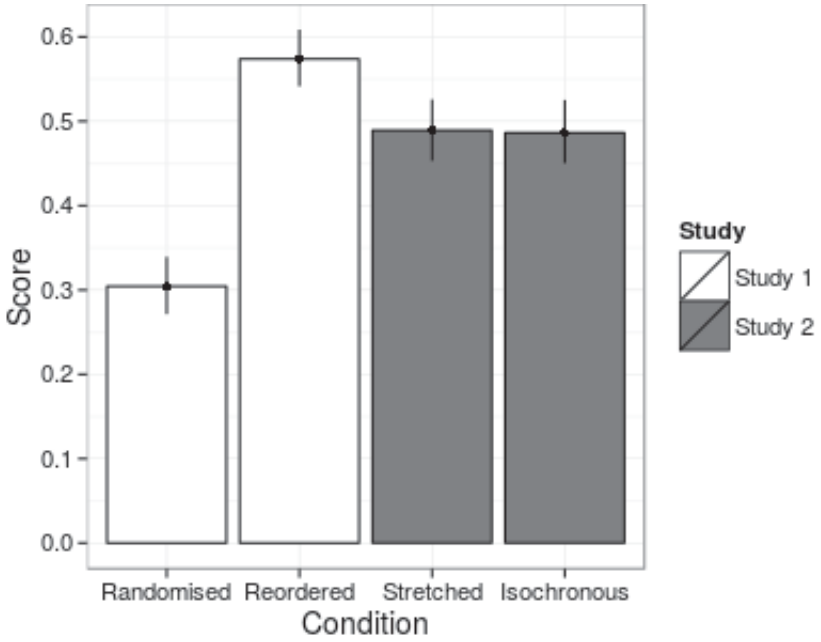


Figure 3–2: Aggregated cross-participant identification scores for stimuli in the altered conditions of Studies 1 and 2. All altered conditions significantly inhibited identification success. There are significant differences in identification scores across all condition pairings except those of Study 2 (isochronous and stretched). Melodies not correctly identified in their unaltered condition are excluded, per-participant. Point ranges indicate bootstrapped 95% confidence intervals (1,000 resamplings).

Partial identifications, misidentifications, "feeling of knowing" responses, and correlations with algorithmic measures were analyzed as in Study 1.

3.2.7 Results

Identification. As in Study 1, identification rates were inhibited in both altered conditions of Study 2 (stretched: M=.49, SD=.5; isochronous: M=.48, SD=.49). Results of the proportional odds analysis confirm a significant effect of the presence of rhythmic alteration on response score ($\beta_3 = 2.93$, $SE = 0.17$,

$z = 17.6$, $p < .0001$), but reveal no significant differences between the stretched and isochronous alterations ($\beta_4 = 0.011$, $SE = 0.11$, $z = 0.1$, $p = 0.92$).

Results of the cross-study two sample z-tests for proportions revealed significant differences in identification rates among all four cross-study pairings of distorted conditions: randomized and stretched, z=-7.24, p<.0001; randomized and isochronous, z=-7.13, p<.0001; reordered and stretched, z=3.27, p<.005; and reordered and isochronous, z=3.38, p<.001.

Partial identification. Participants responded with a partial identification (naming a category without specifying a song) 154 times. Of these responses, 37.6% occurred in the unaltered condition; 19.5% occurred in the stretched condition; and 42.9% occurred in the isochronous condition. Participants again tended to remain within-category, correctly matching the category of the target song in 120 of 154 cases (77.9%).

Misidentification. There were 152 instances of misidentification in Study 2. Of these, 30.9% occurred in the unaltered condition; 34.9% occurred in the isochronous condition; and 34.2% occurred in the stretched condition. Again, pairings between the three songs, "Old MacDonald", "Bingo", and "Yankee Doodle", accounted for a large portion of the misidentifications (37 of 152, or 24.3%). As was the case with partial identifications, and as in Study 1, participants tended to remain within-category with their misidentifications, specifying a song different from but sharing a category with the target in 114 of 152 instances (75%).

"Feeling of knowing" responses. Hit rates (*H.R.*) and false alarm rates (F.A.R.) for responses indicating the presence or absence of a "feeling of knowing" without any further identification attempt (i.e., "yes" or "no" responses) corresponded closely for both altered conditions, resulting in low sensitivity (stretched: $H.R. = .16, F.A.R. = .19, d' = -0.1$; isochronous: $H.R. = .32, F.A.R. = .30, d' = .06$). This suggests that participants' "feeling of knowing" of the distorted stimuli in absence of attempted identification was not a reliable predictor of their actual familiarity with the melodies, as in Study 1.

Correlation with algorithmic measures. Pearson's product-moment correlations were calculated comparing the chronotonic, DTW, and rhytfuzz distance measures between the isochronous and stretched stimuli and their respective unaltered originals with mean identification scores in these altered conditions, aggregated across participants. Neither the chronotonic distance measure nor the DTW measure correlated significantly with participants' performance in either condition (chronotonic distance—isochronous: r=-.07, p=.67; stretched: r=-.26, p=.14; DTW—isochronous: r=.15, p=.4; stretched: r=-.28, p=.1). The rhytfuzz measure's fuzzified rhythms are unaffected by uniform changes in melody tempo—i.e., identical fuzzy classifications would result for any corresponding stretched and unaltered stimuli—and thus, rhytfuzz was not determined for the stretched condition. However, the rhytfuzz measure's outcome correlated significantly with mean identification scores in the isochronous condition (r=.4, p<.05).

3.2.8 Discussion

The results of Study 2 reveal significant adverse effects of the stretched and isochronous distortions on identification rate. There are no significant differences in terms of identification performance between these two conditions. However, the adverse effect of both distortions is significantly stronger than that of the reordered condition and significantly weaker than that of the randomized condition of the first study (see Figure 3–2, p. 105).

A comparison of the results of the randomized condition of Study 1 with those of the stretched condition of Study 2 suggests that while the slow overall tempo of the randomized stimuli may inhibit identification success, a significant part of the adverse effect must be ascribed to the disruption of rhythmic information caused by the randomization of the durational components of the melody.

The findings regarding partial and misidentifications resemble those of the first study, and provide further evidence that participants attended to our stimuli at a superordinate level of contextual category membership. The tendency toward misidentification between the three melodies—"Old MacDonald," "Bingo," and "Yankee Doodle"—in the isochronous condition particularly demonstrates that the imposition of isochrony does not necessarily nullify rhythmic cues. Misidentification pairs chosen from these three melodies feature in 17 of the 58 misidentifications in the isochronous condition (29.3%); in contrast, only 6 of the 62 misidentifications in the randomized condition of Study 1 (9.7%) feature pairings of these songs.

3.3 General discussion: Rhythm in melody identification

Overall, the results of Studies 1 and 2 demonstrate both the inhibitory effect of rhythmic distortion on melody identification, and the greater disruptive effect of the randomized condition, compared to the other conditions. In particular, the significant differences in identification score between the randomized and isochronous conditions strongly suggest that the assumption that imposing isochrony nullifies rhythmic information in familiar melodies is unwarranted.

The significantly lower identification rates of the randomized condition compared to all other conditions, situated within the context of previous findings int he literature, provide new evidence regarding the relative effectiveness of pitch and rhythm as cues for melody identification. Findings from previous studies suggest that while these melodic facets interact in their contributions to identification, there is a considerable asymmetry in this interaction, with pitch information predominating over rhythmic information (Hébert & Peretz, 1997; White, 1960). Correspondingly, identification rates for tasks involving recognition from rhythm, e.g., via the imposition of equitonality (setting all pitches to the same frequency without altering rhythmic information) are dramatically lower than those for tasks involving recognition from pitch via the imposition of isochrony. The results of our randomized and isochronous conditions suggest that the contributions of rhythmic information have been underestimated in this previous work. While identification rates in our randomized condition approximately match those in the equitonal condition of White's study (24-note: 33% correct; 6-note: 32% correct), where participants recognized distorted melodies from a limited, revealed set, they are still markedly higher than

those in the equitonal task of Hébert and Peretz (6% correct) where participants had to access their long-term memory for identification, as in our study. Nevertheless, the extent of this difference is much reduced compared to isochronous identification rates in our Study 2 and in previous studies (see Figure 3–3).

Our participants' tendency to remain within-category in cases of partial identification and misidentification further demonstrates that the influence of connotative, contextual cues beyond the pitch and durational cues in the constituent musical structure must not be discounted when interpreting the results of such studies, even in controlled conditions that withhold contextual information such as lyrics and instrumentation. This tendency to attend at superordinate levels of abstraction and to err within contextual category is consistent with the principle of cognitive economy as outlined by Rosch (1978); in cases where there is a conflict between the desire to attain a fine discrimination between members of a category, and the cognitive resources available—participants had to produce their responses within six seconds of the end of presentation in each trial—the naming of a more abstract (contextual) category, or the choice of a readily accessible exemplar within that category, is a working compromise.

On methodological grounds, the apparent lack of predictive value of responses indicating a "feeling of knowing" in our altered conditions, regarding the subsequent ability to provide any identifying information on the unaltered melodies, suggests that listeners' experiential feelings of familiarity when attending to unidentified melodies may be suspect, and perhaps illusory, in certain conditions. It must be noted that the distorted stimuli are very different from melodies typically encountered in

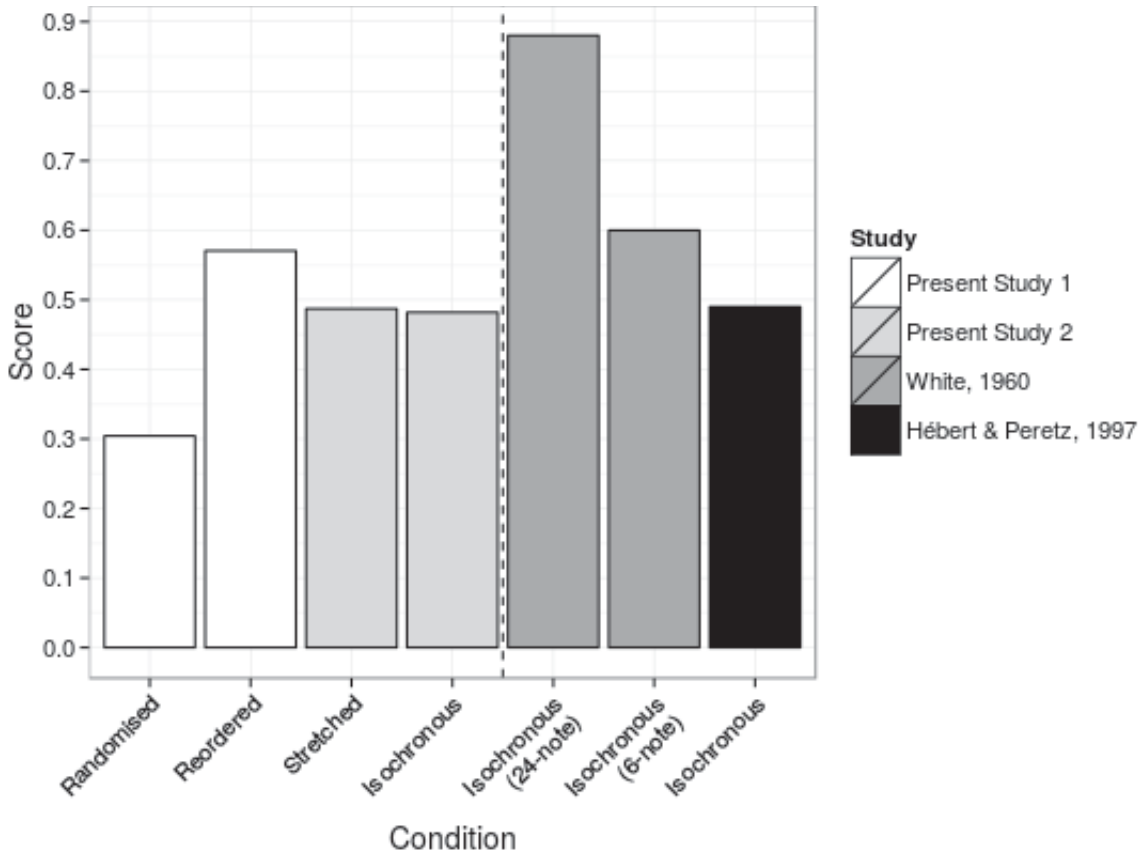


Figure 3–3: Melody identification performance in different rhythmically distorted conditions. N.b.: Bars to the right of the dotted line illustrate results of previous studies in the literature, using slightly different experimental procedures and different stimulus sets, and sampling participants from different populations, than is the case in the present studies.

everyday listening. However, experiments in music perception and cognition involving real, pre-existing musical stimuli commonly attempt to control for confounds of participant familiarity by asking participants to report or rate their familiarity with each musical piece. In light of our results, it may be recommendable to supplement this self-reporting with an identification task, to see whether participants are able to provide any identifying information for pieces deemed "highly familiar." Our findings suggest that ratings of the "feeling of knowing" without such additional validation may be insufficient in controlling for confounds of familiarity. We note the slight parallel to Newton (1990): As her tappers are unable to accurately predict the melody identification performance of other listeners, so our participants do not significantly predict their own level of identification success when attending to an impoverished musical signal.

Finally, the weak overall performance of our three algorithmic measures of rhythmic distance in terms of the correlation with participants' identification performance suggests that they are insufficient as algorithmic measures of experienced melodic identity. Such a measure must be operationalized in order to provide an approach toward quantifying topical relevance for tasks such as query-by-humming, query-by-performance, and cover song detection in the field of Music Information Retrieval (Weigl & Guastavino, 2013; Chapter 2 of this dissertation).

Our results strongly suggest that rhythmic cues, although insufficient in isolation, cannot be disregarded in determining experiential measures of melodic identity. Combinations of distance measures beyond those focusing on rhythm alone have

shown improved correspondence with listener's judgments in melodic similarity rating tasks (Müllensiefen & Frieler, 2004). Future directions might fruitfully include investigations of such combined measures in the context of a melody identification task along the lines presented here.

Furthermore, our results demonstrate significant contributions of connotative, contextual cues to experienced melodic identity. Algorithmic measures may stand to benefit by incorporating such cues. Systems to automatically generate contextual tags, based on data mining of user-contributed annotations on the web, have been proposed in the MIR literature (Bischoff, Firan, Nejdl, & Paiu, 2009). Metadata schemes such as those of the Music Encoding Initiative (Hankinson, Roland, & Fujinaga, 2011), and the Digital Music Objects currently being explored as part of the Fusing Audio and Semantic Technologies for Intelligent Music Production and Consumption (FAST-IMPACt) project (De Roure, Klyne, Page, Pybus, & Weigl, 2015), explicitly provide elements encoding context alongside those encoding musical facets. Future research in this direction is necessary in order to fully address the challenge of measuring experiential relevance in the field of Music Information Retrieval.

Whereas the melody identification task investigated in this chapter relates to querying scenarios in which a known musical item is being sought, we now turn toward an investigation of the role of rhythmic information in a different information seeking scenario; an alternative case that involves the user searching for music to suit the needs of a particular usage context, where situational requirements dictate the desired nature of specific aspects of the music. Correspondingly, the quality of a match is not judged by "topicality" (as in this chapter; the melodic query matches

the information object retrieved), but rather by situational relevance (a measure of the relationship between aspects of the information object, and the particular information need situation at play).

Concretely, Chapter 4 will explore the common musical use case of finding music to move to, by undertaking a series of experiments investigating beat salience, a measure of the perceptual prominence of the beat. In these experiments, we will assess the inter-rater reliability of beat salience ratings, as well as the effect of varying levels of beat salience on the beat induction and sensorimotor synchronization phases forming the cognitive process of moving to music, and the predictive validity of salience ratings on task performance in these contexts. Finally, we will consider possible implementation strategies for an algorithm to operationalize beat salience as a situational relevance criterion for MIR systems.

CHAPTER 4
A Convergent-Methods Investigation of Beat Salience in Sensorimotor Synchronization

Sensorimotor synchronization—the temporal coordination of motor action with a regular, periodic external signal—is an ability shared almost universally from an early age (Drewing, Aschersleben, & Li, 2006), across different levels of musical expertise (Chen, Penhune, & Zatorre, 2008). Indeed, recent evidence suggests this ability is not unique to our species (Fitch, 2013; Patel, 2014; Patel, Iversen, Bregman, & Schulz, 2009). Yet despite its ubiquity in the human population, several authors have reported considerable individual differences in rhythmic ability (Grahn & Schuit, 2012), with some special populations exhibiting rhythmic impairment or "beat deafness" (Foxton, Nandy, & Griffiths, 2006; Phillips-Silver et al., 2011).

Differences in sensorimotor synchronization are of particular interest in the context of music research, since music listening frequently involves the coordination of a motor rhythm with an external rhythm (Repp, 2005). For this reason, the effect of tempo on performance in beat perception and synchronization tasks has received considerable research attention. According to Parncutt (1994), *pulse sensation*, the evocation of a sense of beat, swing, or rhythm in the mind of the listener, occurs when the inter-onset interval (IOI) conforms to a range between about 200 and 1,800 ms. Within this range, pulse sensation is greatest in the so-called dominance region

of about 400-900 ms, reaching a point of maximum pulse salience at around 600 ms (Fraisse, 1982; Parncutt, 1994). Outside of this range, pulse sensations cease.

Researchers have operationalized the perceptual prominence of the beat in various ways. Tzanetakis, Essl, and Cook (2002) refer to *beat strength* as the rhythmic characteristic(s) that allow discrimination between two musical exemplars sharing the same tempo, while Lartillot, Eerola, Toiviainen, and Fornari (2008) define *pulse clarity* as a measure of how easily listeners can perceive the underlying rhythmic or metrical pulsation at a given point in the music. Finally, Toiviainen and Snyder (2003) define *resonance value* as the accentuation strength of a particular tapping mode, i.e., the period and phase that a listener would be likely to adopt when tapping along to the music.

Previous studies have evaluated computational models of the prominence of the beat using subjective rating tasks. Tzanetakis, Essl, and Cook (2002) conducted an experiment where 32 participants with varied levels of musical expertise rated 50 musical excerpts of 15 seconds each, representing a variety of musical styles, on a Likert scale encompassing five levels of beat strength (from *weak* to *strong*). Lartillot et al. (2008) conducted a similar study in which 25 musically trained participants rated the pulse clarity of 100 five-second excerpts from movie soundtrack on a 9-point Likert scale (from *unclear* to *clear*).

Rating tasks possess several advantages in this research context. They are easy to administer, and do not require musical expertise to complete. Both of the above studies demonstrated high levels of inter-rater agreement, suggesting that such ratings are reliable across individuals. However, the approach also holds a number

of limitations: such ratings are retrospective rather than momentary, representing an overall judgement of the stimulus without capturing real-time variation, and as a consequence, suffer from potential confounds relating to attention and memory (Steffens & Guastavino, 2015). What is more, motor coordination is often not considered, though working definitions of beat strength, pulse clarity, etc. often cite the ability to synchronize to the external stimulus as an important factor.

Toiviainen and Snyder (2003) take an alternative approach by employing a tapping task to focus explicitly on the motor processes of sensorimotor synchronization. There is a long tradition of sensorimotor synchronization research employing such a tapping paradigm (Repp, 2005; Repp & Su, 2013). In such experiments, participants tap along to isochronous (metronome-like) pulse patterns, simple synthesized rhythms, or musical recordings. Relevant dependent variables include tap consistency, which represents the degree of regularity of a participant's taps; tap asynchrony, which refers to the temporal offsets between participants' tap times and the "true" beat positions within the stimulus; and tap consensus (Toiviainen & Snyder, 2003), which represents the consistency of the tap times of different individuals for the same stimulus. Participants may be instructed to tap at particular auditory events in the stimulus (e.g., with every metronome click), or to tap along at a pace that feels natural to them. Tasks employing this latter approach are termed pulse-finding tasks, and they typically result in inter-participant variation resulting from the metrical level at which participants decide to tap. Meter in this context refers to nested layers of approximately equally spaced beats or pulses (Krebs, 1999). For a passage notated in common time and played at a moderate tempo (i.e., where each

measure is divided into four equally spaced beats), for example, participants sometimes tap at the level of the quarter note (i.e., on beats 1, 2, 3, and 4 within the bar), at the level of the half note (i.e., on beats 1 and 3), or at the level of the whole note (i.e., on beat 1). In a pulse-finding task, participant tap times therefore implicitly reveal characteristics of the perceived meter, where tapping tends to adhere to a particular level of the metrical hierarchy.

Tapping experiments have typically employed auditory stimuli presenting rhythmically simple or isochronous tone sequences. Such synthetic, lab-generated stimuli offer a great amount of control over the numerous attributes characterizing most musical genres, including timbre, pitch, intonation, and accentuation; however, their ecological validity is limited in a musical context. What is more, pulse-finding experiments employing genuine musical stimuli have generally been confined to a narrowly defined genre and style (Snyder & Krumhansl, 2001), or to different excerpts from the same musical piece (Toiviainen & Snyder, 2003). This more naturalistic approach offers greater ecological validity while maintaining a fair degree of control over confounding facets. However, the range of applicability toward a more varied musical and rhythmic context remains limited.

The act of sensorimotor synchronization is a two-stage process. Sensorimotor synchronization begins with a perceptual process of beat induction in the first stage, in which the listener determines the period and phase of the isochronous pulse at one or more levels of the metrical hierarchy, with the various motor processes involved in synchronization constituting the second stage. One disadvantage of the tapping paradigm is that these perceptual and motor processes are difficult to disentangle

(Toiviainen & Snyder, 2003). The beat alignment test (BAT) proposed by Iversen and Patel (2008) offers an experimental framework to address this issue by combining the tapping approach discussed above with a psychoacoustic task measuring beat perception.

The BAT was conceived as a test battery to "quantify the normal range of beat-based processing abilities in the general population" (p. 465), to assess the degree of association between beat perception and sensorimotor synchronization task performance, and to identify "beat deaf" participants, i.e., individuals that demonstrate significantly impaired abilities of beat perception and synchronization but do not exhibit congenital amusia or "pitch deafness" (see also Phillips-Silver et al., 2011).

The BAT consists of three stages of testing: 1) a synchronization task to collect baseline data, in which participants are first instructed to tap consistently at a tempo of their choosing, without an accompanying stimulus, and then to tap along to metronomic sequences at different tempi; 2) a pulse-finding task, in which participants tap along to each of twelve musical stimuli selected from the rock, jazz, and pop instrumental genres; and 3) a beat perception task, in which participants indicate whether isochronous trains of pulses (beeps) superimposed over each musical excerpt align with one or more levels of the metrical hierarchy. In the perception task, three versions of each excerpt are presented, with pulses either correctly placed "on the beat", or exhibiting one of two types of error: pulse trains that are too fast or too slow ("tempo error" condition), or pulse trains that are out of phase (either too early or too late) with the beat of the music ("phase error" condition). Participants

respond by indicating as quickly as possible whether the pulses are on or off the beat, and accuracy and response time are recorded.

Iversen and Patel (2008) reported on a pilot application of the BAT to 30 participants ranging in musical expertise from professional to self-declared amusic (p. 466). Their results demonstrated significant differences in synchronization performance between individuals (matching tapping tempo against musical tempo), as might be expected given the range of expertise. None of their participants were found to be "beat deaf", with tempo sensitivity remaining clearly evident even in the worst individual performance (an outlier demonstrating relatively poor tempo tracking). In the beat perception task, participants demonstrated a tendency to identify pulse trains as being "on-the-beat", even when tempo or phase errors were in fact present. Although on-beat pulse trains were identified correctly in most cases (mean percent correct = 90%), performance was significantly impaired in both error conditions, with particular poor results in the phase error condition (mean percent correct = 60%, IQR = 30%, range = 18 – 98%; see figure 3, p. 467). Finally, their results demonstrated a moderate correlation between synchronization and perception task performance, when the results of the participant with the worst synchronization performance were removed.

The present study concerns beat perception and sensorimotor synchronization to musical stimuli with beats exhibiting a range of degrees of perceptual salience. Similarly to Tzanetakis, Essl and Cook (2002), we are interested in rhythmic characteristics beyond tempo. However, whereas their study conceives of beat strength as supporting discrimination between different types of beat, we wish to measure the

effect of such characteristics on the process of beat perception and sensorimotor synchronization. Our concerns are also well aligned with those of Lartillot et al.'s study of pulse clarity (2008), in that we wish to investigate the clarity of the rhythmic pulsation underlying the music as it is perceived by the listener. However, rather than focusing on momentary impressions for short musical stimuli, we employ somewhat longer excerpts of 17 seconds, encompassing an average of 6.33 complete measures of music (SD=1.52), in order to involve longer-term rhythm and meter perception to provide greater ecological validity in the context of perceiving and moving to music. Concretely, our study investigates the sensation of *beat salience*, which we define as the quality of the music that affords synchronized rhythmic movement; a measure of the perceptual prominence of the beat in the context of moving to music.

Our motivations for pursuing this study are rooted in the disciplines of psychology, music perception and cognition (MPC), and music information retrieval (MIR). From a psychological perspective, we apply a convergent measures approach in order to study the phenomenon of sensorimotor synchronization by obtaining psychometric measures of the constituent processes of perception and motor synchronization, and experiential measures in the form of self-reported ratings by the listener, across a broad range of task difficulties. From the perspective of MPC, we contribute to a tradition of research in rhythm perception while filling certain gaps in previous work – limitations in the ecological diversity of musical stimuli, and a focus on participants lacking substantive musical training. Non-musicians are of particular interest given the near-universality of the human ability to synchronize movement to a musical beat. Finally, rhythmic synchronization of movement is a common musical "use

case", as in dance, work, or exercise. Musical qualities that afford such movement are thus of particular interest from an MIR perspective, where beat salience may serve as a useful relevance criterion or query mechanism in MIR system design.

To address these goals, the present study employs a convergent approach featuring three experimental tasks. Our participants first complete either a beat perception task (Experiment 1) or a sensorimotor synchronization task (Experiment 2), based on the corresponding tasks of the Beat Alignment Test (Iversen & Patel, 2008). Both experiments employ musical stimuli selected in a preliminary study as exemplars of a range of beat salience levels (*low, medium,* or *high*). After completing their first task, all participants completed a rating task (Experiment 3) based on the approaches of Tzanetakis, Essl, and Cook (2002) and Lartillot et al (2008).

4.1 Preliminary study

We conducted a preliminary study in order to determine a normative set of musical exemplars of low, medium, and high levels of beat salience, to serve as stimuli in our perception, synchronization, and rating tasks (Experiments 1-3).

4.1.1 Participants

Ten participants (four female) were recruited from the McGill University community. All participants described themselves as either amateur (9) or professional musicians (1), averaging 13.1 years of study on a musical instrument. All participants also described themselves as music lovers, listening to an average of six hours of music each week. Participants also reported listening to a variety of musical genres (e.g., classical, pop, rock, techno, hip hop, folk, etc.). All participants reported normal hearing. A standard audiogram was administered before the experiment to

confirm that hearing thresholds were below 20 dB HL (ISO 398-8, 2004). None of the participants indicated they had absolute pitch. All participants gave informed consent. The study was certified for ethical compliance by McGill University's Research Ethics Board II.

4.1.2 Materials

The stimuli consisted of 54 short excerpts (17s) selected from the popular and electronic music repertoire, and a further six excerpts selected to serve as example stimuli in practice trials before data collection began. To ensure familiarity with the excerpts would not affect participant ratings of beat salience, we attempted to select relatively unknown excerpts. Care was also taken to select stimuli in which the perceived beat salience would not vary over the course of the excerpt. Each excerpt was pre-classified into one of three categories by the authors—low, medium, and high salience, 18 excerpts in each category—according to our perception of the clarity and strength of the beat. A one second fade-in and fade-out was also applied to each stimulus to de-emphasize stimulus boundaries. To ensure differences in perceived loudness between excerpts would not affect beat salience ratings, the stimuli were normalized to -.3 dB.

4.1.3 Apparatus

The participants were seated in a double-walled IAC sound-isolation booth (IAC Acoustics, Bronx, NY). The stimuli were reproduced on a Macintosh G5 PowerPC (Apple Computer, Cupertino, CA), output as S/PDIF using an M-Audio Audiophile 192 sound card (Avid, Irwindale, CA), converted to analog using a Grace Design m904 monitor system (Grace Design, Boulder, CO), and presented stereophonically over a

pair of Dynaudio BM6A monitors (Dynaudio, Skanderborg, DK). The stimuli were presented at a comfortable listening level that was kept constant for all participants. The experimental program, subject interface, and data collection were programmed using the Max/MSP program from Cycling 74' (San Francisco, CA), and controlled by the PsiExp software environment (Smith, 1995).

4.1.4 Design and procedure

Upon arriving, each participant was asked to fill out an informed consent form, and then directed into the testing booth to begin the experiment. After listening to each excerpt, participants were instructed to rate the salience of the beat on a 5-point Likert scale labeled from *very low* to *very high*. Beat salience was defined to the participants as "the perceptual strength or prominence of a beat. A value of 5 indicates that you could easily tap to the beat, while a value of 1 indicates that you simply could not tap to the beat." Participants were encouraged to use the full range of the scale over the course of the experiment, and to listen to each excerpt as many times as they wished. Participants were explicitly asked to tap along to the music in order to inform their response.

In addition to the beat salience rating, participants rated their familiarity with the excerpt on 7-point analogical-categorical scales (Weber, 1991). This consisted of an analogue scale subdivided into seven discrete categories labeled from 1 to 7, where a rating of 7 indicated that they had certainly heard the excerpt and knew the song, and 1 indicated that they had never heard the excerpt before.

To familiarize the participants both with the range of stimuli and with the experimental task, the experiment began with a practice session of six additional

excerpts, pre-categorized as exhibiting low, medium, and high beat salience (two additional excerpts per beat salience category). The pre-study was composed of two blocks of 27 trials each. At the end of the first block, participants could take a short break, leaving the testing booth if they wished. After completing the ratings task, participants filled out a questionnaire addressing their music background.

4.1.5 Outcomes

Agreement between participants was fairly high, with a median of eight participants assigning the same rating on stimuli we had pre-classified as exemplifying a high level of beat salience, and a median of five participants on medium and low beat salience level stimuli.

To determine a minimum threshold of inter-case agreement, a chi-square test was calculated to compare the beat salience judgement that received the highest number of responses (e.g., very low) to the sum of the responses for the other four possible judgements. Out of 180 cases within each beat salience category (10 participants x 18 excerpts per category), a minimum of 46 identical responses (25.6%) was necessary to achieve significance, $\chi^2(1) = 4.27, p < .04$. The excerpts pre-classified as high beat salience received "high" responses in 23%, and "very high" responses in over 65% of all cases. Excerpts pre-classified as medium beat salience received "medium" responses in just under 40% of all cases, while excerpts pre-classified as low beat salience received "low" responses in 40%, and "very low" in 25% of all cases.

On average participants rated the excerpts as somewhat familiar ($M = 3.03, SD = 3.8$). However, the mean familiarity ratings displayed a fairly restricted range, falling between 2–4 on the 7-point scale. Unfortunately, the median beat salience ratings

for each excerpt were correlated with ratings of familiarity, $rs(54) = .7, p < .001$, indicating that the participants' familiarity with the excerpts may have affected their beat salience ratings.

The outcomes of this preliminary study were used to obtain a reduced subset of 24 highly representative exemplars of each beat salience level (8 musical excerpts per category). To eliminate unwanted effects of familiarity, care was taken to select excerpts that received low mean familiarity ratings (1 SD below the mean). K-means clustering on our participants' beat salience ratings was then performed in order to identify the 8 most representative excerpts for each category. The 24 resulting excerpts were used in the remaining experiments reported in this chapter, and are listed in Appendix F.

4.2 Experiment 1: Beat induction

We conducted an experiment in order to investigate the effect of beat salience on beat induction, the perceptual phase of sensorimotor synchronization wherein the temporal location of the external beat is established. This experiment employed a response time task modeled on the perceptual judgement task of the Beat Alignment Test (Iversen & Patel, 2008), with the addition of *beat salience level* as an independent variable.

An assumption underlying our experimental design is that completion of this task requires a series of perceptual steps: beat induction, that is, establishing an internal representation of the regular isochronous pattern corresponding to the metric grid of the music; locating the position of the cowbell pulses along this grid; and finally, assessing whether these pulse positions intersect with valid positions of the

beat of the music. We expect the difficulty of the first step to vary according to the clarity with which the beat of the music is perceived. Accordingly, we hypothesize a linear relationship between beat salience level and our outcome measures, where high beat salience levels are associated with fast response times and a high proportion of correct responses, and low beat salience levels are associated with slow response times and a low proportion of correct responses.

4.2.1 Participants

Thirty non-musicians (17 female) from the Montreal community were recruited through the McGill University classified ads calling for non-musicians with no known hearing impairments to participate in a music perception experiment. Interested volunteers were screened in a pre-participation questionnaire to ensure they possessed no more than one year of formal training on a musical instrument. None of the participants had taken part in the preliminary study. The average age of the participants was 25 years (SD = 6). Twenty seven participants self-identified as "non-musicians"; three self-identified as "amateur musicians". All but one participant indicated that they enjoyed listening to popular and electronic music, and participants reported listening to a mean of 2.5 hours of music each day (SD=2.5). All participants exhibited unimpaired binaural hearing at ranges from 250 – 8,000 Hz as assessed by a standardised audiometric test at the beginning of each experimental session. All participants gave informed consent. The study was certified by the McGill University Review Ethics Board.

4.2.2 Materials

Twenty four musical excerpts selected in the preliminary study were presented to participants. This stimulus set comprised eight exemplars each of the low, medium, and high beat salience categories. Each stimulus was presented alongside an overlaid isochronous (metronome-like) pulse train presenting an impulsive cowbell sound (200 ms). The cowbell sample was selected from a range of sounds including other percussive audio samples, pure tones, and noise bursts, as its timbre was determined to be the most easily perceptually separable from the musical stimuli.

4.2.3 Apparatus

The sound isolation booth, along with its computer and audio equipment set-up, was described in the preliminary study. For the purposes of Experiment 1, a high-accuracy response time measuring device (Li, Liang, Kleiner, & Lu, 2010) was connected to the USB port of the computer to accept participants' responses. The outer two of the device's four response buttons were masked with tape, leaving a red and a green response button visible and available to the participant. A marker was placed just below and in between these two buttons to indicate the resting position for the participant's index finger during the experiment. The experimental interface was presented using a MATLAB script and the Psychtoolbox-3 module (Kleiner et al., 2007).

4.2.4 Design

Each experimental trial presented one of the 24 stimuli, along with an overlaid pulse train consisting of an isochronous cowbell beat. Each trial first presented the musical excerpt without the pulse train overlay for five seconds. The pulse train

overlay then began during a measure within the following three seconds, according to the present trial's experimental condition (see below). The musical excerpt and pulse train overlay then continued playing until the participant issued a response, or until the excerpt's duration of 17 seconds was reached. Participants were tasked with responding as quickly and as accurately as possible as to whether or not the pulse train was synchronized to the beat of the music.

The placement of the pulse train corresponded to the experimental condition according to two factor manipulations: meter, in which the first sounding of the pulse train occurred on either a metrically strong or metrically weak position, corresponding to the first or last beat of the measure; and phase, in which the pulse train was either aligned perfectly with the beat of the music, or misaligned with a phase error of $\pm$75 ms (early or late). This duration was selected as it represents the duration of a 32nd note at the ideal tapping tempo of 100bpm as determined by Parncutt (1994).

The study therefore employed a 3 (*beat salience level*) x 3 (*phase condition*) x 2 (*metric condition*) design, where beat salience level corresponded to the classification of the individual musical excerpts in the preliminary study. There were 8 musical excerpts for each beat salience level, and each excerpt was presented to each participant once in every combination of the phase and metric conditions, resulting in a total of 144 trials in the experimental session. Presentation order of excerpts and conditions was randomized for each participant. Participants' response times (the time between the first sounding of the pulse train and the participant's response) and accuracy were recorded.

4.2.5 Procedure

After signing the informed consent form, each participant was led to the testing booth, where an audiogram was performed. Upon passing the audiogram test, participants received instructions regarding the experimental task and the requirement to respond as quickly as possible as soon as they were confident of the correct response was emphasized to each participant.

A practice block consisting of eight trials preceded the main experiment, presenting stimuli to exemplify the different levels of beat salience, using excerpts not included in the main session. During this block, the experimenter was present in the testing booth to ensure that the participant understood the interface and the task. After completion of the practice block, the experimenter exited the testing booth, and the participant was left to complete all 144 trials of the main experiment. Each successive trial's presentation started automatically 500ms after the participant issued a response for the previous trial. In the case that any stimulus played for 17s without receiving a response, the excerpt would fade to silence, and the next trial would only commence after the participant issued a response; participants were instructed to simply press either response button in this case, and any such delayed responses were withheld from analysis.

After completion of the main experimental task, participants completed Experiment 3 (the beat salience rating task), described below. Finally, participants filled out a questionnaire on their musical background. All participants received $10 CAD upon completion of the experimental session as compensation for their time.

4.2.6 Analysis

Response times. As differences in response times are to be expected between correct and incorrect responses (Ratcliff & Rouder, 1998) we constrained our analysis to take into account only correct responses. This necessarily results in an unbalanced dataset, as the proportion of correct responses will vary according to differences in expected task difficulty among the experimental conditions. Further, while it was desirable to treat our musical excerpts as random effects, we expected correlations among the response times for excerpts exhibiting the same level of beat salience. As such, our analysis applied a linear mixed effects modelling approach that is able to address both unbalanced data sets and crossed random effects (Pinheiro & Bates, 2006). The analysis was performed using R (R Core Team, 2015) and the lme4 module (Bates, Maechler, Bolker, & Walker, 2014). Response times were log-transformed in order to address the positive skew inherent in raw response time data.

Following Barr, Levy, Scheepers, and Tily (2013), we first specified a maximal random-effects structure, with sum-coded fixed effects for beat salience level, phase condition, metric condition, and their interactions, and random intercepts with by-participant slopes for all fixed effects, and by-item (musical excerpt) slopes for phase condition and metric condition[1] . Linear mixed effects models (LMEM) are typically fit using an iterative algorithm that estimates parameter values to maximize the likelihood of obtaining the observed data, given the structure of the model (Barr

[1] By-item slopes for beat salience are not justified by our design, as each musical excerpt belongs to only one beat salience level.

et al., 2013, p. 261). This algorithm is not guaranteed to arrive at a solution within a tractable number of iterations, in which case the estimation is said to be "non-convergent". Such convergence failures are more commonly encountered in structurally complex models (p. 261), and in our case, the maximal model failed to converge.

To identify principled means of simplifying the model, we fit a reduced, random-intercept-only linear mixed effects model (LMEM) with fixed effects for beat salience level, phase condition, metric condition, and their interactions, and crossed random intercepts by participant and item. No random slopes were specified. Such random-intercepts-only models are structurally simpler than models specifying slopes, and thus more likely to converge; however, they are prone to producing type 1 errors, as the unspecified random slope variation may account for differences in condition means, resulting in spurious "treatment effects" that are in fact merely statistical artifacts (p. 261). We opted to fit such a model in order to identify any fixed effects that remain insignificant even with this inflated risk of false positives, in order to justify pruning these from a more fully specified model encoding random intercepts and slopes. Significance was determined using the pamer.fnc function of the LMERConvenienceFunctions R module, which computes "upper- and lower-bound p-values for the analysis of variance (or deviance) as well as the amount of deviance explained (%) for each fixed-effect" of an LMEM, given the range of possible degrees of freedom (Tremblay & Ransijn, 2013). The analysis of this intercepts-only model revealed a significant main effect of beat salience level, and a significant

interaction of beat salience level and phase condition. No significant effects were detected for the metric condition, nor for any of its interactions.

We thus arrived at our final model, specifying fixed effects for beat salience level, phase condition, and their interactions, and random intercepts with slopes by-participant for all fixed effects present in the model, and by-item (musical stimulus) for phase condition. This model's structure does not account for the metric condition, but is otherwise identical to our initial, maximally specified model. After convergence, assumptions of homoscedasticity and normality were checked by visual inspection of residual plots; no obvious deviations were detected. Hypothesis testing was conducted using the pamer.fnc module to determine significance (Tremblay & Ransijn, 2013), as above.

In order to control for potential confounds that may be introduced by the range of tempi represented in the stimulus set (see Appendix F), we repeated the above analysis investigating an alternative measure of the response time: the number of inter-onset intervals of the overlaid cowbell pulses before a response was obtained in each trial. Rather than measuring the (tempo-dependent) time period before a response, we thus obtained a measure of the number of (tempo-independent) indications of the overlaid pulse's position relative to the beat of the musical stimulus before a response was obtained. Similarly to response times, the distribution of observations of this measure is positively skewed; we thus log-transformed our data before analysis. As in the response time analysis above, a maximal model failed to converge; however, an intercept-only model indicated no significant main effects of the metric condition or its interactions. The outcomes of a model replicating the

final model of the response time analysis—fixed effects for beat salience level, phase condition, and their interactions, and random intercepts with slopes by-participant for all these fixed effects, and by-item (musical stimulus) for phase condition—were similar to the results obtained in the response time analysis.

Accuracy. A logistic linear mixed effects modeling approach was undertaken using the lme4 R module (Bates et al., 2014) following a similar approach to the one described for the response time analysis above. Following Barr et al. (2013), we specified a maximal model with fixed effects for beat salience level, phase condition, metric condition, and their interactions, and random intercepts with by-participant slopes for all fixed effects, and by-item (musical stimulus) slopes for phase condition and metric condition. This maximal model achieved convergence.

4.2.7 Results

Response times. Our raw data set included 4,320 observations, corresponding to 30 participants' responses to 24 musical excerpts (eight per beat salience level), each presented in three phase conditions and two metric conditions. We removed two observations encoding responses that were given before the sounding of the first cowbell pulse, and 292 observations given after 17 seconds (i.e., after presentation of the stimulus had ended). The resulting data set included 4,026 observations, with a minimum response time of 455 ms after the sounding of the first cowbell pulse.

As timing differences are to be expected between correct and incorrect responses (Ratcliff & Rouder, 1998), this set of observations was further divided into correct (N=2,198) and incorrect (N=1,828) cases in the analysis of response times.

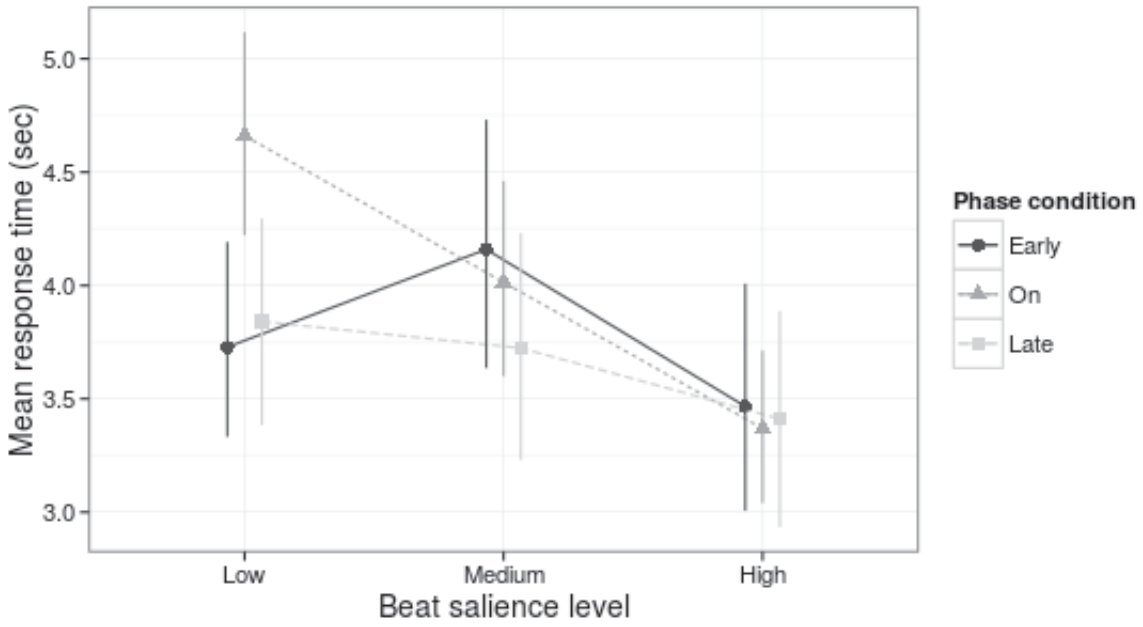


Figure 4–1: Cross-participant mean response times for correct responses across beat salience levels and phase conditions. Response times are measured from the sounding of the first cowbell in the pulse train overlay. Point ranges indicate bootstrapped 95% confidence intervals (1,000 resamplings).

The analysis of our random-intercept-only model of log-transformed response times (RT) for correct responses, fit to identify a principled means of reducing the maximally specified model after non-convergence, revealed a significant main effect of beat salience level, $F = 3.18, p < .05$, and a significant interaction of beat salience level and phase, $F = 8.95, p < .001$; with a range of DF of 2,126 – 2,180, and p-values, adjusted to control for false discovery rate (FDR) for multiple comparisons (Benjamini & Hochberg, 1995), differing by less than .0001 over this range.

As no significant effects of metric condition or its interactions was detected, we arrived at our final model, specifying fixed effects for beat salience level, phase

condition, and their interactions, and random intercepts with slopes by-participant for all fixed effects present in the model, and by-item (musical stimulus) for phase condition. Analysis of this model revealed a significant main effect of beat salience level, $F = 3.31, p < .05$, and a significant interaction of beat salience level and phase, $F = 3.2, p < .05$, with a range of DF of 1,847 – 2,189, and FDR-adjusted p values differing by less than .001 over this range. Figure 4–1 illustrates the mean untransformed RT for correct responses across the different beat salience levels and phase conditions.

Post-hoc testing of the interaction between beat salience level and phase condition was conducted by employing least-squares means using the R lsmeans package (Lenth, 2014), with pairwise contrasts of beat salience level grouped by phase condition (Figure 4–2). This analysis revealed significant differences in log(RT) of responses to low vs. high beat salience stimuli when phase condition was "on": low - high beat salience $log(RT) = 0.38, t(31.16) = 3.65, p < .01$ (Tukey-adjusted for multiple comparisons). We also detected a trend short of significance in the differences in log(RT) of responses to medium vs. high beat salience stimuli when phase condition was "on": medium - high beat salience $log(RT) = 0.2, t(21.96) = 2.22, p < .1$. No significant differences between beat salience levels were detected in those conditions exhibiting phase error ("early" and "late").

Pulse train inter-onset intervals before response. We repeated our analysis of the correct responses (N=2,198) using an alternative, tempo-independent measure of the speed of response – the number of inter-onset intervals of the overlaid

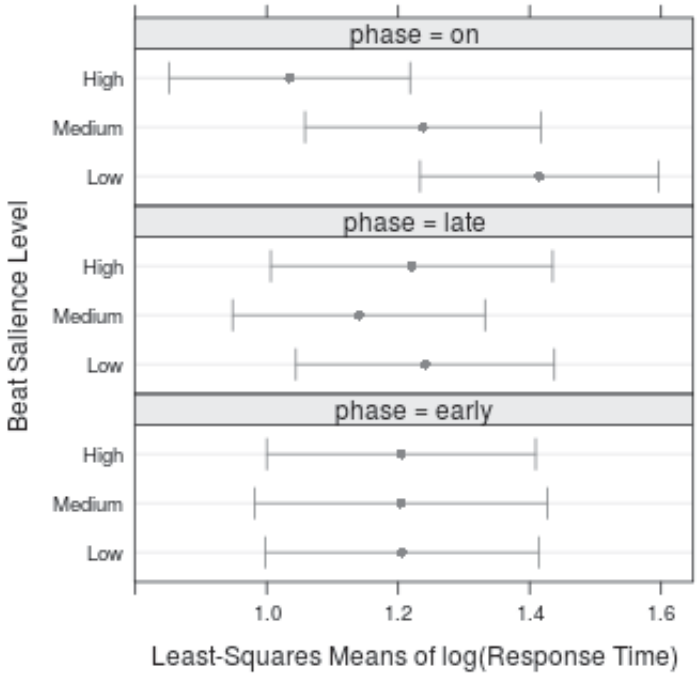


Figure 4–2: Least-squares means and 95% confidence intervals of log(RT) according to beat salience level, grouped by phase condition. Response time measures to low and high beat salience stimuli differ significantly in the phase "on" condition.

cowbell pulses before a response was obtained in each trial. We fit an LMEM specifying fixed effects for beat salience level, phase condition, and their interactions, and random intercepts with slopes by-participant for all fixed effects present in the model, and by-item (musical stimulus) for phase condition, replicating the final model employed in the RT analysis above. This analysis revealed a significant interaction of beat salience level and phase condition, $F = 3.48, p < .001$; with a range of DF of 1,847 – 2,189, and FDR-adjusted p values differing by less than .001 over this range.

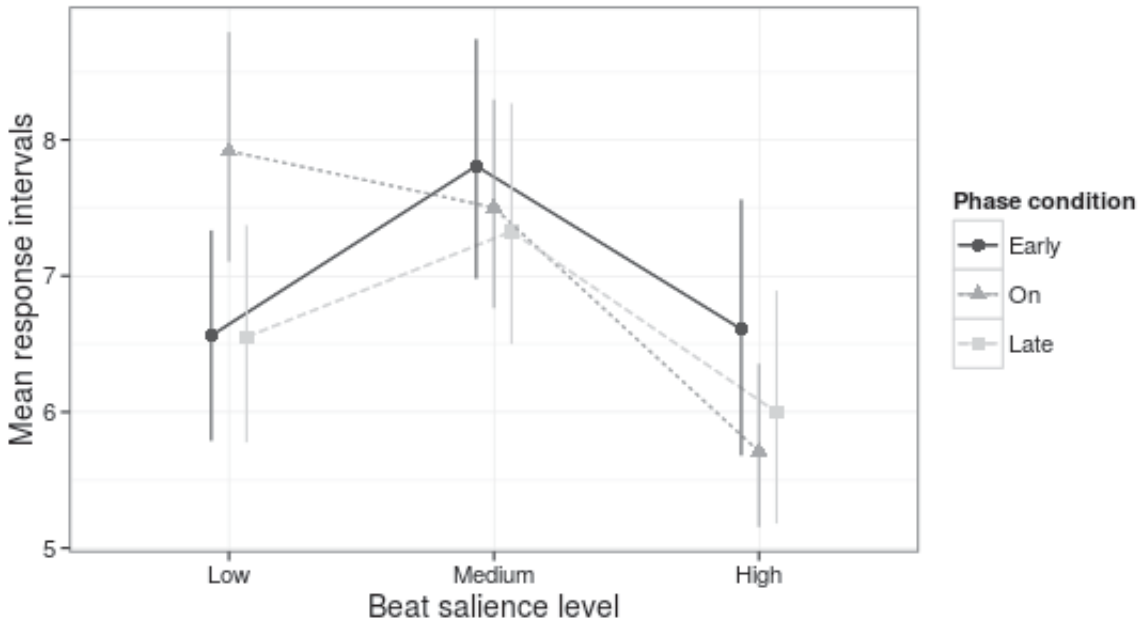


Figure 4–3: Mean number of pulse train inter-onset intervals between first sounding of cowbell beat and participant response. This alternative indicator provides a measure of response speed that is independent of stimulus tempo. Point ranges indicate bootstrapped 95% confidence intervals (1,000 resamplings).

Post-hoc testing conducted in the same manner discussed above for response times was repeated on our inter-onset interval measure, resulting in significant divergence from the grand mean only in the low and high beat salience levels of the phase "on" condition (low beat salience: divergence from grand mean inter-onset intervals $= 0.14, t(28.56) = 1.94, p < .05$; high beat salience: divergence $= 0.23; t(25.24) = 3.3, p < .01$; FDR-adjusted p-values).

The outcomes of our analysis of this alternative measure thus largely replicate the results of our log(RT) analysis presented above, suggesting that our findings are resilient to confounds from the range of tempi in the employed stimuli. Figure 4–3

illustrates the mean number of inter-onset intervals for correct responses across the different beat salience levels and phase conditions.

Accuracy. The data set for the analysis of accuracy (proportion of correct responses) consisted of the undivided set of observations comprising correct and incorrect responses, after the removal of responses obtained before the sounding of the first cowbell pulse or after the presentation of the stimulus had ended (N=4026). Hypothesis testing on our maximally specified logistic LMEM, with fixed effects for beat salience level, phase condition, metric condition, and their interactions, and random intercepts with by-participant slopes for all fixed effects, and by-item (musical stimulus) slopes for phase condition and metric condition, was conducted using a type III Wald χ^2 test (Table 4–1). We determined a significant main effect of phase condition on accuracy, as well as a significant interaction of beat salience level and phase condition, reflecting the findings of the response time analysis. We further detected a significant interaction of phase and metric condition on accuracy, not present in the analysis of response times.

4.2.8 Discussion

Our results support the hypothesized linear relationship between beat salience level and our outcome measures in the phase "on" condition. When the pulse train aligned with the phase of the music, high salience levels corresponded to fast response times and a high proportion of correct responses, and low salience levels corresponded to slow response times and a low proportion of correct responses. The results for trials presenting misaligned pulse trains (i.e., the phase "early" and "late" conditions) do not support this hypothesis. Instead, we observed a significant interaction

Table 4–1: Proportion of correct responses: Type III Wald χ^2 test.

Fixed effect	DF	χ^2
Grand mean	1	5.19*
Beat salience level (BSL)	2	1.54
Phase condition (PC)	2	13.34**
Metric condition (MC)	1	2.13
BSL×PC interaction	4	22.49***
BSL×MC interaction	2	0.99
PC×MC interaction	2	6.84*
BSL×PC×MC interaction	4	0.64

*** $p < .0001$ ** $.001 < p < .01$ * $.01 < p < .05$

Note: N = 4,026. Model specifies random intercepts and slopes by-participant for beat salience level, phase condition, metric condition, and their interaction, and by-item for phase condition, metric condition, and their interaction.

Table 4–2: Proportion of correct responses: parameter estimates.

PC	BSL	Estimate	SE	z-ratio
Early	Low	0.44	0.29	1.51
	Medium	0.37	0.28	1.32
	High	−0.81	0.29	−2.79*
On	Low	−1.31	0.36	−3.65***
	Medium	−0.55	0.35	−1.57
	High	1.86	0.39	4.75***
Late	Low	1.15	0.43	2.66*
	Medium	0.46	0.42	1.1
	High	−1.62	0.46	−3.51**

*** $p < .0001$ ** $.001 < p < .01$ * $.01 < p < .05$ (FDR-adjusted p-values)

Note: Results averaged over the levels of the metric condition.

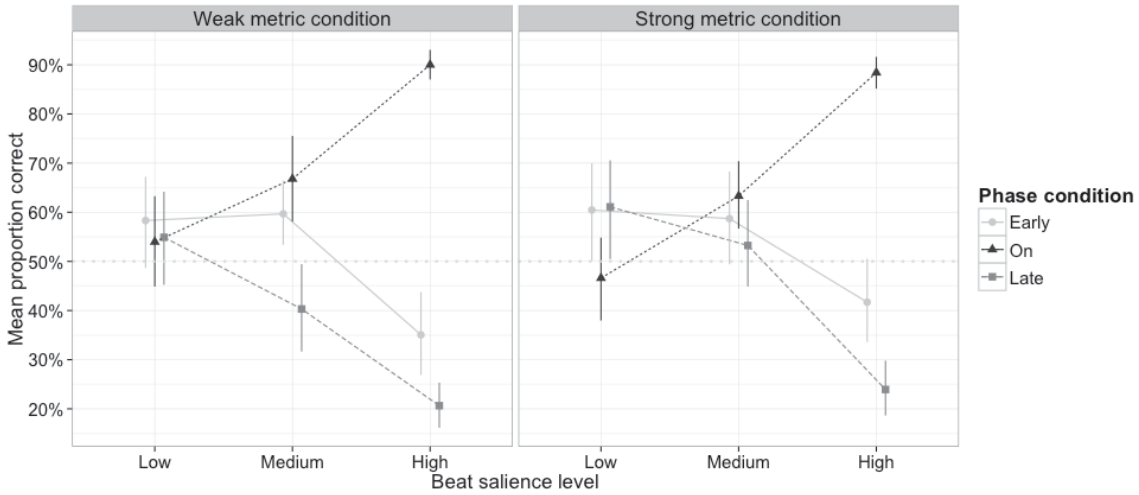


Figure 4–4: Cross-participant mean proportion of correct responses across beat salience levels, phase conditions, and metric conditions. Participants responded as to whether or not the pulse train was synchronized to the beat of the music. Dotted line indicates chance level (50%). Point ranges indicate bootstrapped 95% confidence intervals (1,000 resamplings).

between beat salience level and phase condition, with response times significantly diverging in the "low" and "high" beat salience levels of only the phase "on" condition. Our investigation of the alternative measure of response time controlling for tempo differences between stimuli, which consisted of the number of inter-onset intervals (inter-cowbell soundings) before a response was provided, reveals a similar pattern of results, though the difference between the "low" and "medium" beat salience levels of the phase "on" condition was less pronounced.

For the proportion of correct responses shown in Figure 4–4, the significant interaction between beat salience level and phase condition resulted in a counter-intuitive outcome where participants overwhelmingly responded that the pulse train

was synchronized with the beat of the music in trials presenting high salience stimuli, regardless of the actual presence or absence of phase error. This resulted in a proportion of correct responses that was significantly above the grand mean in the interaction of high beat salience and synchronized phase, as would be expected, but significantly below the grand mean in the interaction of high beat salience and "early" or "late" phase, contrary to expectation (Table 4–2, p. 140). This finding reflects a similar asymmetry in the results obtained in the beat perception task of Iversen and Patel (2008), where participants also more frequently judged off-beat pulses to be on the beat than they judged on-beat pulses to be off the beat. Iversen and Patel's study does not investigate beat salience; their stimuli would likely fall into the high beat salience level in our study's design.

The tendency toward perceiving marginally early or late beat onsets as falling on the beat in the high beat salience condition may be attributable to a captor effect, whereby the overlaid cowbell beats are perceptually grouped with the highly salient beat of the music. From this perspective, we may posit an integration window spanning a short time period around the beat onset time; pulse onsets occurring within this span are captured by the musical stimulus, and perceived as "on the beat". The absence of this effect for the medium and low beat salience stimuli could be accounted for by positing a relationship between the size of the integration window, and the distinctiveness of the metric grid established during beat induction, corresponding to the experienced level of beat salience. This hypothesis could be tested by experimenting with different phase offset sizes in the different salience conditions. We would expect the captor effect to be replicated with smaller phase

offsets in the medium and low salience conditions. Additionally, potential effects of the degree of participants' musical training on the location of the offset thresholds at the transition between the presence and absence of capture could be investigated, with the expectation that a higher degree of training would result in a narrower integration window, and thus heightened sensitivity to phase error.

Following on from our investigation of the effect of beat salience on beat induction, the primary perceptual phase of the sensorimotor synchronization process, we conducted a further experiment focusing on the secondary stage, motor coordination. This second experiment made use of a beat finding methodology commonly employed in tasks investigations of sensorimotor synchronization (Iversen & Patel, 2008; Repp, 2005; 2013; Toiviainen & Snyder, 2003) in order to investigate the effect of beat salience on participants' task performance in terms of consistency of motor coordination when moving to the musical beat.

4.3 Experiment 2: Sensorimotor synchronization

4.3.1 Participants

Thirty two non-musicians (18 female) from the Montreal community were recruited through McGill University classified ads. Participants were screened based on the same criteria as in Experiment 1. None of the selected participants had taken part in either Experiment 1 or in the preliminary study. The average age of the participants was 22 years (SD = 5). Twenty six participants self-identified as "non-musicians" and six self-identified as "amateur musicians". All but two participants indicated that they enjoyed listening to popular and electronic music, and participants reported listening to a mean of 3.1 hours of music each day (SD = 1.8).

All participants exhibited unimpaired binaural hearing, assessed at the beginning of each experimental session as in Experiment 1. All participants gave informed consent. The study was certified by the McGill University Review Ethics Board.

4.3.2 Materials

This task presented the 24 musical excerpts selected in the preliminary study and employed in Experiment 1, exemplifying low, medium, and high levels of beat salience (8 excerpts per level), with a duration of 17s per stimulus. In this task, the excerpts were presented without the pulse train overlay (cowbell beat) used in Experiment 1.

4.3.3 Apparatus

The sound isolation booth, along with its computer and audio equipment set-up, were the same as in the preliminary study. Additionally, a Roland HPD15 HandSonic electronic hand percussion MIDI controller (Roland, Hamamatsu, JP) was connected to the computer's USB port to record participants' taps during the experiment. The MIDI controller was parameterized for optimal sensitivity to finger tapping. A flashing LED indicator on the drum pad gave visual feedback to reassure participants that their taps had been registered.

4.3.4 Design

Each trial in the experimental session presented one of the 24 musical stimuli in three successive sub-trials: i) *naïve tapping*; ii) *attentive listening*; and iii) *informed tapping*. In each case, excerpts were presented in full, i.e., played for 17s per sub-trial. Trial order was randomly determined for each participant. Measurements of the onset times of participants' taps on the MIDI drum were taken, and the duration

between the first sounding of the current sub-trial's musical excerpt and each tap onset was recorded.

4.3.5 Procedure

Upon signing the informed consent form, participants entered the testing booth and completed an audiogram test, as in Experiment 1. After passing the audiogram, participants were instructed on their task, and given the opportunity to ask for clarification from the experimenter. An initial block of practice trials was then completed in the presence of the experimenter to familiarize the participant with the experimental interface and task. This practice block presented three trials featuring musical excerpts exhibiting low, medium, and high beat salience that were chosen from outside the main stimulus set. After completion of the practice block, the experimenter then left the participant to complete the main experimental block of 24 trials.

Each trial was comprised of three successive sub-trials presenting the same stimulus. In the first sub-trial ("naïve tapping"), participants tapped along to the musical excerpt, having not heard it before in the context of the experiment. Participants were instructed to start tapping as soon as they were confident that they had found the beat. In the second sub-trial ("attentive listening"), participants were instructed to listen attentively to the beat of the music, without tapping along. In the third sub-trial ("informed tapping"), participants tapped along again. Participants were explicitly not required to replicate their tapping from the first sub-trial; rather, they were instructed to tap wherever seemed natural.

As after Experiment 1, participants completed a beat salience rating task (Experiment 3) after completion of Experiment 2. Finally, participants filled out a questionnaire on their musical background. All participants received $10 CAD upon completion of the experimental session as compensation for their time.

4.3.6 Analysis

Tapping consistency. In order to determine the effect of beat salience level on participants' consistency while tapping along to the music, and to control for a potential confound of stimulus tempo on tap consistency, we analyzed the standard deviation of intertap intervals (ITI) for each participant and excerpt, log-transformed to address positive skew in the data. We constrained our analysis to ITI obtained during informed tapping, in order to control for potential confounding effects of unexpectedness upon initial (naïve) exposure to each stimulus.

We specified a linear mixed effect model with fixed effects for beat salience level and excerpt tempo (in beats per minute), with random intercepts by-participant and by-excerpt, and a random slope by-participant for beat salience level. Visual inspection of residual plots revealed no apparent violations of the assumptions of normality and homoscedasticity. As in the analysis of response times (Experiment 1), hypothesis testing was conducted using the pamer.fnc module to determine significance (Tremblay & Ransijn, 2013).

Tapping consensus. We visualized the degree of tapping consensus between the individuals within our group of participants by plotting the probability density function $p(t)$ of all participants' tap times for a given stimulus and sub-trial (naïve or informed tapping), where $p(t)$ is the sum of Gaussian kernels placed at each tap

time (Toiviainen & Snyder, 2003). This approach is parameterized by the standard deviation of the Gaussian kernels to be placed; a larger SD results in a smoother distribution, a smaller SD in a more granular distribution. To address the range of tempi and beat salience inherent in our stimuli, we determined this value independently for each stimulus using the Sheather and Jones "Solve-the-Equation Plug-In" approach (Jones, Marron, & Sheather, 1996; Sheather & Jones, 1991), adjusting by a value of 50ms, as used by Toiviainen and Snyder. We found this approach to produce a good compromise between smoothness and resolution of the estimated probability densities.

We hypothesized that the musical stimuli with a highly salient beat would present little ambiguity regarding appropriate tapping times, and stimuli exhibiting low beat salience conversely would present greater ambiguity. Accordingly, we expected the probability distributions of high salience stimuli to be characterized by clearly defined, spiky peaks, as participants are more likely to tap at mutually consistent times; and those of low salience stimuli to be defined by diffuse peaks of unclear definition, as different individuals' taps are not well correlated.

In order to obtain a more objective grasp of tapping consensus and to facilitate analysis, it was desirable to quantify such distributional cues. Toiviainen and Snyder obtained a quantitative consensus measure of their participants' tap consensus based on information-theoretic considerations (Toiviainen & Snyder, 2003, p. 62). In our case, this was less straightforwardly applicable, as our stimuli exhibit a range of tempi and, necessarily, differ widely in beat salience; Toiviainen and Snyder's participants

tapped along to different excerpts of the same piece by J.S. Bach, the last of four organ duettos for solo performers (BWV 805).

As our primary concern was with the degree of regularity and periodicity inherent in the tap densities, we instead employed a digital signal processing approach to quantify the cross-participant tapping consensus for each stimulus. We computed the Fourier transform of the mean-deducted autocorrelation function for each tapping density distribution. We then determined the coefficient of variation of each Fourier plot in order to obtain a normalized measure of dispersion inherent in the frequency distribution. The value of this coefficient is larger when most of the energy in the Fourier plot is concentrated within specific frequency bands, i.e., when participants' taps tend to be placed consistently; it thus accounts for the issue of different participants adhering to different levels of the metric hierarchy. Conversely, the value of the coefficient is smaller as energy is dispersed more arbitrarily across the frequency space of the Fourier plot, i.e., when taps tend to be placed inconsistently across participants (Figure 4–8, p. 153).

Naïve vs. informed tapping. To determine the impact of preparatory exposure to the music on tapping behavior, we conducted two-sample Kolmogorov-Smirnov tests on the cross-participant tap distributions obtained during the naïve tapping and informed tapping phases of each trial.

4.3.7 Results

Tapping consistency. We constrained our analysis to the standard deviation of inter-tap intervals ($SDITI$) obtained in the informed tapping sub-trials. This resulted in 768 data points (32 participants tapping to 24 musical excerpts). The

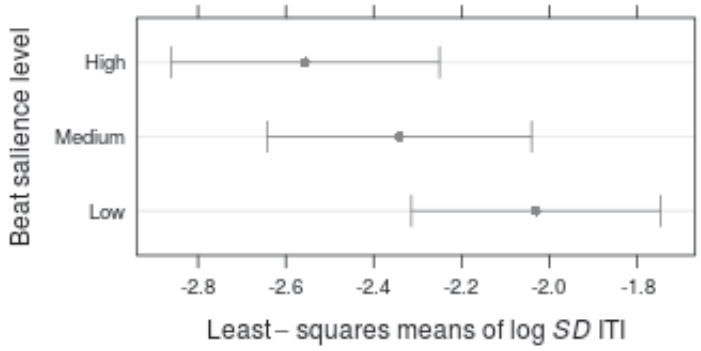


Figure 4–5: Least-squares means and 95% confidence intervals of tap consistency (log standard deviation of intertap interval) according to beat salience level. The tap consistencies of low and high beat salience level stimuli differ significantly.

$SDITI$ ranged from 12 ms to 6.9 seconds across all informed tapping trials, with a median of 103 ms, and a 5% trimmed mean of 126 ms. As the observation with the maximum $SDITI$ value was very distant from the remaining distribution, we removed it as an outlier; the next-largest value was 2.6 seconds. The remaining 767 observations were analysed using a linear mixed effect model with fixed effects for beat salience level and excerpt tempo (in beats per minute), with random intercepts by-participant and by-excerpt, and a random slope by-participant for beat salience level. As hypothesized, we determined a significant main effect of beat salience level on tapping consistency ($logSDITI$), $F = 4.92, p < .01$, with a range of DF of 640 – 760 and p-values differing by less than .001 over this range. No significant effect of tempo was detected.

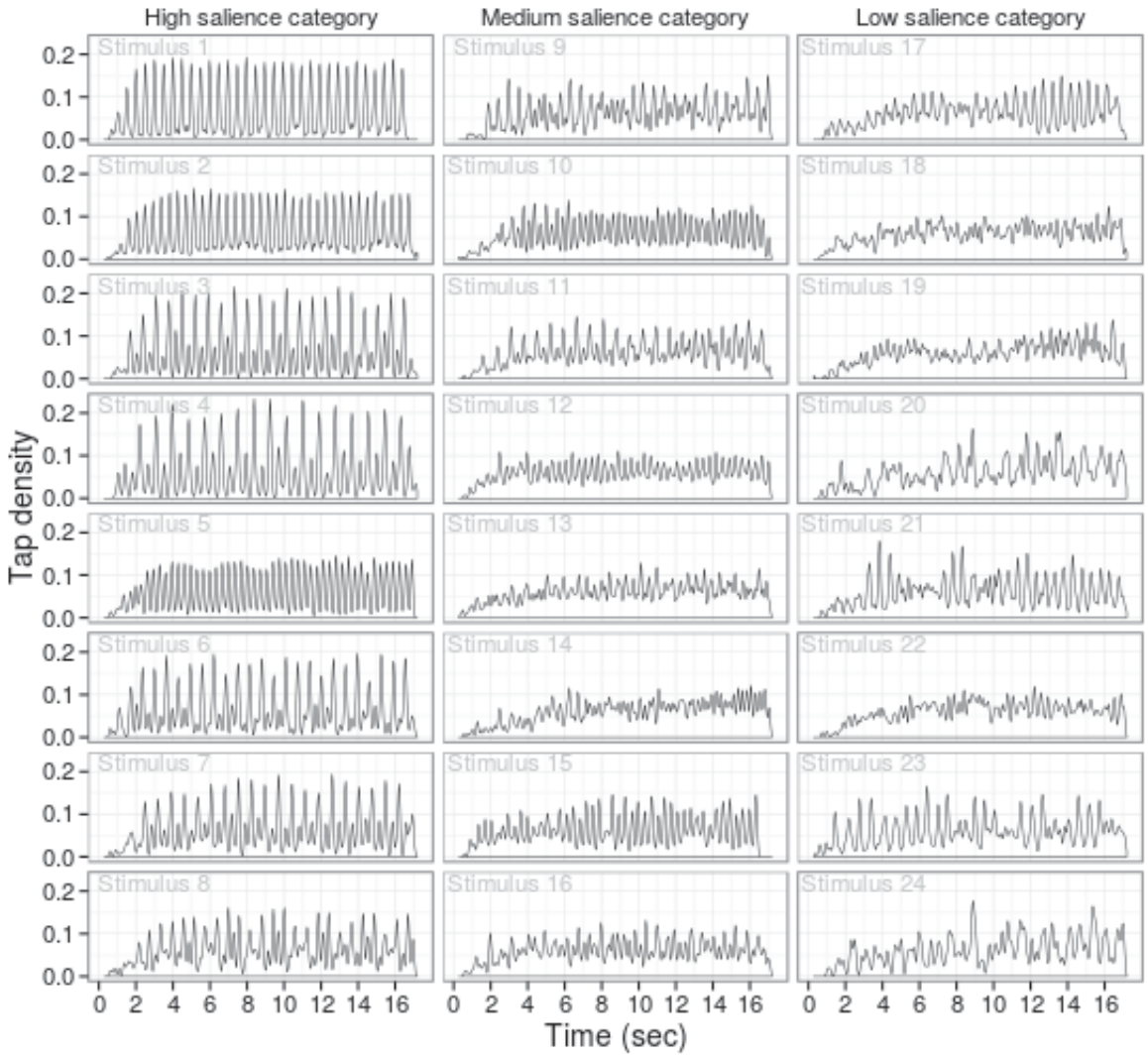


Figure 4–6: Tapping density distributions generated by participants during the naïve tapping sub-trial.

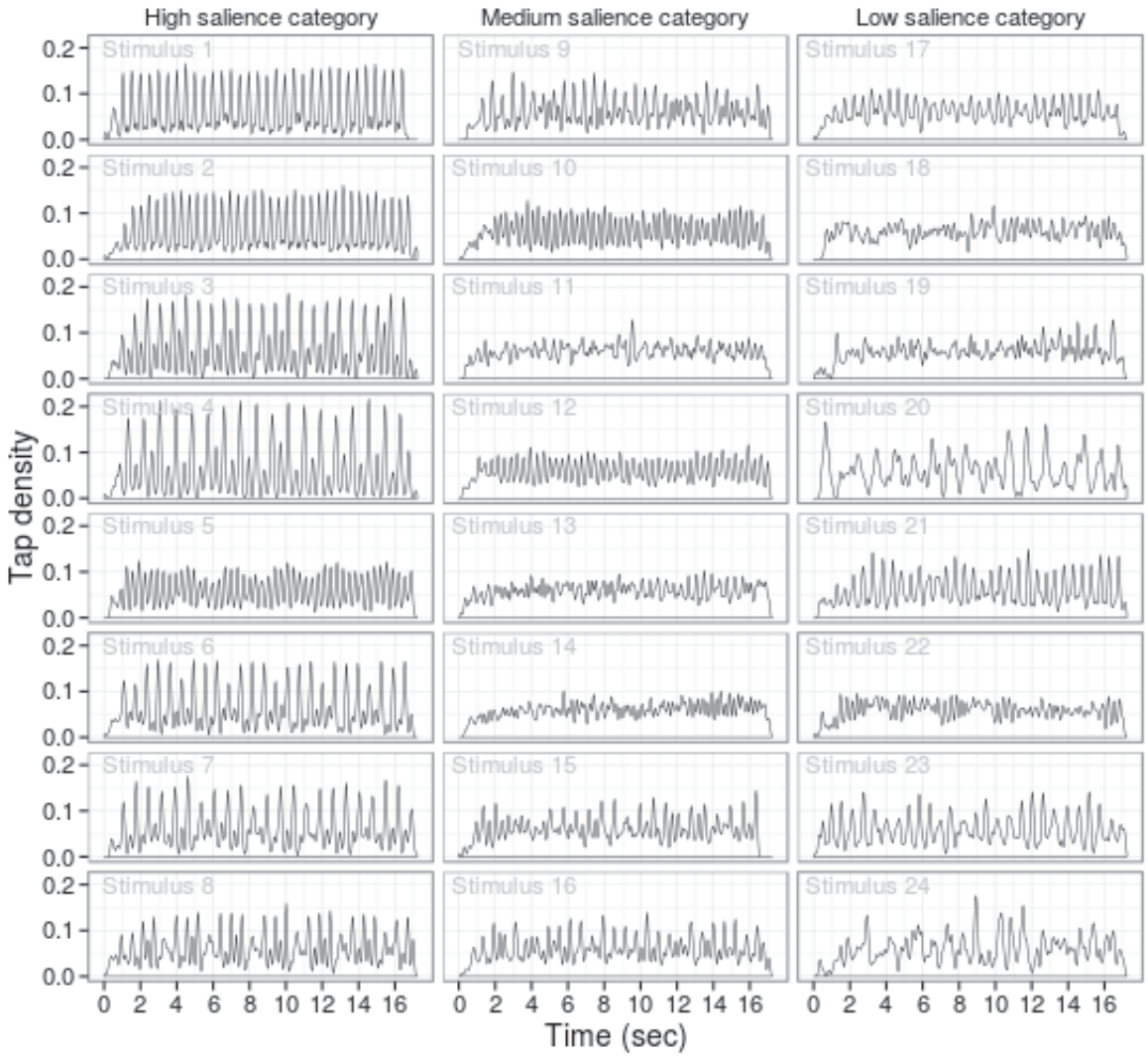


Figure 4–7: Tapping density distributions generated by participants during the informed tapping sub-trial.

Post-hoc testing conducted using pairwise contrasts of least-squares means (Figure 4–5, page 149) revealed a significant difference in $SDITI$ of the low vs. high

beat salience categories: low - high beat salience $SDITI = 0.53, t(22.3) = 3, p < .05$ (Tukey-adjusted for multiple comparisons).

Tapping consensus. The set of observations consists of 50,880 tap events, with 46.8% obtained during the naïve tapping sub-trials, and 52.7% obtained during informed tapping. The remaining 263 observed taps were (erroneously) produced during the attentive listening sub-trial, in-between naïve and informed tapping; they were discarded from analysis.

The tapping density distributions for each stimulus are visualised in Figure 4–6 (naïve tapping sub-trials; page 150) and Figure 4–7 (informed tapping sub-trials; page 151). As hypothesized, the plotted density distributions exhibit a striking difference between the comb-like shape of the stimuli exhibiting high levels of beat salience, and the more diffuse and less regular peaks of the stimuli exhibiting medium or low beat salience. The visual differences between medium and low salience stimuli are less clear-cut, as both categories include some diffuse but regular distributions, and some highly irregular distributions.

To quantify the degree of tapping consensus inherent in each tapping density distribution, we determined the mean-deducted autocorrelation function (ACF) of the distribution, and the Fourier-transform of the ACF: FFT(ACF); see Figure 4–8, p. 153.

We then calculated the coefficient of variation (CoV) of the FFT(ACF) curve to provide a normalised measure of dispersion that is low when taps were placed inconsistently (i.e., low consensus), and high when taps were placed consistently (i.e.,

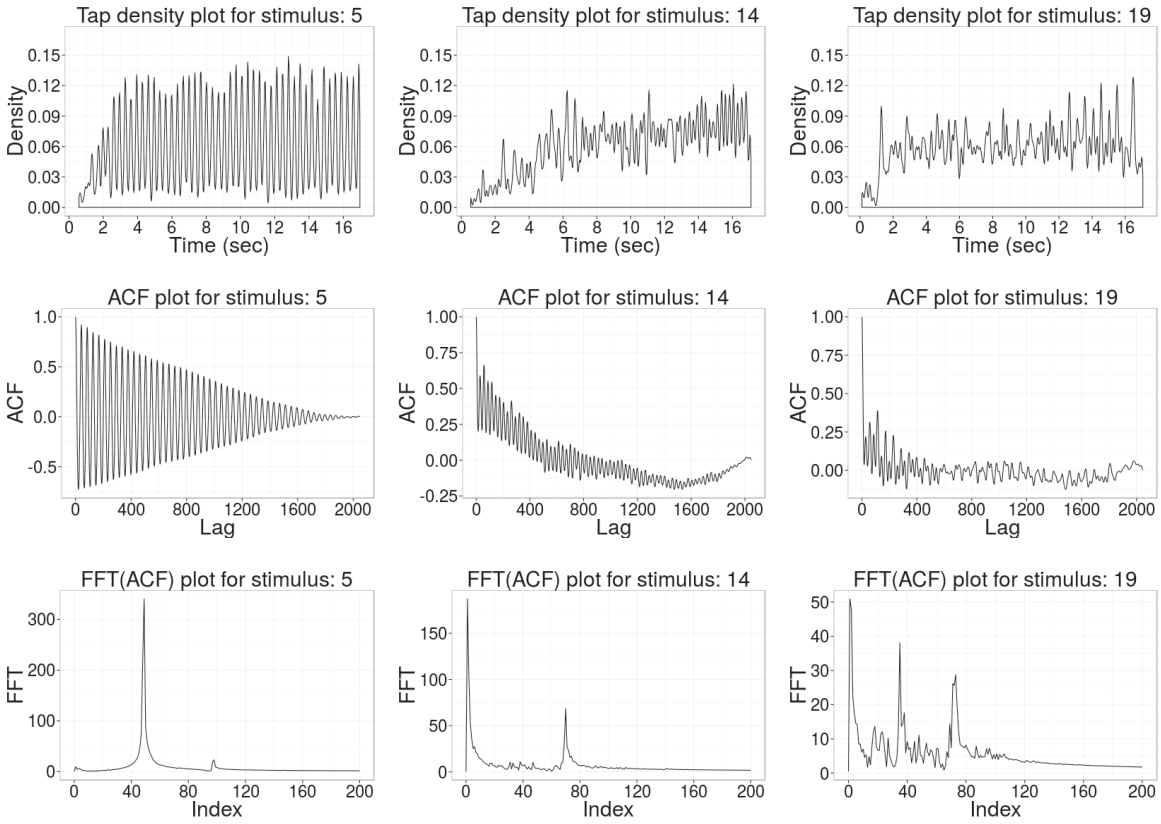


Figure 4–8: Quantifying tapping consensus: tapping density, mean-deducted autocorrelation function (ACF), and Fourier transform – FFT(ACF). *Left*: Plot for stimulus with the highest tapping consensus measure: Stimulus 5 (high beat salience level), sub-trial 1 (naïve tapping), CoV=5.57. Note: Scale of y-axis differs between plots for the different stimuli. *Middle*: Plot for stimulus with the median tapping consensus measure: Stimulus 14 (medium beat salience level), sub-trial 1 (naïve tapping), CoV=3.85. *Right*: Plot for stimulus with the lowest tapping consensus measure, as determined by the CoV of FFT(ACF): Stimulus 19 (low beat salience level), sub-trial 3 (informed tapping), CoV=1.98.

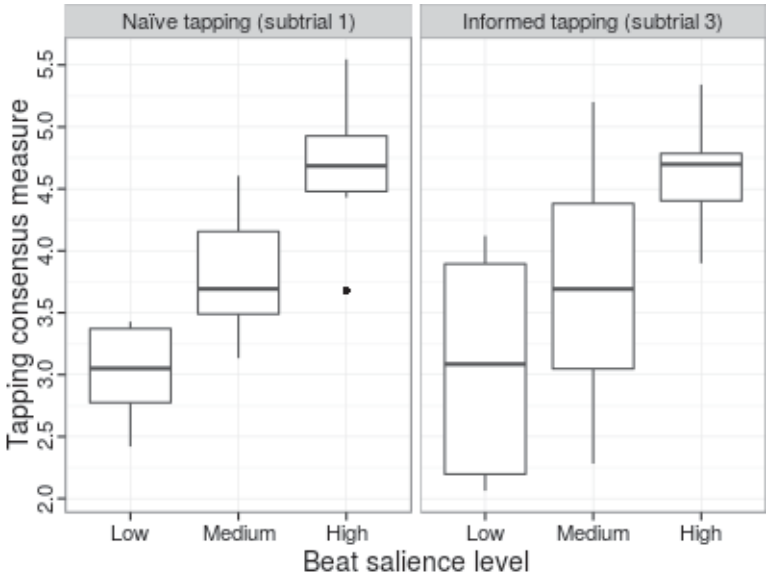


Figure 4–9: Tapping consensus according to beat salience level.

high consensus) between participants. The outcomes of this measure are summarized in Figure 4–9, grouped by tapping sub-trial (naïve and informed tapping), and aggregated according to beat salience level.

Naïve vs. informed tapping. The degree of overlap between the cross-participant tap density distributions obtained in naïve tapping compared to informed tapping sub-trials is visualized for each excerpt in Appendix G (Figures G1 – G3). The distance between these distributions, quantified by calculating the two-sample Kolmogorov-Smirnov statistic for the differences in tap densities between the sub-trials for each stimulus, is presented in Table 4–3. Significant differences in the distributions were detected for all low beat salience stimuli; all but one of the medium beat salience stimuli; and for two of the high beat salience stimuli (eight stimuli per

Table 4–3: Naïve vs. informed tapping: Kolmogov-Smirnov statistic for each stimulus

Beat Salience Level	Stimulus	Kolmogov-Smirnov D
High	1	0.04
	2	0.047
	3	0.044
	4	0.059
	5	0.056*
	6	0.051
	7	0.068*
	8	0.071*
Medium	9	0.076*
	10	0.07 **
	11	0.078*
	12	0.058*
	13	0.075**
	14	0.11 ***
	15	0.048
	16	0.061*
Low	17	0.12 ***
	18	0.083**
	19	0.071*
	20	0.15 **
	21	0.069*
	22	0.12 ***
	23	0.078*
	24	0.11 *

*** $0 < p < .001$ ** $.001 < p < .01$ * $.01 < p < .05$ (FDR-adjusted p-values)

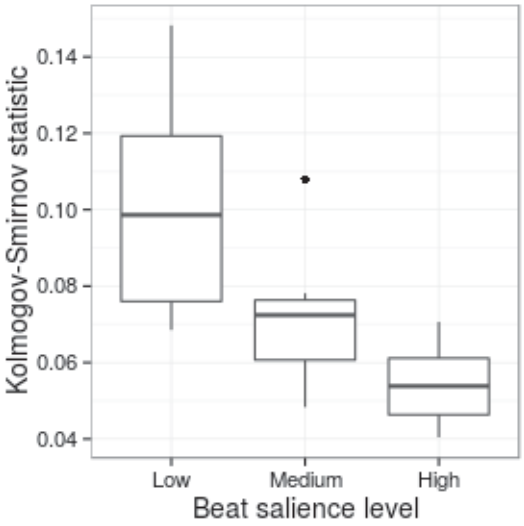


Figure 4–10: Two-sample Kolmogov-Smirnov D statistic quantifying the distance between the cross-participant naïve tapping and informed tapping distributions.

beat salience level). Figure 4–10 visualises the statistic for each stimulus, grouped according to beat salience level.

4.3.8 Discussion

Our results demonstrate a significant correspondence between beat salience level and the consistency at which individuals tap along to the music, as well as between beat salience level and the consensus between individuals (i.e., the inter-participant consistency in tapping behaviour). In particular, tapping consistency and consensus are significantly stronger in stimuli exhibiting high beat salience, compared to those exhibiting low beat salience; with medium beat salience stimuli overlapping on either side of the spectrum. This evidence supports the hypothesized relationship between beat salience and sensorimotor synchronization performance in the context of tapping to the beat in the music.

Our hypothesis regarding the reason for this relationship – that highly salient beats present little ambiguity regarding appropriate tap positions, and conversely, that low beat salience presents a stronger ambiguity in terms of when to tap – is supported by observing the differences in tap distributions generated during naïve and informed tapping for the individual stimuli. These differences remain insignificant for six of the eight high beat salience stimuli; but the distributions differ significantly for seven of the eight medium salience stimuli, and for all eight low salience stimuli. This suggests that little tapping ambiguity was present in most of the high salience stimuli, with participants' tap behaviour during the second tapping attempt (and thus on their third exposure to the stimulus) largely replicating their tap behaviour on initial exposure to the music. Conversely, a larger degree of ambiguity could explain the observed inconsistencies in tapping behaviours exhibited in response to the medium and low beat salience stimuli. The implied differences of degree of ambiguity are apparent on visual inspection of the overlap of naïve and informed tapping distribution plots (Appendix G), with the two curves largely tracing each other neatly in the high salience category, and strikingly lesser overlap of the two distributions in response to medium and (especially) low salience stimuli.

4.4 Experiment 3: Beat salience ratings

We conducted a beat salience rating experiment in order to investigate whether subjective assessment of beat salience was consistent between participants, and to quantify the degree of correspondence between an individual's subjective assessment of beat salience and their beat perception or synchronization task performance.

4.4.1 Participants

After completion of their primary task (either Experiment 1 or Experiment 2), all participants were instructed to complete a beat salience rating experiment. Technical issues prevented one participant, who had successfully completed the sensorimotor synchronization task (Experiment 2), from taking part in the beat salience rating experiment. All other participants completed the experiment (N=62; 34 female; mean age=23.7, SD=5.9).

4.4.2 Materials

This experiment presented the same 24 musical excerpts used in Experiments 1 and 2, as selected in the preliminary study. As in Experiment 2, the excerpts were presented without the pulse train overlay (cowbell beat) featured in Experiment 1.

4.4.3 Apparatus

The equipment and experimental interface were as described for the preliminary study.

4.4.4 Design and procedure

As in the preliminary study, participants listened to each musical excerpt, presented in random order, and were tasked to rate beat salience on a 5-point Likert scale, and familiarity and liking on 7-point analog categorical scales. Beat salience was again defined as "the perceptual strength or prominence of a beat. A value of 5 indicates that you could easily tap to the beat, while a value of 1 indicates that you simply could not tap to the beat." The experiment proceeded as in the preliminary study; however, as there were only 24 excerpts in this case, participants completed only one experimental block following a practice block presenting 6 excerpts drawn

from outside the main stimulus set. During the experimental block, the 24 excerpts of the main set were presented in an order generated randomly for each participant.

4.4.5 Analysis

Inter-rater reliability. Participants' inter-rater reliability was quantified by determining intraclass correlation coefficients (ICC; Shrout & Fleiss, 1979) for both groups of raters: those who had just completed the perception experiment (Experiment 1; N=30); and those who had just completed the sensorimotor synchronization task (Experiment 2; N=31). As all participants rated all musical excerpts, and as we are treating both judges (participants) and targets (musical excerpts) as randomly sampled from a larger population, a two-way random effects anova model corresponding to Shrout & Fleiss' Case 2 was conducted (p. 421):

$$x_{ij} = \mu + a_i + b_j + (ab)_{ij} + e_i$$

where x_{ij} denotes the ith judge's rating of the jth target; μ, the overall population mean of the ratings; a_i, the difference from μ of the mean of the ith judge's ratings; b_j, the difference from μ of the jth target's "true score" (i.e., the mean across many repeated ratings of the jth target); $(ab)_{ij}$, the degree to which the present rating reflects a departure of the ith judge's usual rating tendencies on the jth target; and e_{ij}, the random error in the ith judge's scoring of the jth target (Shrout & Fleiss, 1979, p. 421).

Corresponding ICC measures were calculated to determine the reliability of a single, typical rater, and the average reliability of all raters in aggregate—corresponding to $ICC(2,1)$ and $ICC(2,k)$ in Shrout and Fleiss' notation, respectively. For both

cases, we further determined ICC measures for consistency and for absolute agreement between raters. In the consistency measure, systematic differences between different judges' rating behaviors are discounted, whereas such differences remain relevant in the absolute agreement measure. All measures were determined using the icc function of the irr R package (Gamer, Lemon, Fellows, & Singh, 2012).

Predictive validity. In order to assess whether participants' beat salience ratings significantly relate to task performance in the perception and sensorimotor synchronization tasks (Experiments 1 and 2), we applied further linear mixed effects model analyses, specifying beat salience ratings as fixed effects, and response time (Experiment 1) and the standard deviation of intertap interval (Experiment 2) as dependent variables. We also investigated the proportion of correct responses (Experiment 1) as a dependent variable, specifying a corresponding logistic linear mixed effects model with beat salience ratings as fixed effects. Each of these models was specified with random intercepts for the individual participants.

4.4.6 Results

The beat salience ratings obtained from both participant groups for musical excerpts in each of the three beat salience categories (as established in the pre-study) are summarized in Figure 4–11. ICC agreement and consistency measures for the reliability of a single, typical rater—$ICC(2,1)$—and for the average reliability of all raters in aggregate—$ICC(2,k)$—are listed in Table 4–4.

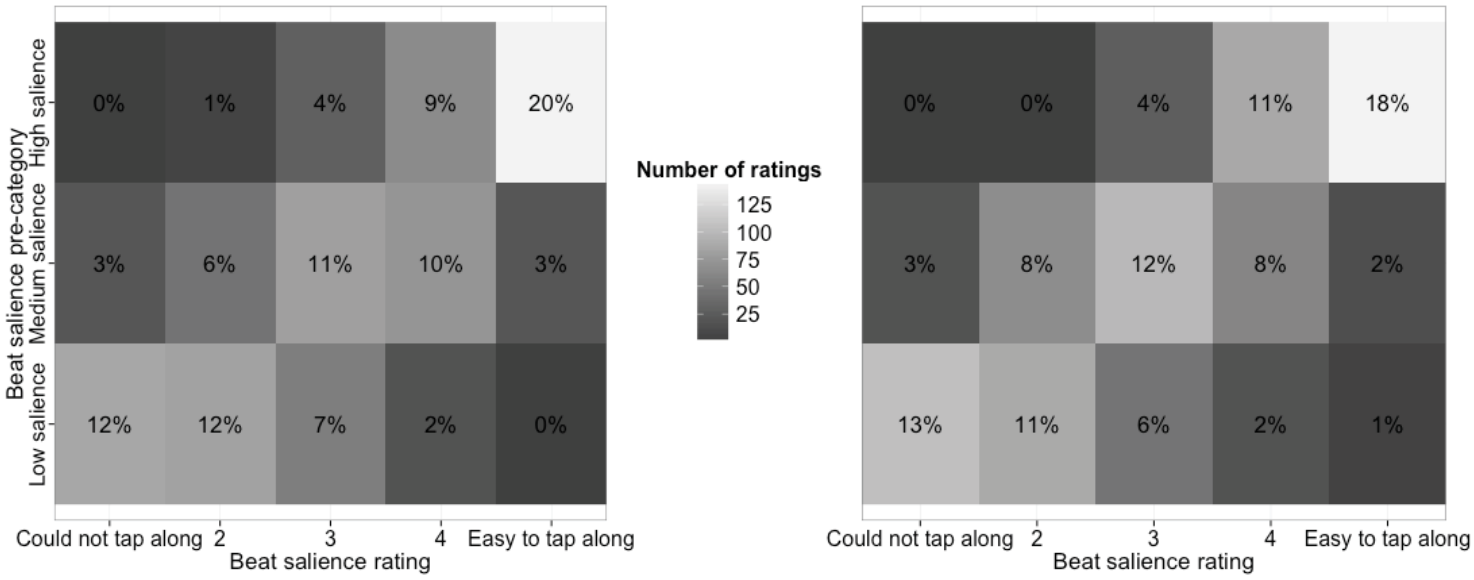


Figure 4–11: Beat salience ratings obtained in Experiment 3. Horizontal axes: participants' ratings of each musical excerpt on a 5-point Likert scale: from 1—low beat salience, could not tap along to the music; to 5—high beat salience, could easily tap along to the music. Vertical axes: beat salience categories determined for each excerpt by expert inter-rater agreement in the preliminary study. *Left:* Beat salience ratings provided by participants after completing the beat perception experiment (Experiment 1 group; total number of ratings: 720). *Right:* Beat salience ratings provided by participants after completing the sensorimotor synchronization experiment (Experiment 2 group; total number of ratings: 744).

Table 4–4: Intraclass correlation coefficients (ICC) quantifying the degree of inter-rater reliability of participants in the beat salience rating experiment (Experiment 3).

Participants	Unit of analysis[a]	Measure[b]	ICC	95% CI
Experiment 1 group (N=30)	Single	Consistency	0.62	$.49 < ICC < .77$
		Agreement	0.57	$.44 < ICC < .73$
	Average	Consistency	0.98	$.97 < ICC < .99$
		Agreement	0.98	$.96 < ICC < .99$
Experiment 2 group (N=31)	Single	Consistency	0.57	$.44 < ICC < .73$
		Agreement	0.56	$.42 < ICC < .71$
	Average	Consistency	0.98	$.96 < ICC < .99$
		Agreement	0.98	$.96 < ICC < .99$

[a] Inter-rater reliability determined for a typical rater (Single) or for all raters in aggregate (Average).
[b] Inter-rater reliability discounting systematic differences in individual raters' behaviors (Consistency) or penalizing such differences (absolute Agreement).

Predictive validity. Our analysis reveals significant main effects of participants' beat salience ratings on both response times ($F = 17.7, p < .0001$, DF lower bound–upper bound: 881–911) and on proportion correct ($\chi^2(4) = 121.2, p < 0.0001$) in the beat perception task (Experiment 1), and on the standard deviation of inter-tap intervals ($F = 6.4, p < .0001$, DF lower bound–upper bound: 704–735) in the sensorimotor synchronization task (Experiment 2).

4.4.7 Discussion

The results of the beat salience rating experiment reveal a high degree of inter-rater consistency, both within and between experimental groups. Furthermore, there is a good correspondence between the ratings obtained in the context of this experiment and the beat salience classifications determined by expert agreement in the preliminary study. Finally, our analysis suggests that the ratings obtained in this

task significantly relate to dependent variables corresponding to task performance in both the beat perception and sensorimotor synchronization experiments. Several implications relating to the validity and generalizability of beat salience ratings may be drawn from these findings.

There were only relatively minor differences between individuals' subjective impressions of beat salience in our exemplars of popular and electronic music. This consistency appears to hold regardless of the presence of musical training, as the ratings of our non-expert participants correspond closely to the pre-categorizations obtained from the clustered ratings of trained music technology graduate students.

Figure 4–11 illustrates the high degree of consistency between responses obtained from participants who had previously completed the beat perception experiment, and from those who had instead completed the sensorimotor synchronization experiment. This consistency is of particular interest as the beat salience rating task was cast in terms of prospective tapping difficulty. The sensorimotor synchronization group had direct evidential access to this information, having spent approximately 45 minutes tapping to these stimuli immediately prior to the start of the beat salience rating experiment; whereas the beat perception group had been critically attending to the music, without the requirement of coordinated motor activity. Nevertheless, responses obtained from both groups are very similar, suggesting that beat salience ratings may be framed in terms of prospective tapping difficulty with a high degree of construct validity.

Our analysis demonstrates a significant correspondence between participants' beat salience ratings and their individual task performance on the same stimuli,

in both the perception and sensorimotor synchronization contexts. It is thus the case that participants' prospective assessments were not only consistent as measured against the responses of other individuals, but also accurate in terms of the rater's performance in the previously completed experimental tasks. This suggests that beat salience ratings framed in terms of prospective tapping difficulty also have a degree of predictive validity, in both perception and synchronization contexts.

4.5 Evaluating a Computational Measure of Beat Salience

The three experiments reported in the preceding sections investigate different psychometric measures of beat salience. One of the motivating goals of this research has been to establish the validity and reliability of beat salience as a potential query mechanism or relevance criterion for music information retrieval systems serving the common music information need of finding music to move to. Given that the suitability of beat salience has been established in this context, an algorithmic measure that approximates human performance on beat salience related tasks is required before such a mechanism can be integrated into a music information retrieval system.

We thus report on a preliminary investigation of one possible method of implementation. The music information retrieval task of automated beat tracking is highly analogous to human beat finding tasks such as the one presented in our sensorimotor synchronization experiment, in that both the algorithmic and human process produce a vector along the musical time-line encoding temporal positions upon which the presence of the beat is asserted.

For our purposes, we apply a committee-based beat tracker (Holzapfel, Davies, Zapata, Oliveira, & Gouyon, 2012) to our musical excerpts. The committee-based approach combines a set of five state-of-the-art beat tracking algorithm (Dixon, 2007; Degara et al., 2012; D. P. Ellis, 2007; Oliveira, Gouyon, Martins, & Reis, 2010; Klapuri, Eronen, & Astola, 2006), each employing their own distinct method of implementation, into a beat tracking committee. Beat annotations are then produced by selecting the output of the component algorithm that provides the closest match to the output of all other component algorithms, i.e., the maximal mutual agreement (MaxMA) among the committee, for each potential beat location. Holzapfel et al. (2012) demonstrate that this MaxMA approach significantly outperforms each individual component beat tracker when evaluated against a large data set of manually annotated beat positions.

An interesting property of the decision making process based on choosing the output with maximal mutual agreement is that the degree of agreement among the committee members is itself informative. Zapata et al. (2012) explore the utility of the mean mutual agreement (MMA) among committee members, calculated over entire musical tracks, demonstrating that MMA may be used to meaningfully assign confidence thresholds on the overall quality of automatic beat annotations of large data sets.

Following the intuition that the degree of confidence in the vector of beat positions along the musical time-line of a given excerpt relates to the level of beat salience present in the music, we conducted an investigation of the relationship between the MMA measure produced by the algorithm for our musical excerpts, and the beat

Table 4–5: Beat salience ratings and mean mutual agreement: Significance testing for fixed effects.

Fixed effect	DF (lower – upper)	F	% deviance explained
Beat salience rating (SR)	1,393 – 1,454	216.24 ***	37.25
Prior study (PS)	1,393 – 1,454	0.8	0.034
SR×PS interaction	1,393 – 1,454	0.54	0.093

*** $0 < p < .001$

salience ratings produced by our participants in the course of Experiment 3. The raw MMA scores produced for our stimuli ranged from 0.72 to 3.74, where higher values indicate better mutual agreement between the component algorithms of the committee. We rescaled these values to a range with minimum of 1 and a maximum of 5, in order to facilitate interpretation against the beat salience ratings obtained from our participants in Experiment 3 on a scale of 1 (very high beat salience) to 5 (very low beat salience). We fit a linear mixed effects model, specifying the rescaled MMA measure as the dependent variable, beat salience ratings, the rater's prior completed study (Experiment 1 or Experiment 2), and their interaction as fixed effects, and by-participant intercepts as random effects. Given the limited number of musical stimuli employed in our study, and thus the small number of MMA data points (one per musical excerpt), neither the individual items nor their beat salience pre-categories were encoded in the random effects structure to avoid over-fitting the model. Difference coding was applied to investigate the relative contributions of adjacent levels of beat salience ratings in terms of their relationship to the MMA measure.

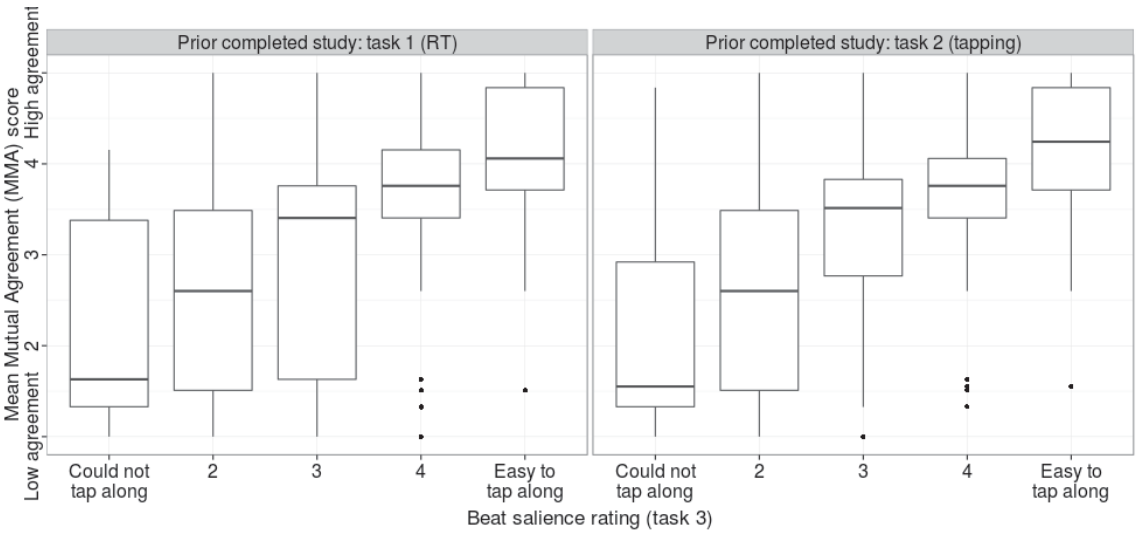


Figure 4–12: Mean mutual agreement (MMA) measure produced by a committee-based beat tracker (Holzapfel et al., 2012), related to beat salience ratings obtained from our participants in Experiment 3 for the same musical excerpts. Participants had previously completed either a response time study measuring beat salience perception (Experiment 1), or a tapping (beat finding) study investigating sensorimotor synchronization (Experiment 2). MMA has been rescaled to a continuous range from 1—lowest agreement between the committee's constituent beat tracking algorithms; to 5—best agreement. Beat salience ratings were obtained on a Likert scale from 1—low beat salience, could not tap along to the music; to 5—high beat salience, could easily tap along to the music.

4.5.1 Results and discussion

Our results demonstrate a significant relationship between the beat salience ratings obtained from participants during the beat salience rating task (Experiment 3), and the MMA measure produced by the committee-based beat tracker, in the context of the 24 musical excerpts forming our stimulus set. As expected from the outcomes of Experiment 3, there were no significant differences for the beat salience ratings obtained from participants who had previously completed Experiment 1, compared to ratings obtained from those who had previously completed Experiment 2, in terms of correspondence with the MMA measure (Table 4–5, p. 166).

We also found significant differences between each successive level of the beat salience rating scale (Table 4–6, p. 169), suggesting that the relationship between beat salience ratings and the MMA measure is present along the entire scale, and that the scale's level of granularity is appropriate for the purpose of this investigation. That said, we observed a fairly large degree of overlap between the range of MMA ratings for each successive salience level (Figure 4–12), suggesting that a coarser conception of beat salience (e.g., a binary indication of low or high beat salience) might be preferable if accuracy is required, as may be the case in the context of music information retrieval system design.

The MMA measure was not specifically developed to correspond to human beat salience judgements; Zapata et al. (2012) discuss its use in improving and evaluating the quality and accuracy of beat tracking algorithms. As such, the preliminary results presented here are especially encouraging – both by demonstrating that algorithmic

Table 4–6: Difference coded contrasts investigating the relative differences of adjacent levels of beat salience ratings (BSR) in terms of their relationship to the MMA measure.

Change in BSR	Change in MMA	Std. Error	t-value
BSR:1 → BSR:2	0.34	0.12	−2.75 **
BSR:2 → BSR:3	0.58	0.11	−5.23 ***
BSR:3 → BSR:4	0.59	0.11	−5.55 ***
BSR:4 → BSR:5	0.52	0.11	−4.86 ***

*** $0 < p < .001$ ** $.001 < p < .01$

Note: The MMA measure was rescaled to a continuous range from 1 (best agreement between committee beat tracker algorithms) to 5 (least agreement). BSR were obtained on a categorical scale of 1 (very low beat salience) to 5 (very high salience).

approximation of human beat salience judgement is feasible, and by suggesting that improved performance is likely feasible with further research in this direction.

4.6 General discussion: Beat salience

We have presented an investigation into beat salience, a measure of the perceptual prominence of the beat in the context of moving to music. Making use of a stimulus set of musical excerpts drawn from the popular and electronic music repertoire and categorized according to their beat salience by expert inter-rater agreement, we have employed a convergent methods approach to investigate the effect of varying levels of beat salience on psychometric measures of the beat induction and motor synchronization phases of sensorimotor synchronization, and on the subjective experience of movement affordance in the context of tapping along to music. With a view toward practical applications in music information retrieval, we have further established the validity and reliability of beat salience ratings among musically untrained individuals, and we have proposed an initial approach to generating algorithmically

derived equivalents of such ratings in order to aide in the implementation of a relevance criterion or query mechanism based on beat salience for the common use case of finding music to move to.

The results of our response time experiment, based on the beat perception task of Iversen and Patel's Beat Alignment Test (2008), suggest a strong correspondence between beat salience level and perception task performance in the "on-beat" condition, where a superimposed cowbell beat is synchronized without phase error with the beat of the music. In the "early" and "late" conditions where a small (75ms) phase error was present, we have found evidence of a temporal captor effect, whereby participants tended to respond that the cowbell was on the beat in high beat salience exemplars, regardless of the actual presence of phase error; this effect appears to be limited to musical stimuli exhibiting high beat salience, and was not significantly evident in responses to trials in the medium and low beat salience conditions. To explain this effect, we have posited the existence of a perceptual integration window around the musical beat, within which musical onsets are perceived as occurring in synchronization with the music; the width of this window may be parameterized by the salience of the beat, such that the 75ms phase offsets of our early and late conditions tend to fall within this range in the beat integration processes of our participants for high salience excerpts, but tend to fall outside for medium and low salience excerpts. We leave the exploration of this hypothesis for future research.

The results of our beat finding task, employing a tapping paradigm with a long tradition in sensorimotor synchronization research (Repp, 2005; Repp & Su, 2013),

further suggest a good correspondence between beat salience level and various measures of tapping performance, both individually (e.g., the participant's tapping consistency in terms of the standard deviation of inter-tap intervals over a given trial), and in terms of cross-participant consensus measured by quantifying the periodicity present in the combined tap density across all individuals' responses to a given musical excerpt. This correspondence is particularly evident when inspecting task performance in the high beat salience condition, compared to responses to trials in the medium and low salience conditions.

The outcomes of our analysis of the beat salience ratings provided by participants after the completion of their respective primary task indicate a high degree of inter-rater consistency both within and across experimental groups (i.e., participants who had previously completed either the respose-time task investigating beat induction, or the beat finding task investigating sensorimotor synchronization). This consistency between individuals in their ratings of beat salience related measures provides a replication of the findings of Tzanetakis, Essl, and Cook (2002) and Lartillot et al. (2008). Further, these ratings, produced by individuals without significant musical training, correspond strongly with the beat salience categorizations determined by inter-rater agreement among the ten music technology post-graduate students who participated in our normative preliminary study. This correspondence provides evidence of the reliability of such a measure across a range of musical expertise. Finally, these experiential ratings correlated with individuals' task performance in both the response-time and beat finding tasks, suggesting that individuals are accurate in their experiential assessment of beat salience in terms of movement affordance.

Taken together, the results of these experiments suggest a consistent role for beat salience, understood in terms of movement affordance in the context of tapping along to music, on an experiential, and on a performance-based level. This is manifested in terms of reliability (consistency within and across responses from individuals sampled from a population of enthusiastic music listeners lacking significant musical training); construct validity (participants' beat salience ratings, responses to the question of how easily they could tap along to the music, correlate strongly with measures of tapping performance across different levels of beat salience determined in our preliminary study); and ecological validity, tapping to music being a common, everyday activity. This suggests a role for beat salience as a relevance criterion or query mechanism for music information retrieval systems, catering to the use case of finding music to move to, as in the contexts of dance, exercise, or work. Our finding of significant correlations between our participants' beat salience ratings and the Mean Mutual Agreement (MMA) measure of Zapata et al.'s ensemble beat tracking algorithm (2012) provide a preliminary suggestion for a possible approach toward implementing such a system.

Several limitations of the present studies must be outlined clearly at this stage, relating mainly to our set of musical stimuli. From the initial phases of study design, we made the conscious decision to emphasize ecological validity by choosing commercial, polyphonic musical excerpts, exhibiting a range of tempi, selected from the repertoire of popular and electronic music commonly listened to by a broadly specified Western musical audience; this decision was made explicitly to broaden our investigation beyond the synthetic, laboratory generated stimuli typically used in

previous research. Such diversity naturally comes at a cost in terms of control over potential musical confounds. We have tried to address this issue by selecting stimuli exhibiting a range of time signatures, metric complexity, and attack density across the different levels of beat salience, by accounting for differing tempi in our analyses where possible, and by attempting to minimise experimenter bias by employing only a subset of the initially compiled collection of musical excerpts, selected as best exemplifying low, medium, and high levels of beat salience in our preliminary study. However, it is certainly possible that confounds present in the interactions particularly of extra-temporal musical facets (e.g., of timbre, orchestration, or melodic complexity) and the cognitive process of sensorimotor synchronization remain unaccounted for.

The relatively small number of musical excerpts, prescribed by the practical need to limit each experimental session to a manageable duration (approximately 1 hour), also needs to be considered; while our experimental design generated many data points for each excerpt and participant, resulting in robust outcomes in the context of the stimulus set employed here, care must be taken in applying these outcomes to popular and electronic music at large without further study. This is particularly true in the analysis of the correlation between our participants' responses and the measures generated by the beat tracking algorithm, for which we have only 24 data points; while results here are promising, further research employing a larger stimulus set, ideally sampled randomly from the musical repertoire, is required before widespread conclusions may be drawn. Given the evidence presented here for the correspondence of beat salience ratings and sensorimotor task performance, a study

that collects human beat salience ratings on a large stimulus set may be sufficient to properly evaluate algorithm performance, even in absence of finer-grained psychometric measures such as those gathered in our beat perception and beat finding tasks. Crowd-sourcing solutions such as Amazon's Mechanical Turk platform could serve to bootstrap such a study.

As the ability to move along to music is widespread in the general population, we have chosen in the present studies to recruit participants without significant musical training. It would be interesting to pursue further research in this direction that explicitly investigates the effect of musical expertise on beat salience ratings and task performance. Such a study might look at musicians in general, and perhaps at highly trained percussionists in particular; we would expect similar overall trends to the ones described here, but with improved reaction times, accuracy measures, and tapping consistency due to greater musical expertise. The tendency for participants to identify the superimposed cowbell as playing on the beat in the high salience condition of Experiment 1, even with phase error present, would be particularly interesting to investigate further; we would expect the hypothesized integration window that spans the range of temporal offsets accepted as "on the beat" to vary with musical training, and thus that the acceptance threshold under which temporal offsets are tolerated would be smaller in magnitude for participants with greater degrees of musical expertise.

We hope that the theoretical and practical outcomes of this research will contribute to ongoing conversations in communities of music perception and cognition,

and of music information retrieval research. By employing experimental methodologies to gather empirical evidence on human information processing in order to inform information system design, we are building on a tradition of information retrieval research employing cognitive perspectives (see e.g., Ingwersen, 1996; Ingwersen & Järvelin, 2005) and on more recent cognitive trends in the music information domain (e.g., Aucouturier & Bigand, 2012; Honing, 2010; Müllensiefen & Frieler, 2004; consider also the annual CogMIR seminar[2] , first held in 2011). We see much potential in the overlap of these disciplines, both from the perspective of the creation of research tools for the music psychologist, and in terms of empirical insights on the user's behaviour to inform the development of music information systems. Calls for a greater focus on the user have cropped up repeatedly in the MIR literature (e.g., Cunningham et al., 2003; Downie et al., 2009; Lee & Cunningham, 2012; Schedl et al., 2013; Taheri-Panah & MacFarlane, 2004; Weigl & Guastavino, 2011), and cognitive approaches in particular are required to address the problem of operationalizing measures of musical similarity and relevance in the highly subjective, mood- and situation-dependent context of music listening behaviour. This problem, formulated in Downie's "multiexperiential challenge" to music information retrieval (Downie, 2003) in the early days of the field, continues to loom formidably today.

[2] Seminar on cognitively based music informatics research (CogMIR): `http://www.cogmir.org`

CHAPTER 5
Conclusion

By successive technological advances over the last two decades, digital music listening has become a pervasive part of everyday life: from readily-available broadband access, to high-quality compressed audio encoding formats, via file-sharing applications of questionable legal status, to more legitimate music streaming services, and powered by the ubiquitous mobile computing platforms carried in the pockets of a large proportion of the population. Massive quantities of music are readily available for our listening pleasure. Consequently, the design and implementation of efficient and effective methods of music information storage, organization, and retrieval to support listeners' music information needs and behaviours has become a research priority attracting focused attention from a wide range of disciplines. Under the banner of music information retrieval, these multifarious influences are united into one coherent field of research.

MIR research has produced algorithmic solutions addressing problems on a spectrum from low-level digital signal processing (e.g. note onset detection) to high-level digital musicology (e.g. structural segmentation). This research has predominantly focused on the design and implementation of the components of music information systems. While investigations of the (potential) users of such systems have remained sparse in comparison, the body of findings relating to listeners' music information needs and behaviours is nevertheless substantive, and growing.

In this dissertation, we have synthesised and built upon this body of research, in order to obtain an understanding of the current state of knowledge on music and relevance, and to inform the operationalization of relevance criteria for music information retrieval. This work has been motivated by Downie's *multiexperiential challenge*: in the abstract information domain of music, lacking a well-defined lexicon and representational semantics, relevance cannot be inferred without taking into account the listener's perception and experience of the music, varying according to mood, situation, and circumstance (Downie, 2003). In Chapter 2, we have presented a wide-ranging, systematic analysis of MIR user studies, focussing on findings informing a broader understanding of the notion of relevance. In doing so, we have conceptualized relevance according to our application of Saracevic's stratified model of relevance interactions (Saracevic, 1997; 2007b; Weigl & Guastavino, 2013).

In Chapters 3 and 4, we have narrowed our focus, targeting the temporal facets of music that we have termed *rhythmic information* in order to investigate their role in the definition of criteria of topical relevance, a measure of "the mapping between a query and a set of information objects," and situational relevance, a measure of the user's judgement of "the relationship between information and information need situations in a certain time" (Jansen & Rieh, 2010, p. 1525).

The overarching goal has expressly *not* been to "solve" the problem of relevance for music—such a lofty aim is precluded by the complex natures both of musical information, and of the notion of relevance more broadly. Rather, this dissertation represents an acknowledgement that relevance, a key notion for information retrieval in general (Saracevic, 2007b), and for MIR in particular (Downie, 2003), has thus far

received relatively sparse attention in the MIR literature, with system design often guided by "anecdotal evidence of user needs, intuitive feelings for user information seeking behavior, and a priori assumptions of typical usage scenarios" (Cunningham et al., 2003), and with user studies generating findings that touch on relevance diffusely, without focussing on this important aspect of MIR explicitly.

We further acknowledge the experiential, and thus psychological, aspects of music relevance: "User studies are useful, but some kind of cognitive framework is required if we are to better understand the music seeking behaviour of MIR users" (Taheri-Panah & MacFarlane, 2004, p. 459). As such, we have chosen to adapt a conceptual model of relevance that explicitly takes into account the user's cognition and affective state; and we have investigated music perception and cognition in two sets of experiments aiming to inform relevance measures at the intersection of (algorithmic) information processing concerning the temporal facets of music, and the perceptive aspects of melody identification, beat perception, and sensorimotor synchronisation—areas that have remained under-explored in MIR user research (section 2.3.2).

5.1 Research outcomes

We now return to the three research questions posed in section 1.1, addressing each in order to reflect on the outcomes of the work presented in this dissertation.

5.1.1 How may the notion of relevance be conceptualized for music information research?

In the preceding discussion, we have shown the conceptualization of relevance for the music information domain to be a non-trivial task. Lacking the concrete semantic mappings afforded by lexical meaning in textual information, and operating in a

space where hedonic, non-goal-oriented information seeking is the norm (Laplante & Downie, 2011), MIR researchers must consider notions of relevance as complex and multifaceted as musical information itself.

In Chapter 2, we have shown Saracevic's stratified model of relevance interactions (Saracevic, 1997; 2007b), applied to the music information domain (Weigl & Guastavino, 2013), to provide an adequate fit for this purpose, acting as conceptual scaffolding to house findings from MIR user research pertaining to relevance. Saracevic's model was chosen as an established framework from the textual IR domain that provides the abstraction and flexibility to facilitate cross-application to music, while retaining a level of analytical granularity sufficient to capture details of interest. We have shown that the conceptualization of music relevance afforded by this framework accommodates the triangulation of findings from diverse user studies, providing an overview of the current state of knowledge in the field, and anticipate that this approach will fruitfully inform both future user research, and MIR system design.

5.1.2 What is the role of rhythmic information in melody identification, and what are the implications in formulating an experiential criterion of topical relevance in MIR?

Music perception and cognition studies investigating melody identification by presenting familiar melodic stimuli with systematically distorted pitches or rhythms to participants for identification (Hébert & Peretz, 1997; Kuusi, 2009; White, 1960) have found identification performance to be substantially more impaired under pitch distortion, as compared to rhythmic distortion, suggesting a strongly diminished role of rhythmic information in melody identification.

The outcomes of such studies have implications on the formulation of experiential measures of topical relevance for MIR, if a system supporting query methods such as query-by-humming is to correspond to human identification judgments. Indeed, the formulation of musical queries based solely on melodic pitch contour—whether successively sounded pitches go up, down, or stay the same, compared to the immediate precedent—is the basis for Parsons coding, one of the earliest proposed means of content-based music information retrieval (Parsons, 1975).

In Chapter 3, we have reassessed the role of rhythmic cues in the context of melody identification, primarily by tackling the assumption implicit in previous research that the imposition of isochrony (setting each note duration to a globally constant value) *nullifies* rhythmic information. Finding strongly inhibited identification performance when rhythmic information is randomized, we demonstrate that rhythm's role in melody identification has been underestimated in previous studies. MIR systems targeting tasks pertaining to melody identification—for instance, query by humming, query by performance, or cover song detection—thus cannot afford to discount rhythmic information if an experiential measure of topical relevance, mapping the user's query to a music information object in a way that corresponds to the user's perception of melodic identity, is to be realized.

5.1.3 Can beat salience inform a valid and reliable criterion of situational relevance in MIR?

Finding music to move to—in the context of manual work, dance, or exercise—is a common music information need, identified in numerous MIR user studies (e.g. Brinegar & Capra, 2011; Cunningham et al., 2003; Cunningham et al., 2007; Greasley & Lamont, 2009; Lonsdale & North, 2011; Robertson, 2006). Such activities all

involve sensorimotor synchronization - the act of synchronizing one's movements to the external musical stimulus. As such, the predictability and clarity of the musical beat is of particular importance, potentially overriding other criteria such as matching the listener's taste profile, to the point where a user's preferred "gym music" may not be appreciated in other listening contexts (Cunningham et al., 2007).

In Chapter 4, we have presented an investigation of *beat salience*, a measure of the perceptual prominence of the beat in the context of moving to music. Employing a convergent-methods approach to investigate the distinct stages of perceptual beat induction, sensorimotor synchronization, and beat salience judgement, we have demonstrated the validity and reliability of beat salience as a situational relevance criterion for use cases involving synchronized movement to music; beat salience ratings demonstrated good consistency between musically trained and untrained individuals, and were predictive of task performance in both the beat perception and sensorimotor synchronization tasks. Promisingly, the pre-existing committee-based beat-tracking implementation of Holzapfel et al. (2012) produces a secondary output measure—the mean mutual agreement of the constituent algorithms forming the beat-tracking committee—that aligns pleasingly well with human beat salience judgements. We propose this algorithm as a starting point for the operationalization of beat salience as a measure of situational relevance in MIR.

5.2 Contributions

We now review the theoretical, methodological, and practical contributions generated by the work presented in this thesis.

5.2.1 Theoretical contributions

In our application of Saracevic's stratified model of relevance interactions to the music information domain, we have established a first comprehensive conceptualization of the notion of relevance for MIR. In the systematic analysis and synthesis presented in Chapter 2, we have provided an overview of the state of knowledge on relevance in a large subsection of the MIR user literature, as well as identifying several underexplored areas, pointing the way for future research. Further, our study represents a validation of the stratified model: in formulating the model, Saracevic acknowledged that it "has not yet enough details for experimentation and verification" (1997, p. 318). While our work in Chapter 2 lacks experimental components, it nevertheless represents a comprehensive case study of the model's use in representing the relevance interactions implicit in a large corpus of knowledge.

In Chapter 3, we have addressed misconceptions about the nullification of rhythmic information in previous literature in order to re-evaluate the role of rhythm in melody identification, finding it to provide significantly stronger identification cues than previously acknowledged. The tendency of participants to err within the melody's category in cases of partial identification and misidentification further demonstrates the influence of connotative, contextual cues beyond the pitch and durational cues in the constituent musical structure, demonstrating that listeners attend at superordinate levels of abstraction.

In Chapter 4, we have identified beat salience as a valid and reliable measure of situational relevance in the context of finding music to move to. In the course of our experimentation, we have also discovered some surprising aspects of beat perception.

In particular, the tendency of participants to overwhelmingly identify the cowbell pulse on trials presenting high salience stimuli as playing on the beat, even in phase error conditions, suggests the presence of an integration window spanning a short time period around the beat onset time, with pulse onsets occurring within this span being "captured" by the musical stimulus, resulting in a perception of being "on the beat". The presence of this effect in high salience beat stimuli, as well as its absence in trials presenting medium and low beat salience stimuli—perhaps due to a relationship between the size of this integration window, and the distinctiveness of the metric grid established during beat induction, corresponding to the experienced level of beat salience—could be usefully explored in future research.

5.2.2 Methodological contributions

We have implemented a novel coding tool to support the coding activities underlying Chapter 2; this tool provides a means of coordinating the activities of multiple researchers applying the stratified model of relevance interactions to findings identified in large collections extracted from the literature, and could easily be modified to accommodate alternative conceptual models.

In Chapter 3, we build on (and largely replicate) the methods of preceding melody identification studies, but apply complex rhythmic alterations to our stimuli. Alterations such as the "reordered" condition have previously been applied in research on melody perception (e.g. Prince, 2011), but we are the first to apply them in a melody identification context comparable to the work of White (1960) and Hébert and Peretz (1997).

The surprising lack of predictive value of responses indicating a "feeling of knowing" in the altered rhythmic conditions of Chapter 3, regarding the subsequent ability to provide any identifying information on the unaltered melodies, carries broader methodological implications for music perception and cognition research, as it suggests that listeners' experiential feelings of familiarity when attending to unidentified melodies may be unreliable in certain conditions. Experiments in music perception and cognition employing "real" musical stimuli taken from the wider musical catalogue commonly attempt to control for confounds of participant familiarity by asking participants to self-report their familiarity with each musical piece. In light of our results, it may be recommendable to supplement this self-reporting with an identification task, to see whether participants are able to corroborate their experienced feelings of familiarity with any identifying information.

In Chapter 4, we avail ourselves of methods already applied in previous research (e.g. Iversen & Patel, 2008). However, we believe our application of convergent methodologies (response-time based beat induction task; tapping-based beat finding task; experiential beat salience rating task), analyzing the degree of correspondence in task performance in comparable experimental conditions in order to inform the construct validity and generalizability of a potential relevance criterion for MIR applications, to be a novel approach. In adopting these convergent methodologies, we provide a means of directly addressing the gap between algorithmic information processes and the listener's perception and cognition, an important area that has remained under-represented in the literature (section 2.3.2).

5.2.3 Practical contributions

The work presented in the preceding chapters has produced a number of practical contributions with direct applicability in future MIR research and systems design. The tools developed during the coding and analytical stages of the work presented in Chapter 2, as well as the corpus of coded studies and findings established during this work, has been made available to the MIR research community[1] . We anticipate that the analytical interface, enabling the exploration of our coded findings at the level of stratum interactions, or at the level of co-occurring sub-stratum descriptors, to be of particular interest to designers of specific MIR systems or algorithms interested in obtaining a quick overview of the available user research pertaining to their particular implementation area. We anticipate that these tools will prove valuable to MIR user researchers, to assist in hypothesis generation for future research. Further, we hope that other MIR user researchers might be interested in contributing their own findings to the database, in order to grow the corpus and make it more fully representative of the state of knowledge in the field. Given sufficient interest, this could form part of a web-service, managed by multiple stakeholders, addressing the lack of systematic synthesis of results in MIR user research, and the disconnect between system designers and user studies researchers, as proposed by Lee and Cunningham (2012).

Insights regarding the role of rhythmic information as a relevance criterion, in the case of known-item musical queries (e.g. query by humming), and in the case of

[1] Available at: http://relevance.linkedmusic.org

situational queries revolving around a specific musical facet in a particular use context (requesting music exhibiting highly salient beats, as part of a query to find music to move to), have direct practical applications for MIR systems design. In particular, we demonstrate that rhythmic information cannot be neglected in determinations of relevance as a response to such queries; and we propose a pre-existing set of algorithms—the committee-based beat-tracking implementation of Holzapfel et al. (2012)—as a promising starting point in the implementation of a computational measure of beat salience as an experiential situational relevance criterion.

5.3 Limitations

We must acknowledge a number of limitations arising from the methodologies and paradigmatic approach applied in the work presented here.

The coding activities underlying the systematic analysis and synthesis of the MIR literature presented in Chapter 2 necessarily rely on interpretations by the coders employed in this research. Care was taken to adopt an approach that was as rigorous and systematic as possible, by means of an iterative, mutual coding process, followed by group discussions among the four researchers until convergence in our coding approach was reached, and making use of a coding tool that propagated new terms added by any one coder to all others, in order to promote consistency. Nevertheless, we acknowledge the potential bias inherent in the individual judgements underlying our coding decisions; a different group of individuals would likely not have produced a corpus of coded findings identical to the one presented in Chapter 2, although we are confident that the overall distribution of studies and findings identified would remain comparable.

This issue could be addressed to some extent by enabling the authors of MIR user studies to perform the coding process for their own studies in future work, should sufficient interest to do so arise. Of course, this would not eliminate the need for interpretation, but it would guarantee that the process was driven by optimal background knowledge of the studies in question.

In Chapters 3 and 4, we make use of empiricist approaches focussed on psychometric measurement of the cognitive and perceptual processes of our participants as they complete our experimental tasks. As this work was conducted with the purpose of producing practical insights to inform relevance criteria for MIR in mind, we employ a post-positivist paradigm, rejecting the use of artificially generated stimuli and instead adopting excerpts of "real" music that might be encountered in everyday listening. In this decision, we promote ecological validity at the expense of tight control over potential confounds from musical facets not pertaining to rhythmic information.

Practical considerations relating to the length of our experiments, and thus the demands on our participants, further prescribed limitations on the amount of musical stimuli that could be presented. While our stimulus selection was guided by normative studies, again to promote ecological validity, it is conceivable that the stimulus sets are not entirely representative of the musical catalogue at large.

The relatively limited size of our stimulus sets is a particular issue in the analysis of the correlation between the mean mutual agreement (section 4.5), which, although promising, is based on a small number of observations, limited to the 24 musical stimuli employed in Chapter 4. Given the correspondence of beat salience ratings and sensorimotor task performance, algorithmic implementations of beat salience

measures could be evaluated more thoroughly if human beat salience ratings were obtained on a larger scale, even in absence of finer-grained psychometric measures such as those obtained in the experimental work presented in Chapter 4. Such a corpus of beat salience ratings could be bootstrapped using crowdsourcing platforms such as the Amazon Mechanical Turk, the suitability of which has been previously demonstrated in MIR research, e.g. in the context of obtaining music similarity judgements (Lee, 2010).

5.4 Future work

The work presented in this dissertation can be extended in a number of directions.

Depending on community response, we hope to maintain and extend the corpus of findings collected in the work presented in Chapter 2, and thus to contribute toward increasing the cohesiveness and impact of the user-centric aspects of MIR research.

The surprising lack of predictive value of "feelings of knowing" of distorted melodies in Chapter 3 could be usefully explored in further experimental work; partly to illuminate this finding in itself, but also because self-reported familiarity with musical stimuli is part of standard pre-screening questionnaires in many music perception and cognition experiments. If these turn out to be unreliable, this may have broad methodological consequences.

The outcomes of Chapter 3 demonstrate the contributions of connotative, contextual cues to experienced melodic identity. At time of writing, I am conducting research on enriching musical information with extra-musical, contextual semantics,

by employing Linked Data and Semantic Web technologies, as a research associate on the multi-institutional Fusing Audio and Semantic Technologies for Intelligent Music Production and Consumption (*FAST-IMPACt*) project (e.g. De Roure et al., 2015). I anticipate that this work will provide a means of progressing toward enhanced relevance measures in MIR, further occupying the gap between information processing and music perception and cognition.

Two directions clearly suggest themselves as avenues toward building on the work presented in Chapter 4. On the one hand, a large-scale crowdsourcing of human beat salience judgements would provide a useful dataset to function as "ground-truth" in the evaluation of algorithms estimating beat salience. We have shown the inter-rater reliability of such judgements, among musically trained and untrained individuals, as well as the correspondence between these judgements and perception and synchronization task performance. Such a crowd-sourced corpus could thus be deemed valid and reliable in terms of reflecting a situational relevance criterion in the context of finding music to move to with some degree of confidence

On the other hand, an important continuation of our investigations would study the correspondence between performance in a tapping task, as compared to performance in a task involving whole-body coordination. The tapping task has been employed as a proxy for the idea of moving to music in the research presented here, largely for practical reasons: it is very easy to employ. However, the assumed correspondence between tapping performance and, say, performance running on a treadmill, needs to be verified. Future experimental work could perform such verification,

hopefully to corroborate the validity of beat salience as a relevance criterion beyond "mere" tapping.

This dissertation has provided a broad account of the vital notion of relevance in music information retrieval, establishing the first comprehensive conceptualization of relevance for the domain of musical information, and describing areas that have received greater, or lesser, degrees of research attention thus far. At a lower level, the work presented herein has focussed on very specific aspects of music and of relevance, investigating the role of rhythmic information as a criterion for topical and situational relevance in MIR. Motivated by Downie's multiexperiential challenge (Downie, 2003), we acknowledge that we have not "solved" the problem of relevance and music. Instead, we have provided a solid grounds for future work specifying this notion, and a significant advance toward formulating the "rigorous and practical theories" concerning the nature of relevance that the challenge demands. We hope that this work will be useful to the MIR research community.

Appendix A: Distribution of findings according to relevance interactions (Chapter 2)

The following tables display the distribution of findings ($N = 866$) according to relevance interactions. Table A-1 displays this distribution ordered by number of interacting strata (i.e. dimensionality of interaction), and then by stratum class from left to right. Table A-2 displays the same distribution, ordered by number of findings encoding the corresponding interaction. For the columns encoding the stratum classes, 1 indicates presence of this stratum in an interaction, and 0 indicates its absence.

Table A-1: Distribution of findings by number of interacting strata.

Content	Processing	Engineering	Interface	Cognitive	Affective	Situational	SocContext	CultContext	#Findings	%Findings	#InteractingStrata
1	0	0	0	0	0	0	0	0	29	3.35%	1
0	1	0	0	0	0	0	0	0	8	0.92%	1
0	0	0	1	0	0	0	0	0	59	6.81%	1
0	0	0	0	1	0	0	0	0	42	4.85%	1
0	0	0	0	0	1	0	0	0	12	1.39%	1
0	0	0	0	0	0	1	0	0	36	4.16%	1
0	0	0	0	0	0	0	1	0	19	2.19%	1
0	0	0	0	0	0	0	0	1	2	0.23%	1
1	1	0	0	0	0	0	0	0	9	1.04%	2
1	0	1	0	0	0	0	0	0	4	0.46%	2
1	0	0	1	0	0	0	0	0	90	10.39%	2
1	0	0	0	1	0	0	0	0	43	4.97%	2
1	0	0	0	0	1	0	0	0	9	1.04%	2
1	0	0	0	0	0	1	0	0	11	1.27%	2
1	0	0	0	0	0	0	1	0	7	0.81%	2
1	0	0	0	0	0	0	0	1	9	1.04%	2
0	1	0	1	0	0	0	0	0	17	1.96%	2
0	1	0	0	1	0	0	0	0	22	2.54%	2
0	1	0	0	0	1	0	0	0	2	0.23%	2
0	1	0	0	0	0	1	0	0	2	0.23%	2
0	0	1	1	0	0	0	0	0	4	0.46%	2
0	0	1	0	1	0	0	0	0	1	0.12%	2
0	0	1	0	0	0	1	0	0	3	0.35%	2
0	0	0	1	1	0	0	0	0	69	7.97%	2
0	0	0	1	0	1	0	0	0	5	0.58%	2
0	0	0	1	0	0	1	0	0	14	1.62%	2
0	0	0	1	0	0	0	1	0	18	2.08%	2
0	0	0	1	0	0	0	0	1	4	0.46%	2
0	0	0	0	1	1	0	0	0	9	1.04%	2
0	0	0	0	1	0	1	0	0	25	2.89%	2
0	0	0	0	1	0	0	1	0	11	1.27%	2
0	0	0	0	1	0	0	0	1	2	0.23%	2
0	0	0	0	0	1	1	0	0	9	1.04%	2
0	0	0	0	0	1	0	1	0	1	0.12%	2
0	0	0	0	0	1	0	0	1	2	0.23%	2
0	0	0	0	0	0	1	1	0	8	0.92%	2
0	0	0	0	0	0	1	0	1	5	0.58%	2
0	0	0	0	0	0	0	1	1	1	0.12%	2
1	1	0	1	0	0	0	0	0	24	2.77%	3
1	1	0	0	1	0	0	0	0	6	0.69%	3
1	1	0	0	0	0	1	0	0	2	0.23%	3
1	1	0	0	0	0	0	0	1	1	0.12%	3
1	0	1	1	0	0	0	0	0	2	0.23%	3
1	0	1	0	1	0	0	0	0	1	0.12%	3
1	0	1	0	0	0	1	0	0	2	0.23%	3
1	0	1	0	0	0	0	1	0	1	0.12%	3
1	0	0	1	1	0	0	0	0	27	3.12%	3
1	0	0	1	0	1	0	0	0	3	0.35%	3
1	0	0	1	0	0	1	0	0	6	0.69%	3
1	0	0	1	0	0	0	1	0	6	0.69%	3

Table A-1: Distribution of findings by number of interacting strata.

Content	Processing	Engineering	Interface	Cognitive	Affective	Situational	SocContext	CultContext	#Findings	%Findings	#InteractingStrata
1	0	0	1	0	0	0	0	1	7	0.81%	3
1	0	0	0	1	1	0	0	0	3	0.35%	3
1	0	0	0	1	0	1	0	0	8	0.92%	3
1	0	0	0	1	0	0	1	0	2	0.23%	3
1	0	0	0	1	0	0	0	1	3	0.35%	3
1	0	0	0	0	1	1	0	0	5	0.58%	3
1	0	0	0	0	1	0	1	0	1	0.12%	3
1	0	0	0	0	1	0	0	1	2	0.23%	3
1	0	0	0	0	0	1	1	0	1	0.12%	3
1	0	0	0	0	0	1	0	1	1	0.12%	3
1	0	0	0	0	0	0	1	1	1	0.12%	3
0	1	0	1	1	0	0	0	0	12	1.39%	3
0	1	0	1	0	1	0	0	0	3	0.35%	3
0	1	0	1	0	0	1	0	0	4	0.46%	3
0	1	0	0	1	1	0	0	0	2	0.23%	3
0	1	0	0	1	0	0	1	0	2	0.23%	3
0	1	0	0	0	0	1	1	0	1	0.12%	3
0	0	1	1	1	0	0	0	0	1	0.12%	3
0	0	1	1	0	0	1	0	0	3	0.35%	3
0	0	1	1	0	0	0	0	1	1	0.12%	3
0	0	1	0	0	0	0	1	1	1	0.12%	3
0	0	0	1	1	1	0	0	0	7	0.81%	3
0	0	0	1	1	0	1	0	0	7	0.81%	3
0	0	0	1	1	0	0	1	0	4	0.46%	3
0	0	0	1	1	0	0	0	1	1	0.12%	3
0	0	0	1	0	1	1	0	0	1	0.12%	3
0	0	0	1	0	0	1	1	0	4	0.46%	3
0	0	0	1	0	0	1	0	1	2	0.23%	3
0	0	0	1	0	0	0	1	1	2	0.23%	3
0	0	0	0	1	1	1	0	0	8	0.92%	3
0	0	0	0	1	1	0	0	1	1	0.12%	3
0	0	0	0	1	0	1	1	0	8	0.92%	3
0	0	0	0	1	0	0	1	1	5	0.58%	3
0	0	0	0	0	1	1	1	0	2	0.23%	3
0	0	0	0	0	0	1	1	1	2	0.23%	3
1	1	0	1	1	0	0	0	0	4	0.46%	4
1	1	0	1	0	1	0	0	0	1	0.12%	4
1	1	0	1	0	0	1	0	0	1	0.12%	4
1	1	0	0	1	0	1	0	0	5	0.58%	4
1	0	1	0	1	0	0	1	0	1	0.12%	4
1	0	0	1	1	1	0	0	0	1	0.12%	4
1	0	0	1	1	0	1	0	0	3	0.35%	4
1	0	0	1	1	0	0	1	0	1	0.12%	4
1	0	0	1	0	1	1	0	0	1	0.12%	4
1	0	0	1	0	0	1	1	0	1	0.12%	4
1	0	0	0	1	1	1	0	0	2	0.23%	4
1	0	0	0	1	1	0	0	1	1	0.12%	4
1	0	0	0	0	0	1	1	1	2	0.23%	4
0	1	0	1	1	0	1	0	0	4	0.46%	4
0	0	0	1	1	1	1	0	0	4	0.46%	4
0	0	0	1	1	0	1	0	1	1	0.12%	4
0	0	0	0	1	1	1	1	0	3	0.35%	4
0	0	0	0	1	1	1	0	1	4	0.46%	4
0	0	0	0	1	1	0	1	1	1	0.12%	4
0	0	0	0	1	0	1	1	1	1	0.12%	4
1	0	0	1	1	1	0	1	0	1	0.12%	5
1	0	0	0	1	1	1	0	1	1	0.12%	5

Table A-2: Distribution of findings by number of findings corresponding to interaction.

Content	Processing	Engineering	Interface	Cognitive	Affective	Situational	SocContext	CultContext	#Findings	%Findings	#InteractingStrata
1	0	0	1	0	0	0	0	0	90	10.39%	2
0	0	0	1	1	0	0	0	0	69	7.97%	2
0	0	0	1	0	0	0	0	0	59	6.81%	1
1	0	0	0	1	0	0	0	0	43	4.97%	2
0	0	0	0	1	0	0	0	0	42	4.85%	1
0	0	0	0	0	0	1	0	0	36	4.16%	1
1	0	0	0	0	0	0	0	0	29	3.35%	1
1	0	0	1	1	0	0	0	0	27	3.12%	3
0	0	0	0	1	0	1	0	0	25	2.89%	2
1	1	0	1	0	0	0	0	0	24	2.77%	3
0	1	0	0	1	0	0	0	0	22	2.54%	2
0	0	0	0	0	0	0	1	0	19	2.19%	1
0	0	0	1	0	0	0	1	0	18	2.08%	2
0	1	0	1	0	0	0	0	0	17	1.96%	2
0	0	0	1	0	0	1	0	0	14	1.62%	2
0	1	0	1	1	0	0	0	0	12	1.39%	3
0	0	0	0	0	1	0	0	0	12	1.39%	1
1	0	0	0	0	0	1	0	0	11	1.27%	2
0	0	0	0	1	0	0	1	0	11	1.27%	2
1	1	0	0	0	0	0	0	0	9	1.04%	2
1	0	0	0	0	1	0	0	0	9	1.04%	2
1	0	0	0	0	0	0	0	1	9	1.04%	2
0	0	0	0	1	1	0	0	0	9	1.04%	2
0	0	0	0	0	1	1	0	0	9	1.04%	2
1	0	0	0	1	0	1	0	0	8	0.92%	3
0	1	0	0	0	0	0	0	0	8	0.92%	1
0	0	0	0	1	1	1	0	0	8	0.92%	3
0	0	0	0	1	0	1	1	0	8	0.92%	3
0	0	0	0	0	0	1	1	0	8	0.92%	2
1	0	0	1	0	0	0	0	1	7	0.81%	3
1	0	0	0	0	0	0	1	0	7	0.81%	2
0	0	0	1	1	1	0	0	0	7	0.81%	3
0	0	0	1	1	0	1	0	0	7	0.81%	3
1	1	0	0	1	0	0	0	0	6	0.69%	3
1	0	0	1	0	0	1	0	0	6	0.69%	3
1	0	0	1	0	0	0	1	0	6	0.69%	3
1	1	0	0	1	0	1	0	0	5	0.58%	4
1	0	0	0	0	1	1	0	0	5	0.58%	3
0	0	0	1	0	1	0	0	0	5	0.58%	2
0	0	0	0	1	0	0	1	1	5	0.58%	3
0	0	0	0	0	0	1	0	1	5	0.58%	2
1	1	0	1	1	0	0	0	0	4	0.46%	4
1	0	1	0	0	0	0	0	0	4	0.46%	2
0	1	0	1	1	0	1	0	0	4	0.46%	4
0	1	0	1	0	0	1	0	0	4	0.46%	3
0	0	1	1	0	0	0	0	0	4	0.46%	2
0	0	0	1	1	1	1	0	0	4	0.46%	4
0	0	0	1	1	0	0	1	0	4	0.46%	3
0	0	0	1	0	0	1	1	0	4	0.46%	3
0	0	0	1	0	0	0	0	1	4	0.46%	2
0	0	0	0	1	1	1	0	1	4	0.46%	4
1	0	0	1	1	0	1	0	0	3	0.35%	4
1	0	0	1	0	1	0	0	0	3	0.35%	3
1	0	0	0	1	1	0	0	0	3	0.35%	3
1	0	0	0	1	0	0	0	1	3	0.35%	3
0	1	0	1	0	1	0	0	0	3	0.35%	3
0	0	1	1	0	0	1	0	0	3	0.35%	3
0	0	1	0	0	0	1	0	0	3	0.35%	2
0	0	0	0	1	1	1	1	0	3	0.35%	4
1	1	0	0	0	0	1	0	0	2	0.23%	3
1	0	1	1	0	0	0	0	0	2	0.23%	3

Table A-2: Distribution of findings by number of findings corresponding to interaction.

Content	Processing	Engineering	Interface	Cognitive	Affective	Situational	SocContext	CultContext	#Findings	%Findings	#InteractingStrata
1	0	1	0	0	0	1	0	0	2	0.23%	3
1	0	0	0	1	1	1	0	0	2	0.23%	4
1	0	0	0	1	0	0	1	0	2	0.23%	3
1	0	0	0	0	1	0	0	1	2	0.23%	3
1	0	0	0	0	0	1	1	1	2	0.23%	4
0	1	0	0	1	1	0	0	0	2	0.23%	3
0	1	0	0	1	0	0	1	0	2	0.23%	3
0	1	0	0	0	1	0	0	0	2	0.23%	2
0	1	0	0	0	0	1	0	0	2	0.23%	2
0	0	0	1	0	0	1	0	1	2	0.23%	3
0	0	0	1	0	0	0	1	1	2	0.23%	3
0	0	0	0	1	0	0	0	1	2	0.23%	2
0	0	0	0	0	1	1	1	0	2	0.23%	3
0	0	0	0	0	1	0	0	1	2	0.23%	2
0	0	0	0	0	0	1	1	1	2	0.23%	3
0	0	0	0	0	0	0	0	1	2	0.23%	1
1	1	0	1	0	1	0	0	0	1	0.12%	4
1	1	0	1	0	0	1	0	0	1	0.12%	4
1	1	0	0	0	0	0	0	1	1	0.12%	3
1	0	1	0	1	0	0	1	0	1	0.12%	4
1	0	1	0	1	0	0	0	0	1	0.12%	3
1	0	1	0	0	0	0	1	0	1	0.12%	3
1	0	0	1	1	1	0	1	0	1	0.12%	5
1	0	0	1	1	1	0	0	0	1	0.12%	4
1	0	0	1	1	0	0	1	0	1	0.12%	4
1	0	0	1	0	1	1	0	0	1	0.12%	4
1	0	0	1	0	0	1	1	0	1	0.12%	4
1	0	0	0	1	1	1	0	1	1	0.12%	5
1	0	0	0	1	1	0	0	1	1	0.12%	4
1	0	0	0	0	1	0	1	0	1	0.12%	3
1	0	0	0	0	0	1	1	0	1	0.12%	3
1	0	0	0	0	0	1	0	1	1	0.12%	3
1	0	0	0	0	0	0	1	1	1	0.12%	3
0	1	0	0	0	0	1	1	0	1	0.12%	3
0	0	1	1	1	0	0	0	0	1	0.12%	3
0	0	1	1	0	0	0	0	1	1	0.12%	3
0	0	1	0	1	0	0	0	0	1	0.12%	2
0	0	1	0	0	0	0	1	1	1	0.12%	3
0	0	0	1	1	0	1	0	1	1	0.12%	4
0	0	0	1	1	0	0	0	1	1	0.12%	3
0	0	0	1	0	1	1	0	0	1	0.12%	3
0	0	0	0	1	1	0	1	1	1	0.12%	4
0	0	0	0	1	1	0	0	1	1	0.12%	3
0	0	0	0	1	0	1	1	1	1	0.12%	4
0	0	0	0	0	1	0	1	0	1	0.12%	2
0	0	0	0	0	0	0	1	1	1	0.12%	2

Appendix B: Stratum co-occurrence: distribution of findings and of studies (Chapter 2)

The following tables provide an overview of stratum pairings in terms of the distribution of findings ($N = 866$), and of studies ($N = 176$) encoding at least one corresponding finding. Table B-1 is ordered by stratum pairing. Table B-2 is ordered according to the magnitude of rank difference for the corresponding stratum interaction between the two distributions (|*studies ranking*−*findings ranking*|).

Table B-1: Stratum co-occurrence (distribution of findings & studies), by stratum pairing.

Stratum A	Stratum B	Num. Findings	Num. Studies	Findings Ranking	Studies Ranking	Rank Diff.
Content	Content	352	109	3	3	0
Content	Processing	53	33	16	13	-3
Content	Engineering	11	8	35	31	-4
Content	Interface	179	65	5	6	1
Content	Cognitive	113	64	9	7	-2
Content	Affective	31	19	23	23	0
Content	Situational	52	30	17	16	-1
Content	Social Context	25	15	27	25	-2
Content	Cultural Context	28	13	24	27	3
Processing	Processing	132	59	7	8	1
Processing	Engineering	0	0	45	44	-1
Processing	Interface	70	38	12	12	0
Processing	Cognitive	57	32	14	14	0
Processing	Affective	8	6	38	34	-4
Processing	Situational	19	11	30	29	-1
Processing	Social Context	3	2	41	41	0
Processing	Cultural Context	1	1	43	42	-1
Engineering	Engineering	25	15	26	26	0
Engineering	Interface	11	7	34	33	-1
Engineering	Cognitive	4	3	39	39	0
Engineering	Affective	0	0	44	45	1
Engineering	Situational	8	6	37	35	-2
Engineering	Social Context	3	3	40	38	-2
Engineering	Cultural Context	2	1	42	43	1
Interface	Interface	430	118	1	2	1
Interface	Cognitive	147	70	6	5	-1
Interface	Affective	27	20	25	22	-3
Interface	Situational	56	31	15	15	0
Interface	Social Context	37	18	21	24	3
Interface	Cultural Context	18	5	31	37	6
Cognitive	Cognitive	370	135	2	1	-1
Cognitive	Affective	48	25	18	19	1
Cognitive	Situational	84	45	11	11	0
Cognitive	Social Context	40	26	20	17	-3

Table B-1: Stratum co-occurrence (distribution of findings & studies), by stratum pairing.

Stratum A	Stratum B	Num. Findings	Num. Studies	Findings Ranking	Studies Ranking	Rank Diff.
Cognitive	Cultural Context	21	12	28	28	0
Affective	Affective	107	56	10	9	-1
Affective	Situational	40	26	19	18	-1
Affective	Social Context	9	6	36	36	0
Affective	Cultural Context	12	3	33	40	7
Situational	Situational	213	76	4	4	0
Situational	Social Context	33	21	22	20	-2
Situational	Cultural Context	19	8	29	32	3
Social Context	Social Context	119	49	8	10	2
Social Context	Cultural Context	16	9	32	30	-2
Cultural Context	Cultural Context	66	21	13	21	8

Table B-2: Stratum co-occurrence by rank difference: $|Studies\ Ranking - Findings\ Ranking|$.

Stratum A	Stratum B	Num. Findings	Num. Studies	Findings Ranking	Studies Ranking	Rank Diff.
Cultural Context	Cultural Context	66	21	13	21	8
Affective	Cultural Context	12	3	33	40	7
Interface	Cultural Context	18	5	31	37	6
Processing	Affective	8	6	38	34	-4
Content	Engineering	11	8	35	31	-4
Interface	Affective	27	20	25	22	-3
Cognitive	Social Context	40	26	20	17	-3
Content	Processing	53	33	16	13	-3
Situational	Cultural Context	19	8	29	32	3
Content	Cultural Context	28	13	24	27	3
Interface	Social Context	37	18	21	24	3
Engineering	Social Context	3	3	40	38	-2
Engineering	Situational	8	6	37	35	-2
Social Context	Cultural Context	16	9	32	30	-2
Content	Social Context	25	15	27	25	-2
Situational	Social Context	33	21	22	20	-2
Content	Cognitive	113	64	9	7	-2
Social Context	Social Context	119	49	8	10	2
Processing	Engineering	0	0	45	44	-1
Processing	Cultural Context	1	1	43	42	-1
Engineering	Interface	11	7	34	33	-1
Processing	Situational	19	11	30	29	-1
Affective	Situational	40	26	19	18	-1
Content	Situational	52	30	17	16	-1
Affective	Affective	107	56	10	9	-1
Interface	Cognitive	147	70	6	5	-1
Cognitive	Cognitive	370	135	2	1	-1
Engineering	Affective	0	0	44	45	1
Engineering	Cultural Context	2	1	42	43	1
Cognitive	Affective	48	25	18	19	1
Processing	Processing	132	59	7	8	1
Content	Interface	179	65	5	6	1
Interface	Interface	430	118	1	2	1
Processing	Social Context	3	2	41	41	0
Engineering	Cognitive	4	3	39	39	0
Affective	Social Context	9	6	36	36	0
Cognitive	Cultural Context	21	12	28	28	0
Engineering	Engineering	25	15	26	26	0
Content	Affective	31	19	23	23	0
Interface	Situational	56	31	15	15	0
Processing	Cognitive	57	32	14	14	0
Processing	Interface	70	38	12	12	0
Cognitive	Situational	84	45	11	11	0
Situational	Situational	213	76	4	4	0
Content	Content	352	109	3	3	0

Appendix C: List of articles (Chapter 2)

The following list of all articles subject to the systematic analysis presented in Chapter 2 was obtained from the "List of Resources on User Aspects in MIR" available at http://www.jinhalee.com/miruserstudies/ (retrieved on December 7th, 2015). We assigned numeric article IDs (sequentially according to alphabetic order), and use these to refer to specific articles in the synthesis of findings.

1. Adamczyk, P. D. (2004). Seeing sounds: exploring musical social networks. Paper presented at the 12th annual ACM international conference on Multimedia, New York, NY, USA.
2. Allen, M., Gluck, J., Maclean, K., & Tang, E. (2005). An initial usability assessment for symbolic haptic rendering of music parameters. Paper presented at the 7th international conference on Multimodal interfaces (ICMI '05).
3. Andric, A., & Haus, G. (2006). Automatic playlist generation based on tracking user's listening habits. Multimedia Tools and Applications, 29(2), 127-151.
4. Andric, A., Xech, P.-L., & Fantasia, A. (2006). Music Mood Wheel: Improving Browsing Experience on Digital Content through an Audio Interface. Paper presented at the Automated Production of Cross Media Content for Multi-Channel Distribution, International Conference on, Los Alamitos, CA, USA.
5. Ankolekar, A., & Sandholm, T. (2011). Foxtrot: a soundtrack for where you are. Paper presented at the Interacting with Sound Workshop: Exploring Context-Aware, Local and Social Audio Applications, New York, NY, USA.
6. Arhippainen, L., & Hickey, S. (2011). Classifying music user groups and identifying needs for mobile virtual music services. Paper presented at the 15th International Academic MindTrek Conference: Envisioning Future Media Environments, New York, NY, USA.
7. Bainbridge, D., Cunningham, S. J., & Downie, J. S. (2003). How people describe their music information needs: A grounded theory analysis of music queries. Paper presented at the 4th International Conference on Music Information Retrieval (ISMIR 2003), Baltimore, MD.
8. Bainbridge, D., Novak, B. J., & Cunningham, S. J. (2010). A user-centered design of a personal digital library for music exploration. Paper presented at the 10th annual joint conference on Digital libraries, Gold Coast, Queensland, Australia.

9. Barrett, S., & Inskip, C. (2008). Deriving context from users' evaluations to inform software development. Paper presented at the 2nd International Symposium on Information Interaction in Context, London, United Kingdom.
10. Barrington, L., Reid, O., & Lanckriet, G. (2009). Smarter Than Genius? Human Evaluation of Music Recommender Systems. Paper presented at the 10th International Society for Music Information Retrieval Conference, Kobe, Japan.
11. Baumann, S., & Klüter, A. (2002). Super Convenience for Non-Musicians: Querying MP3 and the Semantic Web. Paper presented at the 3rd International Conference on Music Information Retrieval (ISMIR 2002), Paris, France.
12. Baur, D., Steinmayr, B., & Butz, A. (2010). SongWords: Exploring Music Collections Through Lyrics. Paper presented at the 11th International Society for Music Information Retrieval Conference (ISMIR 2010), Utrecht, Netherlands.
13. Bentley, F., Metcalf, C., & Harboe, G. (2006). Personal vs. commercial content: the similarities between consumer use of photos and music. Paper presented at the SIGCHI conference on Human Factors in computing systems (CHI'06).
14. Berenzweig, A., Logan, B., Ellis, D. P. W., & Whitman, B. (2004). A Large-Scale Evaluation of Acoustic and Subjective Music-Similarity Measures. Computer Music Journal, 28(2), 63-76.
15. Bergman, J., Kauko, J., & Keränen, J. (2009). Hands on music: physical approach to interaction with digital music. Paper presented at the 11th International Conference on Human-Computer Interaction with Mobile Devices and Services, New York, NY, USA.
16. Bischoff, K., Firan, C. S., Nejdl, W., & Paiu, R. (2009). How do you feel about "dancing queen"?: deriving mood & theme annotations from user tags. Paper presented at the 9th ACM/IEEE-CS joint conference on Digital libraries Austin, TX, USA.
17. Boltz, M., Schulkind, M., & Kantra, S. (1991). Effects of Background Music on the Remembering of Filmed Events. Memory & Cognition, 19(6), 593-606.
18. Bonardi, A. (2000). IR for contemporary music: What the musicologist needs. Paper presented at the 1st Annual International Symposium on Music Information Retrieval (ISMIR 2000), Plymouth, MA.
19. Braunhofer, M., Kaminskas, M., & Ricci, F. (2011). Recommending music for places of interest in a mobile travel guide. Paper presented at the 5th ACM conference on Recommender systems, New York, NY, USA.
20. Brinegar, J., & Capra, R. (2011). Managing music across multiple devices and computers. Paper presented at the 2011 iConference, Seattle, WA, USA.
21. Carlisle, J. (2007). Digital music and generation Y: discourse analysis of the online music information behaviour talk of five young Australians. Information Research, 12(4).

22. Chen, Y.-X., & Butz, A. (2009). Musicsim: integrating audio analysis and user feedback in an interactive music browsing ui. Paper presented at the 14th International Conference on Intelligent user interfaces Sanibel Island, FL, USA.

23. Cunningham, S., & Bainbridge, D. (2010). A Search Engine Log Analysis of Music-Related Web Searching. In N. Nguyen, R. Katarzyniak & S.-M. Chen (Eds.), Advances in Intelligent Information and Database Systems (Vol. 283, pp. 79-88): Springer Berlin / Heidelberg.

24. Cunningham, S. J. (2002). What People Do When They Look for Music: Implications for Design of a Music Digital Library. Lecture notes in computer science(2555), 177-178.

25. Cunningham, S. J. (2003). User studies: A first step in designing an MIR testbed. The MIR/MDL Evaluation Project White Paper Collection Edition #3, 17–19.

26. Cunningham, S. J., Bainbridge, D., & Falconer, A. (2006). "More of an art than a science": Supporting the creation of playlists and mixes. Paper presented at the 7th International Conference on Music Information Retrieval (ISMIR 2006), Victoria, Canada.

27. Cunningham, S. J., Bainbridge, D., & McKay, D. (2007). Finding new music: a diary study of everyday encounter with novel songs. Paper presented at the 8th International Conference on Music Information Retrieval (ISMIR 2007), Vienna, Austria.

28. Cunningham, S. J., Jones, M., & Jones, S. (2004). Organizing digital music for use: an examination of personal music collections. Paper presented at the 5th International Conference on Music Information Retrieval (ISMIR 2004), Barcelona, Spain.

29. Cunningham, S. J., & Nichols, D. M. (2009). Exploring Social Music Behavior: an Investigation of Music Selection at Parties. Paper presented at the 10th International Society for Music Information Retrieval Conference (ISMIR 2009), Kobe, Japan.

30. Cunningham, S. J., Reeves, N., & Britland, M. (2003). An ethnographic study of music information seeking: implications for the design of a music digital library. Paper presented at the 3rd ACM/IEEE-CS joint conference on Digital libraries, Washington, DC, USA.

31. Cunningham, S. J., & Zhang, Y. E. (2008). Development of a music organizer for children. Paper presented at the 9th International Conference on Music Information Retrieval (ISMIR 2008), Philadelphia, PA.

32. Dachselt, R., & Frisch, M. (2007). Mambo: a facet-based zoomable music browser. Paper presented at the 6th international conference on Mobile and ubiquitous multimedia, New York, NY, USA.

33. Dias, R., & Fonseca, M. J. (2010). MuVis: an application for interactive exploration of large music collections. Paper presented at the international conference on Multimedia, New York, NY, USA.

34. Downie, J. S. (1994). The MusiFind musical information retrieval project, Phase II: User assessment survey. Paper presented at the 22nd Annual Conference of the Canadian Association for Information Science, Montreal, Quebec.

35. Downie, J. S., & Cunningham, S. J. (2002). Toward a theory of music information retrieval queries: System design implications. Science And Technology, 18, 29.

36. Duggan, B., & O'Shea, B. (2010). Tunepal-Disseminating a Music Information Retrieval System to the Traditional Irish Music Community. Paper presented at the 11th International Society for Music Information Retrieval Conference (ISMIR 2010), Utrecht, Netherlands.

37. Ellis, D. P. W., & Whitman, B. (2002). The quest for ground truth in musical artist similarity. Paper presented at the 3rd International Conference on Music Information Retrieval (ISMIR 2002), Paris, France.

38. Fernström, M., & Ó Maidín, D. (2001). Computer-supported Browsing for MIR. Paper presented at the 2nd Annual International Symposium on Music Information Retrieval (ISMIR 2001), Bloomington, IN.

39. Fink, E. L., Robinson, J. P., & Dowden, S. (1985). The Structure of Music Preference and Attendance. Communication Research, 12(3), 301-318.

40. Garg, R., Smith, M. D., & Telang, R. (2011). Discovery of Music through Peers in an Online Community. Paper presented at the 2011 44th Hawaii International Conference on System Sciences, Washington, DC, USA.

41. Gatewood, E. (1921). An Experiment in the Use of Music in An Architectural Drafting Room. Journal of Applied Psychology, 5(4), 350-358.

42. Goel, S., Broder, A., Gabrilovich, E., & Pang, B. (2010). Anatomy of the long tail: ordinary people with extraordinary tastes. Paper presented at the third ACM international conference on Web search and data mining, New York, NY, USA.

43. Greasley, A. E., & Lamont, A. (2011). Exploring engagement with music in everyday life using experience sampling methodology. Musicae Scientiae, 15(1), 45-71.

44. Hakansson, M., Rost, M., Jacobsson, M., & Holmquist, L. E. (2007). Facilitating Mobile Music Sharing and Social Interaction with Push!Music. Paper presented at the 40th Annual Hawaii International Conference on System Sciences, Washington, DC, USA.

45. Hijikata, Y., Iwahama, K., & Nishida, S. (2006). Content-based music filtering system with editable user profile. Paper presented at the 2006 ACM symposium on Applied computing, New York, NY, USA.

46. Hoashi, K., Hamawaki, S., Ishizaki, H., Takishima, Y., & Katto, J. (2009). Usability Evaluation of Visualization Interfaces for Content-Based Music Retrieval Systems. Paper presented at the 10th International Society for Music Information Retrieval Conference (ISMIR 2009), Kobe, Japan.

47. Holm, J., Aaltonen, A., & Seppänen, J. (2009). Associating fonts with musical genres. Paper presented at the 6th International Conference on Computer Graphics, Virtual Reality, Visualisation and Interaction in Africa, New York, NY, USA.

48. Holm, J., Aaltonen, A., & Siirtola, H. (2009). Associating Colours with Musical Genres. Journal of New Music Research, 38(1), 87-100.

49. Holm, J., Holm, H., & Seppänen, J. (2010). Associating emoticons with musical genres. Paper presented at the Proceedings of International Conference on New Interfaces for Musical Expression, Sydney, Australia.

50. Holm, J., Lehtiniemi, A., & Eronen, A. (2010). Evaluating an avatar-based user interface for discovering new music. Paper presented at the 9th International Conference on Mobile and Ubiquitous Multimedia (MUM'10), Limassol, Cyprus.

51. Holm, J., Siirtola, H., & Laaksonen, L. (2010). Associating Avatars with Musical Genres. Paper presented at the 2010 14th International Conference Information Visualisation, Washington, DC, USA.

52. Hu, X., Downie, J. S., & Ehmann, A. F. (2006). Exploiting recommended usage metadata: Exploratory analyses. Paper presented at the 7th International Conference on Music Information Retrieval (ISMIR 2006), Victoria, Canada.

53. Hu, X. and Lee, J. H. (2012). A cross-cultural study of music mood perception between American and Chinese listeners. Proceedings of the 13th International Conference on Music Information Retrieval: ISMIR 2012, Porto, Portugal.

54. Huron, D., & Aarden, B. (2002). Cognitive Issues and Approaches in Music Information Retrieval. Retrieved from http://musicog.ohio-state.edu/Huron/Publications/huron.aarden.MIR.html

55. Inskip, C., Butterworth, R., & MacFarlane, A. (2008). A study of the information needs of the users of a folk music library and the implications for the design of a digital library system. Information Processing & Management, 44(2), 647–662.

56. Inskip, C., MacFarlane, A., & Rafferty, P. (2008). Content or context?: searching for musical meaning in task-based interactive information retrieval. Paper presented at the 2nd International Symposium on Information Interaction in Context, London, United Kingdom.

57. Inskip, C., MacFarlane, A., & Rafferty, P. (2008). Music, Movies and Meaning: Communication in Film-makers' Search for Pre-existing Music, and the Implications for Music Information Retrieval. Paper presented at the 9th International Conference on Music Information Retrieval (ISMIR 2008), Philadelphia, PA.

58. Inskip, C., MacFarlane, A., & Rafferty, P. (2010). Creative professional users' musical relevance criteria. Journal of Information Science, 36(4), 517-529.

59. Inskip, C., MacFarlane, A., & Rafferty, P. (2010). Organising music for movies. Aslib Proceedings, 62(4/5), 489-501.

60. Itoh, M. (2000). Subject search for music: quantitative analysis of access point selection. Paper presented at the 1st Annual International Symposium on Music Information Retrieval (ISMIR 2000), Plymouth, MA.

61. Kaji, K., Hirata, K., & Nagao, K. (2005). A Music Recommendation System Based on Annotations about Listeners' Preferences and Situations. Paper presented at the First International Conference on Automated Production of Cross Media Content for Multi-Channel Distribution, Washington, DC, USA.

62. Kaminskas, M. (2009). Matching information content with music. Paper presented at the 3rd ACM conference on Recommender systems (RecSys '09), New York, New York, USA.

63. Kampfe, J., Sedlmeier, P., & Renkewitz, F. (2010). The impact of background music on adult listeners: A meta-analysis. Psychology of Music, 39(4), 424-448.

64. Keränen, J., Bergman, J., & Kauko, J. (2009). Gravity sphere: gestural audio-tactile interface for mobile music exploration. Paper presented at the 27th international conference on Human factors in computing systems, New York, NY, USA.

65. Kibby, M. (2009). Collect Yourself. Information, Communication & Society, 12(3), 428-443.

66. Kim, J.-H., Jung, K.-Y., Ryu, J.-K., Kang, U.-G., & Lee, J.-H. (2008). Design of Ubiquitous Music Recommendation System Using MHMM. Paper presented at the 2008 Fourth International Conference on Networked Computing and Advanced Information Management – Volume 02, Washington, DC, USA.

67. Kim, J.-Y., & Belkin, N. J. (2002). Categories of Music Description and Search Terms and Phrases Used by Non-Music Experts. Paper presented at the 3rd International Conference on Music Information Retrieval (ISMIR 2002), Paris, France.

68. Kinnally, W., Lacayo, A., McClung, S., & Sapolsky, B. (2008). Getting up on the download: college students' motivations for acquiring music via the web. New Media & Society, 10(6), 893-913.

69. Koenigstein, N., Shavitt, Y., Weinsberg, E., & Weinsberg, U. (2010). On the applicability of peer-to-peer data in music information retrieval research. Paper presented at the 11th International Society for Music Information Retrieval Conference (ISMIR 2010), Utrecht, Netherlands.

70. Kolhoff, P., Preuss, J., & Loviscach, J. (2006). Music Icons: Procedural Glyphs for Audio Files. Paper presented at the Computer Graphics and Image Processing, 2006. SIBGRAPI '06. 19th Brazilian Symposium on.

71. Koskela, T., Järvinen, S., Liu, M., & Ylianttila, M. (2010). User Experience in Added Value Location-Based Mobile Music Service. Paper presented at the 2010 IEEE International Conference on Web Services, Washington, DC, USA.

72. Koskela, T., Julkunen, J., Keranen, V., Kostamo, N., & Ylianttila, M. (2009). User Experiences on a Community-Based Music Voting Service. Paper presented at the 2009 Congress on Services – I, Washington, DC, USA.

73. Krause, A. E., & Hargreaves, D., J. (2012). myTunes: Digital Music Library Users and Their Self-Images. Psychology of Music.

74. Kuhn, M., Wattenhofer, R., & Welten, S. (2010). Social audio features for advanced music retrieval interfaces. Paper presented at the International conference on Multimedia (MM'10), Firenze, Italy.

75. Kukka, H., Patino, R., & Ojala, T. (2009). UbiRockMachine: a multimodal music voting service for shared urban spaces. Paper presented at the 8th International Conference on Mobile and Ubiquitous Multimedia, New York, NY, USA.

76. Kuo, F.-F., & Shan, M.-K. (2002). A Personalized Music Filtering System Based on Melody Style Classification. Paper presented at the 2002 IEEE International Conference on Data Mining, Washington, DC, USA.

77. Kuo, F.-F., & Shan, M.-K. (2004). Looking for new, not known music only: music retrieval by melody style. Paper presented at the 4th ACM/IEEE-CS joint conference on Digital libraries, New York, NY, USA.

78. Kusama, K., & Itoh, T. (2011). MusCat: a music browser featuring abstract pictures and zooming user interface. Paper presented at the 2011 ACM Symposium on Applied Computing, New York, NY, USA.

79. Lai, K., & Chan, K. (2010). Do You Know Your Music Users' Needs? A Library User Survey that Helps Enhance a User-Centered Music Collection. Journal of Academic Librarianship, 36(1), 63-69.

80. Lamont, A. (2011). University students' strong experiences of music: Pleasure, engagement, and meaning. Musicae Scientiae, 15(2), 229-249.

81. Lang, A., & Masoodian, M. (2007). Graphic designers' quest for the right music. Paper presented at the 7th ACM SIGCHI New Zealand chapter's international conference on Computer-human interaction: design centered HCI, Hamilton, New Zealand.

82. Laplante, A. (2010). Users' Relevance Criteria in Music Retrieval in Everyday Life: An Exploratory Study. Paper presented at the 11th International Society for Music Information Retrieval Conference (ISMIR 2010), Utrecht, Netherlands.

83. Laplante, A., & Downie, J. S. (2006). Everyday Life Music Information-Seeking Behaviour of Young Adults. Paper presented at the 7th International Conference on Music Information Retrieval (ISMIR 2006), Victoria, Canada.

84. Laplante, A., & Downie, J. S. (2011). The utilitarian and hedonic outcomes of music information-seeking in everyday life. Library & Information Science Research, 33(3), 202-210.

85. Laube, V., Moewes, C., & Stober, S. (2008). Browsing music by usage context. Paper presented at the 2nd Workshop on Learning the Semantics of Audio Signals (LSAS 2008), Paris, France.

86. Lee, J. H. (2008). Analysis of the accuracy of user-provided information in natural language queries for music information retrieval. Journal of the Korean Society for Information Management, 25(4), 149-164.

87. Lee, J. H. (2010). Analysis of user needs and information features in natural language queries seeking music information. Journal of the American Society for Information Science and Technology, 61(5), 1025-1045. doi: http://doi.wiley.com/10.1002/asi.21302

88. Lee, J. H. (2010). Crowdsourcing music similarity judgments using Mechanical Turk. Proceedings of the 11th International Conference on Music Information Retrieval: ISMIR 2010, Utrecht, Netherlands, 183-188.

89. Lee, J. H. (2011). How Similar Is Too Similar?: Exploring Users' Perceptions of Similarity in Playlist Evaluation. Paper presented at the 12th International Society for Music Information Retrieval Conference (ISMIR 2012), Miami, FL, USA.

90. Lee, J. H. and Cunningham, S. J. (2012). The impact (or non-impact) of user studies in music information retrieval. Proceedings of the 13th International Conference on Music Information Retrieval: ISMIR 2012, Porto, Portugal.

91. Lee, J. H., & Downie, J. S. (2004). Survey of music information needs, uses, and seeking behaviours: Preliminary findings. Paper presented at the 5th International Conference on Music Information Retrieval (ISMIR 2004), Barcelona, Spain.

92. Lee, J. H., Downie, J. S., & Cunningham, S. J. (2005). Challenges in Cross-Cultural/Multilingual Music Information Seeking. Paper presented at the 6th International Conference on Music Information Retrieval (ISMIR 2005), London, UK.

93. Lee, J. H., Downie, J. S., & Jones, M. C. (2007). Preliminary analyses of information features provided by users for identifying music. Paper presented at the 8th International Conference on Music Information Retrieval (ISMIR 2007), Vienna, Austria.

94. Lee, J. H., Hill, T., & Work, L. (2011). What does music mood mean for real users? Paper presented at the 2012 iConference, Toronto, Canada.

95. Lee, J. H., Jones. M. C., and Downie, J. S. (2006). Factors affecting the response rates of real-life MIR queries. Proceedings of the 7th International Society for Music Information Retrieval Conference: ISMIR 2006, Victoria, Canada, 371-372.

96. Lee, J. H., & Waterman, N. M. (2012). Understanding user requirements for music information services. Proceedings of the 13th International Conference on Music Information Retrieval: ISMIR 2012, Porto, Portugal.

97. Lee, J. H., & Hu, X. (2012). Generating ground truth for music mood classification using Mechanical Turk. ACM/IEEE-CS Joint Conference on Digital Libraries.

98. Lehtiniemi, A., & Holm, J. (2011). Easy access to recommendation playlists: selecting music by exploring preview clips in album cover space. Paper presented at the 10th International Conference on Mobile and Ubiquitous Multimedia, New York, NY, USA.

99. Lehtiniemi, A., & Holm, J. (2011). Evaluating a Cube-Based User Interface for Exploring Music Collections. Paper presented at the Information Visualisation (IV), 2011 15th International Conference on.

100. Lehtiniemi, A., & Holm, J. (2011). Evaluating a Potentiometer-Based Graphical User Interface for Interacting with a Music Recommendation Service. Paper presented at the 2011 15th International Conference on Information Visualisation, Washington, DC, USA.

101. Lehtiniemi, A., & Seppänen, J. (2007). Evaluation of automatic mobile playlist generator. Paper presented at the 4th international conference on mobile technology, applications, and systems and the 1st international symposium on Computer human interaction in mobile technology, New York, NY, USA.

102. Leong, T., Howard, S., & Vetere, F. (2008). Choice: abidcating or exercising? Paper presented at the twenty-sixth annual SIGCHI conference on Human factors in computing systems, New York, NY, USA.

103. Leong, T. W., Vetere, F., & Howard, S. (2005). The serendipity shuffle. Paper presented at the 17th Australia conference on Computer-Human Interaction: Citizens Online: Considerations for Today and the Future, Narrabundah, Australia, Australia.

104. Leong, T. W., Vetere, F., & Howard, S. (2012). Experiencing coincidence during digital music listening. ACM Trans. Comput.-Hum. Interact., 19(1), 6:1–6:19.

105. Lesaffre, M., Moelants, D., Leman, M., Hewlett, W., & Selfridge-Field, E. (2004). Spontaneous user behavior in "vocal" queries for music-information retrieval Computing in musicology (pp. 129-146). Cambridge, Mass.: MIT Press.

106. Lesaffre, M., Voogdt, L. D., Leman, M., Baets, B. D., Meyer, H. D., & Martens, J.-P. (2008). How potential users of music search and retrieval systems describe the semantic quality of music. Journal of the American Society for Information Science and Technology, 59(5), 695-707.

107. Levitin, D. J. (1994). Absolute memory for musical pitch: evidence from the production of learned melodies. Perception & psychophysics, 56(4), 414-423.

108. Liu, H., Hu, J., & Rauterberg, M. (2009). Music Playlist Recommendation Based on User Heartbeat and Music Preference. Paper presented at the 2009 International Conference on Computer Technology and Development – Volume 01, Washington, DC, USA.

109. Lonsdale, A. J., & North, A. C. (2011). Why do we listen to music? A uses and gratifications analysis. British Journal of Psychology, 102(1), 108-134.

110. Lu, C.-C., & Tseng, V. S. (2009). A novel method for personalized music recommendation. Expert Systems with Applications, 36(6), 10035-10044.

111. Magaudda, P. (2011). When materiality 'bites back': Digital music consumption practices in the age of dematerialization. Journal of Consumer Culture, 11(1), 15-36.

112. Maguire, M., Motson, D. E., Wilson, G., & Wolfe, J. (2004). Searching for Nirvana: Cataloging and the Digital Collection at the Experience Music Project. Journal of Internet Cataloging, 7(1), 9-31.

113. McNab, R. J., Smith, L. A., Witten, I. H., Henderson, C. L., & Cunningham, S. J. (1996). Towards the Digital Music Library: Tune Retrieval From Acoustic Input. Paper presented at the 1st ACM international conference on Digital libraries.

114. McPherson, J. R., & Bainbridge, D. (2001). Usage of the MELDEX Digital Music Library. Paper presented at the 2nd Annual International Symposium on Music Information Retrieval (ISMIR 2001), Bloomington, IN.

115. Meintanis, K., & Shipman, F. M. (2010). Visual expression for organizing and accessing music collections in MusicWiz. Paper presented at the 14th European conference on Research and advanced technology for digital libraries (ECDL'10), Glasgow, UK.

116. Nettamo, E., Nirhamo, M., & Häkkilä, J. (2006). A cross-cultural study of mobile music: retrieval, management and consumption. Paper presented at the 8th Australia conference on Computer-Human Interaction: Design: Activities, Artefacts and Environments, Sydney, Australia.

117. North, A. C., Hargreaves, D. J., & Jon, J. H. (2004). Uses of Music in Everyday Life. Music Perception: An Interdisciplinary Journal, 22(1), 41-77.

118. North, A. C., Hargreaves, D. J., & McKendrick, J. (2006). The Effects of Music on Atmosphere in a Bank and a Bar. Journal of Applied Social Psychology, 30(7), 1504-1522.

119. North, A. C., Hargreaves, D. J., & O'Neill, S. A. (2010). The importance of music to adolescents. British Journal of Educational Psychology, 70(2), 255-272.

120. Omojokun, O., Genovese, M., & Isbell, J. C. (2008). Impact of user context on song selection. Paper presented at the 16th ACM international conference on Multimedia, New York, NY, USA.

121. Oudenne, A. M., Kim, Y. E., & Turnbull, D. S. (2010). Meerkat: exploring semantic music discovery using personalized radio. Paper presented at the International conference on Multimedia information retrieval (MIR '10), Philadelphia, PA, USA.

122. Pauws, S. (2003). Effects of song familiarity, singing training and recent song exposure on the singing of melodies. Paper presented at the 4th International Conference on Music Information Retrieval (ISMIR 2003), Baltimore, MD.

123. Pauws, S., & Eggen, B. (2003). Realization and User Evaluation of an Automatic Playlist Generator. Journal of New Music Research, 32(2), 179-192.

124. Pauws, S., & van de Wijdeven, S. (2005). User Evaluation of a New Interactive Playlist Generation Concept. Paper presented at the 6th International Conference on Music Information Retrieval (ISMIR 2005), London, UK.

125. Pennycook, B. W. (1985). Computer-music interfaces: a survey. ACM Computing Surveys, 17(2), 267-289.

126. Rana, S. A., & North, A. C. (2007). The Role of Music in Everyday Life Among Pakistanis. Music Perception: An Interdisciplinary Journal, 25(1), 59-73.

127. Saito, Y., & Itoh, T. (2011). MusiCube: a visual music recommendation system featuring interactive evolutionary computing. Paper presented at the 2011 Visual Information Communication – International Symposium, New York, NY, USA.

128. Schäfer, T., & Sedlmeier, P. (2010). What makes us like music? Determinants of music preference. Psychology of Aesthetics Creativity and the Arts, 4(4), 223-234.

129. Schlitter, N., & Falkowski, T. (2009). Mining the Dynamics of Music Preferences from a Social Networking Site. Paper presented at the 2009 International Conference on Advances in Social Network Analysis and Mining, Washington, DC, USA.

130. Schuller, B., Zobl, M., Rigoll, G., & Lang, M. (2003). A hybrid music retrieval system using belief networks to integrate multimodal queries and contextual knowledge. Paper presented at the 2003 International Conference on Multimedia and Expo – Volume 2, Washington, DC, USA.

131. Sease, R., & McDonald, D. W. (2009). Musical fingerprints: collaboration around home media collections. Paper presented at the ACM 2009 international conference on Supporting group work (GROUP'09), Sanibel Island, FL, USA.

132. Secord, A., Winnemoeller, H., Li, W., & Dontcheva, M. (2010). Creating collections with automatic suggestions and example-based refinement. Paper presented at the 23nd annual ACM symposium on User interface software and technology, New York, NY, USA.

133. Selfridge-field, E. (2000). What Motivates a Musical Query? Paper presented at the 1st Annual International Symposium on Music Information Retrieval (ISMIR 2000), Plymouth, MA.

134. Shao, B., Wang, D. D., Li, T., & Ogihara, M. (2009). Music Recommendation Based on Acoustic Features and User Access Patterns. IEEE Transactions on Audio, Speech, and Language Processing, 17(8), 1602-1611.

135. Shiroi, S., Misue, K., & Tanaka, J. (2011). A Tool to Support Finding Favorite Music by Visualizing Listeners' Preferences. Paper presented at the 2011 15th International Conference on Information Visualisation, Washington, DC, USA.

136. Silfverberg, S., Liikkanen, L. A., & Lampinen, A. (2011). "I'll press play, but I won't listen": profile work in a music-focused social network service. Paper presented at the ACM 2011 conference on Computer supported cooperative work, New York, NY, USA.

137. Snyder, T. (2010). Music Materials in a Faceted Catalog: Interviews with Faculty and Graduate Students. Music Reference Services Quarterly, 13(3/4), 66-95.

138. Sordo, M., Celma, O., Blech, M., & Guaus, E. (2008). The quest for musical genres: Do the experts and the wisdom of crowds agree? Paper presented at the 9th International Conference on Music Information Retrieval (ISMIR 2008), Philadelphia, PA.

139. Stober, S., Steinbrecher, M., & Nürnberger, A. (2009). A Survey on the Acceptance of Listening Context Logging for MIR Applications. Paper presented at the 3rd International Workshop on Learning the Semantics of Audio Signals (LSAS), Graz, Austria.

140. Stumpf, S., & Muscroft, S. (2011). When users generate music playlists: When words leave off, music begins? Paper presented at the Multimedia and Expo, IEEE International Conference on, Los Alamitos, CA, USA.

141. Taheri-Panah, S., & MacFarlane, A. (2004). Music Information Retrieval systems: why do individuals use them and what are their needs? Paper presented at the 5th International Conference on Music Information Retrieval (ISMIR 2004), Barcelona, Spain.

142. Tarulli, L. (2010). Exploring Public Library Music Collections Through Social Technologies. Fontes Artis Musicae, 57(3), 267-274.

143. Ter Bogt, T. F. M., Mulder, J., Raaijmakers, Q. A. W., & Nic Gabhainn, S. (2010). Moved by music: A typology of music listeners. Psychology of Music, 39(2), 147-163.

144. Uitdenbogerd, A., & Schyndel, R. (2002). A review of factors affecting music recommender success. Paper presented at the 3rd International Conference on Music Information Retrieval (ISMIR 2002), Paris, France.

145. Uitdenbogerd, A. L., & Yap, Y. W. (2003). Was Parsons right? An experiment in usability of music representations for melody-based music retrieval. Paper presented at the 4th International Conference on Music Information Retrieval (ISMIR 2003), Baltimore, MD.

146. Vignoli, F. (2004). Digital Music Interaction concepts: a user study. Paper presented at the 5th International Conference on Music Information Retrieval (ISMIR 2004), Barcelona, Spain.

147. Vignoli, F., & Pauws, S. (2005). A music retrieval system based on user-driven similarity and its evaluation. Paper presented at the 6th International Conference on Music Information Retrieval (ISMIR 2005), London, UK.

148. Voida, A., Grinter, R. E., Ducheneaut, N., Edwards, W. K., & Newman, M. W. (2005). Listening In: Practices Surrounding iTunes Music Sharing. Paper presented at the SIGCHI conference on Human factors in computing systems (CHI'05).

149. Voong, M., & Beale, R. (2007). Music organisation using colour synaesthesia. Paper presented at the CHI '07 extended abstracts on Human factors in computing systems, New York, NY, USA.

150. Vuoskoski, J. K., Thompson, W. F., McIlwain, D., & Eerola, T. (2012). Who Enjoys Listening to Sad Music and Why? Music Perception: An Interdisciplinary Journal, 29(3), 311-317.

151. Weigl, D. M., & Guastavino, C. (2012). User studies in the Music Information Retrieval literature. Paper presented at the 12th International Society for Music Information Retrieval Conference (ISMIR 2012), Miami, FL, USA.

152. Wiering, F., de Nooijer, J., Volk, A., & Tabachneck-Schijf, H. J. M. (2009). Cognition-based Segmentation for Music Information Retrieval Systems. Journal of New Music Research, 38(2), 139-154.

153. Williams, C. (2001). Does It Really Matter? Young People and Popular Music. Popular Music, 20(2), 223-242.

154. Woelfer, J. and Lee, J. H. (2012). The role of music in the lives of homeless young people: A preliminary report. Proceedings of the 13th International Conference on Music Information Retrieval: ISMIR 2012, Porto, Portugal.

155. Yang, Y.-H., Su, Y.-F., Lin, Y.-C., & Chen, H. H. (2007). Music emotion recognition: the role of individuality. Paper presented at the international workshop on Human-centered multimedia, New York, NY, USA.

156. Yi, Y., Zhou, Y., & Wang, Y. (2011). A tempo-sensitive music search engine with multimodal inputs. Paper presented at the 1st international ACM workshop on Music information retrieval with user-centered and multimodal strategies, New York, NY, USA.

157. Yoo, M.-J., & Lee, I.-K. (2011). Affecticon: Emotion-Based Icons for Music Retrieval. IEEE Computer Graphics & Applications, 31(3), 89-95.

158. Zheleva, E., Guiver, J., Rodrigues, E. M., & Milic-Frayling, N. (2010). Statistical models of music-listening sessions in social media. Paper presented at the 19th International World Wide Web Conference (WWW).

159. Zhu, J., & Lu, L. (2005). Perceptual Visualization of a Music Collection. Paper presented at the Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on.

Appendix D: Mean identification scores for melodies in the stimulus set (Chapter 3)

The following tables display the mean identification scores for melodies in the stimulus set (altered conditions). Table D-1 is ordered by melody name. Tables D-2 and D-3 are ordered by mean identification score in the isochronous and randomized conditions, respectively.

Values listed for a particular melody correspond to mean identification score across all participants who were familiar with this melody (i.e., participants able to identify the melody in its unaltered version). Individual responses were scored 1 for correct identifications, .5 for partial identifications, and 0 for misidentifications or non-identifications. Two melodies, marked with asterisks (*) in the following tables, were erroneously presented in their reordered version in both the reordered and randomized conditions of Study 1. Responses to these melodies were accordingly withdrawn from the analyses of all conditions of both studies.

Table D-1: Mean identification scores by melody name.

	Mean identification score			
	Study 1		Study 2	
Melody name	Reordered	Randomized	Stretched	Isochronous
Addams Family Theme	.00	.00	.31	.23
Bingo	.41	.05	.42	.58
Blue Danube Waltz	.45	.36	.53	.47
Camp Town Racing	.69	.17	.17	.44
Coming 'round the Mountain*	.15	*	.74	.47
Deck the Halls	.26	.00	.08	.32
Doe, a Deer	.09	.00	.12	.22
For He's a Jolly Good Fellow	.00	.00	.00	.00
Frère Jacques	.91	.50	.57	.74
Happy Birthday	.94	.97	1.00	1.00
Hark the Herald Angels	.24	.21	.22	.62
Hey Jude	.82	.36	.56	.17
I Will Survive	.05	.00	.00	.00
If You're Happy and You Know It*	.75	*	.95	.59

Table D-1: Mean identification scores by melody name.

	Mean identification score			
	Study 1		Study 2	
Melody name	Reordered	Randomized	Stretched	Isochronous
Jingle Bells	.87	.19	.55	.73
London Bridge	.28	.13	.36	.61
Lone Ranger Theme (William Tell)	.65	.09	.20	.50
Mary Had a Little Lamb	.96	.29	.60	.95
Michael Row the Boat	.00	.00	.50	.50
My Bonnie	.71	.35	.81	.81
Ode To Joy	.40	.40	.64	1.00
O Canada	.92	.50	.60	.30
Oh Suzanna	.65	.08	.42	.58
Old MacDonald	.89	.37	.57	.86
Rock a Bye Baby	.39	.37	.63	.58
Row Row Row Your Boat	.14	.00	.15	.20
Rudolph the Red Nosed Reindeer	.79	.09	.19	.52
Russian Dance (Nutcracker)	.22	.00	.13	.00
Silent Night	.78	.92	.98	.19
The Itsy Bitsy Spider	.02	.00	.32	.00
The Song That Never Ends	.00	.00	.20	.20
Three Blind Mice	.41	.41	.81	.19
Twinkle Twinkle Little Star	.87	.63	.72	.96
We Wish You A Merry Christmas	.95	.17	.50	.66
Wedding March (Here Comes the Bride)	.77	.65	.68	.20
When The Saints Go Marching In	.59	.33	.88	.64
White Christmas	.21	.14	.72	.00
Yankee Doodle	.86	.69	.86	.86

Table D-2: Mean identification scores by performance in isochronous condition.

	Mean identification score			
	Study 1		Study 2	
Melody name	Reordered	Randomized	Stretched	Isochronous
Happy Birthday	.94	.97	1.00	1.00
Ode To Joy	.40	.40	.64	1.00
Twinkle Twinkle Little Star	.87	.63	.72	.96
Mary Had a Little Lamb	.96	.29	.60	.95
Old MacDonald	.89	.37	.57	.86
Yankee Doodle	.86	.69	.86	.86
My Bonnie	.71	.35	.81	.81
Frere Jacques	.91	.50	.57	.74
Jingle Bells	.87	.19	.55	.73
We Wish You A Merry Christmas	.95	.17	.50	.66
When The Saints Come Marching In	.59	.33	.88	.64
Hark the Herald Angels	.24	.21	.22	.62
London Bridge	.28	.13	.36	.61
If You're Happy and You Know It*	.75	*	.95	.59
Bingo	.41	.05	.42	.58
Oh Suzanna	.65	.08	.42	.58
Rock a Bye Baby	.39	.37	.63	.58
Rudolph the Red Nosed Reindeer	.79	.09	.19	.52
Lone Ranger Theme (William Tell)	.65	.09	.20	.50
Michael Row the Boat	.00	.00	.50	.50
Blue Danube Waltz	.45	.36	.53	.47
Coming 'round the Mountain*	.15	*	.74	.47
Camp Town Racing	.69	.17	.17	.44
Deck the Halls	.26	.00	.08	.32
Oh Canada	.92	.50	.60	.30
Addams Family Theme	.00	.00	.31	.23
Doe, a Deer	.09	.00	.12	.22
Row Row Row Your Boat	.14	.00	.15	.20
The Song That Never Ends	.00	.00	.20	.20
Wedding March (Here Comes the Bride)	.77	.65	.68	.20

Table D-2: Mean identification scores by performance in isochronous condition.

	Mean identification score			
	Study 1		Study 2	
Melody name	Reordered	Randomized	Stretched	Isochronous
Silent Night	.78	.92	.98	.19
Three Blind Mice	.41	.41	.81	.19
Hey Jude	.82	.36	.56	.17
For He's a Jolly Good Fellow	.00	.00	.00	.00
I Will Survive	.05	.00	.00	.00
Russian Dance (Nutcracker)	.22	.00	.13	.00
The Itsy Bitsy Spider	.02	.00	.32	.00
White Christmas	.21	.14	.72	.00

Table D-3: Mean identification scores by performance in randomized condition.

	Mean identification score			
	Study 1		Study 2	
Melody name	Reordered	Randomized	Stretched	Isochronous
Happy Birthday	.94	.97	1.00	1.00
Silent Night	.78	.92	.98	.19
Yankee Doodle	.86	.69	.86	.86
Wedding March (Here Comes the Bride)	.77	.65	.68	.20
Twinkle Twinkle Little Star	.87	.63	.72	.96
Frere Jacques	.91	.50	.57	.74
Oh Canada	.92	.50	.60	.30
Three Blind Mice	.41	.41	.81	.19
Ode To Joy	.40	.40	.64	1.00
Old MacDonald	.89	.37	.57	.86
Rock a Bye Baby	.39	.37	.63	.58
Blue Danube Waltz	.45	.36	.53	.47
Hey Jude	.82	.36	.56	.17
My Bonnie	.71	.35	.81	.81
When The Saints Come Marching In	.59	.33	.88	.64
Mary Had a Little Lamb	.96	.29	.60	.95
Hark the Herald Angels	.24	.21	.22	.62
Jingle Bells	.87	.19	.55	.73
We Wish You A Merry Christmas	.95	.17	.50	.66
Camp Town Racing	.69	.17	.17	.44
White Christmas	.21	.14	.72	.00
London Bridge	.28	.13	.36	.61
Rudolph the Red Nosed Reindeer	.79	.09	.19	.52
Lone Ranger Theme (William Tell)	.65	.09	.20	.50
Oh Suzanna	.65	.08	.42	.58
Bingo	.41	.05	.42	.58
Michael Row the Boat	.00	.00	.50	.50
Deck the Halls	.26	.00	.08	.32
Addams Family Theme	.00	.00	.31	.23
Doe, a Deer	.09	.00	.12	.22

Table D-3: Mean identification scores by performance in randomized condition.

	Mean identification score			
	Study 1		Study 2	
Melody name	Reordered	Randomized	Stretched	Isochronous
Row Row Row Your Boat	.14	.00	.15	.20
The Song That Never Ends	.00	.00	.20	.20
For He's a Jolly Good Fellow	.00	.00	.00	.00
I Will Survive	.05	.00	.00	.00
Russian Dance (Nutcracker)	.22	.00	.13	.00
The Itsy Bitsy Spider	.02	.00	.32	.00
If You're Happy and You Know It*	.75	*	.95	.59
Coming 'round the Mountain*	.15	*	.74	.47

Appendix E: Contextual categories for melodies in the stimulus set (Chapter 3)

Table E-1: Contextual categorization of melodies in the stimulus set.

Melody Name	Category
For He's a Jolly Good Fellow	Ceremonial
Happy Birthday	
O Canada	
Wedding March (Here Comes The Bride)	
Bingo	Children
Camp Town Racing	
Coming 'round The Mountain	
Frere Jacques	
If You're Happy And You Know It	
London Bridge	
Mary Had A Little Lamb	
Michael Row the Boat	
My Bonnie	
Old MacDonald	
Oh Suzanna	
Rock a Bye Baby	
Row Row Row Your Boat	
The Itsy Bitsy Spider	
The Song That Never Ends	
Three Blind Mice	
Twinkle Twinkle Little Star	
Yankee Doodle	
Deck the Halls	Christmas
Hark The Herald Angels Sing	
Jingle Bells	
Rudolph the Red Nosed Reindeer	
Silent Night	
We Wish You A Merry Christmas	
When The Saints Come Marching In	
White Christmas	
Blue Danube	Classical
Ode To Joy	
Russian Dance (Nutcracker)	
Hey Jude	Pop
I Will Survive	
Addams Family Theme	Theme
Doe, a Deer	
Lone Ranger Theme (William Tell)	

Note: Categorizations determined by mutual agreement between co-authors.

Table E-2: Contextual categorization of melodies named in misidentifications.

Melody Name	Category
Amazing Grace	Ceremonial
America the Beautiful	
God Save The Queen	
My Country 'tis of Thee	
National Anthem	
Star Spangled Banner	
US National Anthem	
Alouette	Children
Alphabet Song	
Baa Baa Black Sheep	
Bed Time Song	
Do Your Ears Hang Low?	
Fais Dodo	
Farmer in the Dell	
Head, Shoulders, Knees, & Toes	
Hickory Dickory Dock	
Hot Cross Buns	
Hot Potato	
Jack & Jill	
Kumbaya	
La Cucaracha	
Over The Hills We Go	
Pop Goes The Weasel	
Railroad	
Ring Around The Rosie	
Swanee River	
The Bear Went Over The Mountain	
There Was An Old Lady Who Swallowed a Fly	
There's A Hole in My Bucket	
Turkey In The Straw	
Angels We Have Heard On High	Christmas
Away in a Manger	
Frosty The Snowman	
Good King Wenceslas	
Little Drummer Boy	
O Tannenbaum	
Once In Royal David's City	
Silver Bells	
Au Clair de la Lune	Classical
Greensleeves	
Mozart	
Waltz Of The Flowers	
Bang Bang, Wanda Jackson	Christmas
How Much Is That Doggy In The Window	
Macarena	
You Are My Sunshine	
Alfred Hitchcock Theme	Theme
Barney Theme	
Baseball Game Song	
Bonanza Theme Song	
Elliot the Moose Theme	
Hockey Game	
I'd Do Anything (Oliver)	
The Munsters Theme	
Pink Panther Theme	
Somewhere Over The Rainbow	
Theme from My Fair Lady	

Note: Melodies listed in this table were named in participants' misidentifying responses, i.e. none of these melodies were presented during the experiment. Categories determined by mutual co-author agreement.

Appendix F: Stimulus details (Chapter 4)

Table F-1: Stimulus details

Beat Salience Level	Stimulus	Song Title	Artist	Album	Excerpt time	Tempo (BPM)
High	1	The Maker	Omar S	Single	4:45 - 5:02	121
	2	Alpha Male	Royksopp	The Understanding	2:36 - 2:53	135
	3	Sparks	Royksopp	Melody AM	0:32 - 0:49	85
	4	Up There	Centovalley	August	0:42 - 0:59	135
	5	Dirty Dancehall	The Zutons	Who Killed the Zutons	0:49 - 1:05	88
	6	Dwrcan	Bibio	Ambivalance Avenue	1:12 - 1:29	93
	7	Everybody's Stalking	Badly Drawn Boy	The Hour of Bewilderbeast	0:23 - 0:40	83
	8	Fall Apart	Lukid	Foma	1:22 - 1:39	99
Medium	9	Contact Note	Jon Hopkins	Contact Note	3:53 - 4:10	107
	10	The Black Forest Bear	Kobaya (Sub-Opt)	Ocean of Orbs	1:47 - 2:03	100
	11	Closing In	Imogen Heap	Speak for Yourself	2:17 - 2:34	112
	12	Children's Limbo	Venetian Snares	Find Candace	1:03 - 1:20	170
	13	Ratsback2 (Saitone Remix)	Plaid	Tekkonkinkreet OST	1:51 - 2:08	141
	14	The Sentinel	Horror Inc.	Horrorama EP	2:00 - 2:17	125
	15	Matter of Time	ASC	Nothing is Certain	0:45 - 1:02	85
	16	Myxamatosis	Radiohead	Hail to the Thief	0:38 - 0:55	99
Low	17	When I look at your face I laugh and cry	A Setting Sun & Shigeto	Table for Two	2:32 - 2:48	120
	18	Omgyjya Switch 7	Aphex Twin	Drukqs Disk 1	0:00 - 0:17	95
	19	The Black Page #1	Frank Zappa	Läther (Disc 2)	0:37 - 0:54	124
	20	Indigo	Monolake	Single	0:29 - 0:46	118
	21	Mouse Bums	Mu-ziq	Bilious Paths	0:20 - 0:37	120
	22	Bucephalus Bouncing Ball	Aphex Twin	Come to Daddy	3:02 - 3:19	83
	23	Distant Father Torch	Clark	Growl's Garden	0:37 - 0:54	77
	24	Pleasure Is All Mine	BjÃűrk	Medúlla	2:50 - 3:09	95
High	Practice 1	Satellite	Guster	Gangin up on the Sun	3:09 - 3:26	132
Medium	Practice 2	Heysatan	Sigur Ros	Takk	0:40 - 0:57	78
Low	Practice 3	Nightjar	Jon Hopkins	Contact Note	0:37 - 0:54	90
High	Practice 4	Cool for Cats	Squeeze	Greatest Hits	00:32 - 0:49	144
Medium	Practice 5	Gonna Be Sick	Beardyman	I Done A Album	1:23 - 1:40	140
Low	Practice 6	Diabolical Minds	Buckethead	Bucketheadland	0:08 - 0:25	90

Appendix G: Contrasting naïve and informed tap densities (Chapter 4)

The three plots on the following pages display tap density distributions generated during naïve tapping (sub-trial 1) and informed tapping (sub-trial 3), for stimuli respectively exhibiting high, medium, and low levels of beat salience. Black-shaded curves represents naïve tapping; white-shaded curves represent informed tapping (sub-trial 3); grey shade represents areas of overlap.

Figure G-1: Tap density distributions generated in response to high beat salience stimuli.

Figure G-2: Tap density distributions generated in response to medium beat salience stimuli.

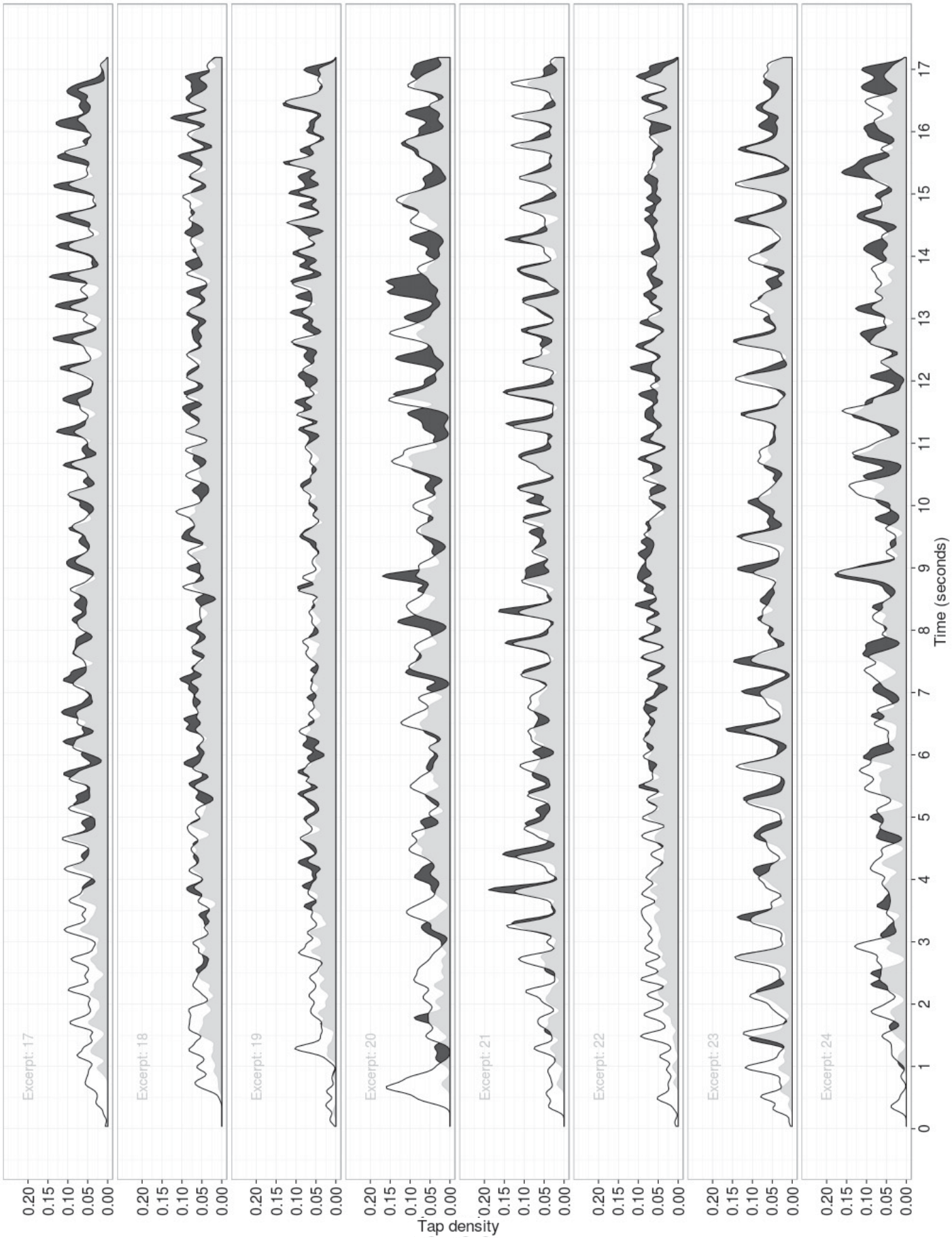


Figure G-3: Tap density distributions generated in response to low beat salience stimuli.

Appendix H: Informed consent forms (Chapters 3 & 4)

The following two pages provide facsimiles of the blank forms signed by participants to indicate informed consent in the melody identification studies reported in Chapter 3, and in the beat salience studies reported in Chapter 4.

Informed consent form - McGill University

WHY ARE WE DOING THIS RESEARCH? Our goal is to establish scientific knowledge about our ability to recognize familiar melodies. It has nothing to do with your personality or motivations or intelligence.

PRIVACY. All the responses that we collect will be aggregated for statistical analysis and so your own responses will be averaged with those of many other people. We know that you value your privacy. You will not be identified as an individual in any scientific report of this research, and your name will not be linked to your responses in this study.

DISCUSSION OF RESEARCH IDEAS. We cannot discuss our ideas with you before the experiment takes place, but we will be happy to talk with you about our hypotheses and theories afterwards.

WHAT WILL HAPPEN DURING THE EXPERIMENT? You will be seated in a quiet room, hear melodies and write down identifying information about them. Specific instructions will follow. The sounds will not be loud enough to cause you discomfort or to adversely affect your hearing. There is no known risk associated with the experiment. You will be free to discontinue your participation at any time without penalty.

The whole experiment will take approximately one hour. Credit will be granted for participating.

Participant's Statement:
"I have read the description of the research project and hereby agree to participate. I am aware that the results will be used for research purposes only, that my identity will remain confidential, and that I can withdraw at any time, if I so wish."

Name: _____ Date: _____

Signature: _____

ID #:

Room 106, School of Information Studies, 3661 Rue Peel [C2]

This research is being conducted by David Weigl under the supervision of McGill professors Catherine Guastavino and Daniel Levitin. Contact david.weigl@mail.mcgill.ca for more information.
McGill University, 555 Sherbrooke Street West, Montreal, QC H3A 1E3. Phone: (514) 398-4535 x. 0300.
Fax: (514) 398-2962.

McGill

Participant #: _____

INFORMED CONSENT FORM

Perception of Beat Salience

Purpose of the project: The goal of this experiment is to investigate the underlying mechanisms responsible for the perception of the beat.

Nature of the Experiment: You will read the instructions before proceeding with the experiment. After the experiment, you will be invited to fill out a background questionnaire. The experiment itself will take approximately 1 hour.

Participation: Participation in this study is not expected to pose any risks to you. Your participation is completely voluntary, and you are free to terminate the experiment at any time for any reason.

Confidentiality: Your confidentiality and anonymity for serving in this experiment will be protected. A record of your participation will be saved using coded identification numbers independent of your name or any other identifying information. Only the principal investigator will have access to the records associating participants' names with their respective identification numbers. Your name will not be mentioned in any report on this study.

Participant's Statement: I have read and understood the above details of the experiment, and I freely consent to participate.

_____ _____

Name of Participant (print) Date

_____ _____

Participant signature Name of Experimenter (print)

David Sears, PhD student (Music Theory), McGill University
david.sears@mail.mcgill.ca
David Weigl, PhD Candidate (Information Studies), McGill University
david.weigl@mail.mcgill.ca
Jason Hockman, PhD Candidate (Music Technology), McGill University
jason.hockman@mail.mcgill.ca
Prof. Stephen McAdams, Schulich School of Music, McGill University
smc@music.mcgill.ca

References

Aljanaki, A., Wiering, F., & Veltkamp, R. (2014). *Collecting annotations for induced musical emotion via online game with a purpose emotify* (Tech. Rep. No. UU-CS-2014-015). Department of Information and Computing Sciences, Utrecht University.

Andrews, M. W., Dowling, W. J., Bartlett, J. C., & Halpern, A. R. (1998). Identification of speeded and slowed familiar melodies by younger, middle-aged, and older musicians and nonmusicians. *Psychology and Aging*, *13*(3), 462.

Aucouturier, J.-J., & Bigand, E. (2012). Mel Cepstrum & Ann Ova: The difficult dialog between mir and music cognition. In *Proceedings of the International Society for Music Information Retrieval* (pp. 397–402).

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using eigen and s4. [r package] version 1.1-7.* http://CRAN.R-project.org/package=lme4.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289–300.

Bertin-Mahieux, T., Ellis, D. P. W., Whitman, B., & Lamere, P. (2011). The million song dataset. In *Proceedings of the 12th International Society for Music Information Retrieval Conference* (p. 591-596).

Bertin-Mahieux, T., Hoffman, M. D., & Ellis, D. P. W. (2011). Tutorial 1: Million song dataset. In *Proceedings of the 12th International Society for Music Information Retrieval Conference* (p. 15). (Abstract)

Bhatara, A., Tirovolas, A. K., Duan, L. M., Levy, B., & Levitin, D. J. (2011). Perception of emotional expression in musical performance. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(3), 921.

Bischoff, K., Firan, C. S., Nejdl, W., & Paiu, R. (2009). How do you feel about dancing queen?: deriving mood & theme annotations from user tags. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries* (pp. 285–294).

Boulgouris, N. V., Plataniotis, K. N., & Hatzinakos, D. (2004). Gait recognition using dynamic time warping. In *Multimedia Signal Processing, 2004 IEEE 6th Workshop on* (pp. 263–266).

Bregman, A. S. (1990). Auditory scene analysis: The perceptual organization of sound. In (pp. 1–45). MIT Press.

Bregman, A. S., & Campbell, J. (1971). Primary auditory stream segregation and perception of order in rapid sequences of tones. *Journal of Experimental Psychology*, *89*, 244-249.

Brinegar, J., & Capra, R. (2011). Managing music across multiple devices and computers. In *Proceedings of the 2011 iConference* (pp. 489–495). ACM. Retrieved

2015-06-29, from http://dl.acm.org/citation.cfm?id=1940828

Brown, J. C., Hodgins-Davis, A., & Miller, P. J. (2006). Classification of vocalizations of killer whales using dynamic time warping. *Journal of the Acoustical Society of America*, *119*(3), EL34.

Burgoyne, J. A., Wild, J., & Fujinaga, I. (2011). An expert ground truth set for audio chord recognition and music analysis. In *Proceedings of the 12th International Society for Music Information Retrieval Conference* (p. 633-38).

Burke, S. M., & Conklin, D. (2010). *MIDI-Perl (v. 0.83) [perl module].* http://search.cpan.org/~conklin/MIDI-Perl/lib/MIDI.pm.

Byrd, D., & Crawford, T. (2002). Problems of music information retrieval in the real world. *Information Processing & Management*, *38*(2), 249-272.

Chapados, C., & Levitin, D. J. (2008). Cross-modal interactions in the experience of musical performances: Physiological correlates. *Cognition*, *108*(3), 639–651.

Chen, J. L., Penhune, V. B., & Zatorre, R. J. (2008). Moving on time: brain network for auditory-motor synchronization is modulated by rhythm complexity and musical training. *Journal of Cognitive Neuroscience*, *20*(2), 226–239.

Christensen, R. H. B. (2013). *ordinal—regression models for ordinal data.* (R package version 2013.9-30. http://www.cran.r-project.org/package=ordinal/)

Copland, A. (1957). *What to listen for in music.* McGraw-Hill Book Company.

Cuddy, L. L., Balkwill, L.-L., Peretz, I., & Holden, R. R. (2005). Musical difficulties are rare. *Annals of the New York Academy of Sciences*, *1060*(1), 311–324.

Cunningham, S. J. (2002). User studies: A first step in designing an MIR testbed. In *The MIR/MDL evaluation project white paper collection edition# 2* (pp.

17–19).

Cunningham, S. J., Bainbridge, D., & McKay, D. (2007). Finding new music: a diary study of everyday encounter with novel songs. In *Proceedings of the 8th International Conference on Music Information Retrieval* (pp. 83–88).

Cunningham, S. J., Jones, M., & Jones, S. (2004). Organizing digital music for use: an examination of personal music collections. In *Proceedings of the 5th International Symposium on Music Information Retrieval.*

Cunningham, S. J., & Nichols, D. M. (2009). Exploring social music behaviour: An investigation of music selection at parties. In *Proceeding of 10th International Society for Music Information Retrieval Conference* (pp. 26–30).

Cunningham, S. J., Reeves, N., & Britland, M. (2003). An ethnographic study of music information seeking: implications for the design of a music digital library. In *Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital Libraries* (pp. 5–16). IEEE Computer Society. Retrieved 2014-06-06, from `http://dl.acm.org/citation.cfm?id=827142` (ID: 30)

de Bruijn, O., & Spence, R. (2001). Serendipity within a ubiquitous computing environment: A case for opportunistic browsing. In *Ubiquitous Computing* (pp. 362–369).

Degara, N., Rúa, E. A., Pena, A., Torres-Guijarro, S., Davies, M. E., & Plumbley, M. D. (2012). Reliability-informed beat tracking of musical signals. *Audio, Speech, and Language Processing, IEEE Transactions on*, *20*(1), 290–301.

De Roure, D., Klyne, G., Page, K. R., Pybus, J. P. N., & Weigl, D. M. (2015). Music and science: Parallels in production. In

Proceedings of the 2nd International Workshop on Digital Libraries for Musicology (pp. 17–20). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2785527.2785530 doi: 10.1145/2785527.2785530

Dervin, B., & Nilan, M. (1986). Information needs and uses. *Annual Review of Information Science and Technology*, *21*, 3.

Dixon, S. (2007). Evaluation of the audio beat tracking system beatroot. *Journal of New Music Research*, *36*(1), 39–50.

Dowling, W. J., & Harwood, D. L. (1987). Music cognition. *Psychomusicology*, *7*(1), 91.

Downie, J. S. (2003). Music information retrieval. *Annual Review of Information Science and Technology*, *37*(1), 295–340.

Downie, J. S. (2004a). A sample of music information retrieval approaches. *Journal of the American Society for Information Science and Technology*, *55*(12), 1033-1036.

Downie, J. S. (2004b). The scientific evaluation of music information retrieval systems: foundations and future. *Computer Music Journal*, *28*(2), 12-23.

Downie, J. S. (2008). The music information retrieval evaluation exchange (2005-2007): A window into music information retrieval research. *Acoustic Science & Technology*, *29*(4).

Downie, J. S., Byrd, D., & Crawford, T. (2009). Ten years of ISMIR: reflections on challenges and opportunities. In *Proceedings of the 10th International Society for Music Information Retrieval Conference* (pp. 13–18).

Downie, J. S., Futrelle, J., & Tcheng, D. (2004). The international music information retrieval systems evaluation laboratory: Governance, access, and security. In *Proceedings of the 5th International Symposium on Music Information Retrieval* (p. 9-14).

Downie, J. S., West, K., Ehmann, A., & Vincent, E. (2005). The 2005 music information retrieval evaluation exchange (mirex 2005): Preliminary overview. In *Proceedings of the International Conference on Music Information Retrieval* (pp. 320–323).

Drewing, K., Aschersleben, G., & Li, S.-C. (2006). Sensorimotor synchronization across the life span. *International Journal of Behavioral Development*, *30*(3), 280–287.

Dubois, D. (2000). Categories as acts of meaning: The case of categories in olfaction and audition. *Cognitive Science Quarterly*, *1*(1), 35–68.

Ehrenfels, C. v. (1937). On gestalt-qualities. *Psychological Review*, *44*(6), 521.

Ellis, D. (2003). *Dynamic Time Warp (DTW) in Matlab [web resource].* http://www.ee.columbia.edu/~dpwe/resources/matlab/dtw/.

Ellis, D. P. (2007). Beat tracking by dynamic programming. *Journal of New Music Research*, *36*(1), 51–60.

Fitch, W. T. (2013). Rhythmic cognition in humans and animals: distinguishing meter and pulse perception. *Frontiers in Systems Neuroscience*, *7*.

Foxton, J. M., Nandy, R. K., & Griffiths, T. D. (2006). Rhythm deficits in 'tone deafness'. *Brain and Cognition*, *62*(1), 24–29.

Futrelle, J., & Downie, J. S. (2002). Interdisciplinary communities and research issues

in music information retrieval. In *Proceedings of the International Conference on Music Information Retrieval* (pp. 215–221).

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2012). *irr: Various coefficients of interrater reliability and agreement. [r package] version 0.84.* http://CRAN.R-project.org/package=irr.

Gjerdingen, R., & Perrott, D. (2008). Scanning the dial: The rapid recognition of music genres. *Journal of New Music Research*, *37*(2), 93–100.

Grahn, J. A., & Schuit, D. (2012). Individual differences in rhythmic ability: Behavioral and neuroimaging investigations. *Psychomusicology: Music, Mind, and Brain*, *22*(2), 105.

Greasley, A., & Lamont, A. (2009). Exploring Engagement with Music in Everyday Life using Experience Sampling Methodology. In *Proceedings of the 7th Triennial Conference of European Society for the Cognitive Sciences of Music* (pp. 165–174). Jyväskylä, Finland.

Guastavino, C. (2007). Categorization of environmental sounds. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *61*(1), 54.

Guastavino, C., Gómez, F., Toussaint, G., Marandola, F., & Gómez, E. (2009). Measuring similarity between flamenco rhythmic patterns. *Journal of New Music Research*, *38*(2), 129–138.

Gustafson, K. (1988). The graphical representation of rhythm. *Progress Reports from Oxford Phonetics*, *3*, 6–26.

Halpern, A. R. (1988). Perceived and imagined tempos of familiar songs. *Music*

Perception, 193–202.

Halpern, A. R. (1989). Memory for the absolute pitch of familiar songs. *Memory & Cognition*, *17*(5), 572–581.

Hankinson, A., Roland, P., & Fujinaga, I. (2011). The music encoding initiative as a document-encoding framework. In *Proceedings of the 12th International Society for Music Information Retrieval Conference* (pp. 293–298).

Harman, D. K. (2005). The TREC test collections. In *TREC: Experiment and evaluation in information retrieval* (chap. 2). MIT Press.

Hébert, S., & Peretz, I. (1997). Recognition of music in long-term memory: Are melodic and temporal patterns equal partners? *Memory & Cognition*, *25*(4), 518–533.

Hjørland, B. (2010). The foundation of the concept of relevance. *Journal of the American Society for Information Science and Technology*, *61*(2), 217–237.

Holzapfel, A., Davies, M. E., Zapata, J. R., Oliveira, J. L., & Gouyon, F. (2012). Selective sampling for beat tracking evaluation. *Audio, Speech, and Language Processing, IEEE Transactions on*, *20*(9), 2539–2548.

Honing, H. (2010). Lure(d) into listening: The potential of cognition-based music information retrieval. *Empirical Musicology Review*, *5*, 121–126.

Huron, D. (1988). Error categories, detection and reduction in a musical database. *Computers and the Humanities*, *22*(4).

Huron, D. (2000). Perceptual and cognitive applications in music information retrieval. In *Proceedings of the International Symposium on Music Information Retrieval.*

Huron, D. (2006). *Sweet anticipation: music and the psychology of expectation*. MIT Press.

Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: elements of a cognitive IR theory. *Journal of Documentation*, *52*(1), 3–50.

Ingwersen, P., & Järvelin, K. (2005). *The Turn: integration of information seeking and retrieval in context* (1st ed.). Springer.

Inskip, C., MacFarlane, A., & Rafferty, P. (2010, June). Creative professional users' musical relevance criteria. *Journal of Information Science*, *36*(4), 517–529.

Iversen, J. R., & Patel, A. D. (2008). The Beat Alignment Test (BAT): Surveying beat processing abilities in the general population. In *Proceedings of the 10th International Conference on Music Perception & Cognition.*

Jansen, B. J., & Rieh, S. Y. (2010). The seventeen theoretical constructs of information searching and information retrieval. *Journal of the American Society for Information Science and Technology*, *61*(8), 1517–1534.

Jones, C., Marron, J., & Sheather, S. (1996). Progress in data-based bandwidth selection for kernel density estimation. *Computational Statistics*(11), 337–381.

Kassler, M. (1966, April). Toward musical information retrieval. *Perspectives of New Music*, *4*(2), 59–67.

Kim, J.-K., & Levitin, D. J. (2002). Configural processing in melody recognition. *Canadian Acoustics*, *30*(3), 156–157.

Klapuri, A. P., Eronen, A. J., & Astola, J. T. (2006). Analysis of the meter of acoustic musical signals. *Audio, Speech, and Language Processing, IEEE Transactions on*, *14*(1), 342–355.

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in psychtoolbox-3. *Perception*, *36*(14), 1.

Knees, P., & Widmer, G. (2008). Searching for music using natural language queries and relevance feedback. In *Adaptive Multimedia Retrieval: Retrieval, User, and Semantics* (pp. 109–121). Springer. Retrieved 2015-12-14, from http://link.springer.com/chapter/10.1007/978-3-540-79860-6_9

Krebs, H. (1999). *Fantasy pieces : Metrical dissonance in the music of robert schumann: Metrical dissonance in the music of robert schumann*. Oxford University Press, USA.

Kroher, N. (2013). *The flamenco cante: Automatic characterization of flamenco singing by analyzing audio recordings* (Unpublished doctoral dissertation). Master thesis, Universitat Pompeu Fabra, Barcelona, Spain.

Kubovy, M., & Van Valkenburg, D. (2001). Auditory and visual objects. *Cognition*, *80*, 97–126.

Kuusi, T. (2009). Tune recognition from melody, rhythm and harmony. In *7th Triennial Conference of European Society for the Cognitive Sciences of Music.*

Laplante, A. (2010a). The role people play in adolescents' music information acquisition. In *Workshop on Music Recommendation and Discovery.*

Laplante, A. (2010b). User's relevance criteria in music retrieval in everyday life: An exploratory study. In *Proceedings of the 11th International Society for Music Information Retrieval Conference* (pp. 601–606).

Laplante, A. (2011). Social capital and music discovery: an examination of the ties through which late adolescents discover new music. In *Proceedings of the*

12th International Society for Music Information Retrieval Conference (pp. 341–346).

Laplante, A., & Downie, J. S. (2006). Everyday life music information-seeking behaviour of young adults. In *Proceedings of the Seventh International Conference on Music Information Retrieval* (pp. 381–382).

Laplante, A., & Downie, J. S. (2011, July). The utilitarian and hedonic outcomes of music information-seeking in everyday life. *Library & Information Science Research*, *33*(3), 202–210. doi: 10.1016/j.lisr.2010.11.002

Lartillot, O., Eerola, T., Toiviainen, P., & Fornari, J. (2008). Multi-feature modeling of pulse clarity: Design, validation and optimization. In *Proceedings of the 9th International Conference for Music Information Retrieval* (pp. 521–526).

Lee, J. H. (2010). Analysis of user needs and information features in natural language queries seeking music information. *Journal of the American Society for Information Science and Technology*, *61*(5), 1025–1045.

Lee, J. H., & Cunningham, S. J. (2012). The impact (or non-impact) of user studies in music information retrieval. In *Proceedings of the International Society for Music Information Retrieval* (pp. 391–396).

Lee, J. H., & Cunningham, S. J. (2013). Toward an understanding of the history and impact of user studies in music information retrieval. *Journal of Intelligent Information Systems*, *41*(3), 499–521.

Lee, J. H., & Waterman, N. M. (2012). Understanding user requirements for music information services. In *Proceedings of the 13th International Society for Information Retrieval Conference* (pp. 253–258).

Lenth, R. (2014). *lsmeans: Least-squares means. [r package] version 2.00-1.* http://CRAN.R-project.org/package=lsmeans.

Levering, M. (2000). Intellectural property rights in musical works. In *Proceedings of the International Symposium on Music Information Retrieval.*

Levitin, D. J. (1994). Absolute memory for musical pitch: Evidence from the production of learned melodies. *Perception & Psychophysics*, *56*(4), 414–423.

Li, X., Liang, Z., Kleiner, M., & Lu, Z.-L. (2010). RTbox: a device for highly accurate response time measurements. *Behavior Research Methods*, *42*(1), 212–.

Lonsdale, A. J., & North, A. C. (2011, February). Why do we listen to music? A uses and gratifications analysis: Music uses and gratifications. *British Journal of Psychology*, *102*(1), 108–134. doi: 10.1348/000712610X506831

Macrae, R., & Dixon, S. (2010). Accurate real-time windowed time warping. In *Proceedings of the 11th International Society for Music Information Retrieval Conference* (pp. 423–428).

Marslen-Wilson, W. D. (1987). The temporal structure of spoken language understanding. *Cognition*, *25*, 1-71.

Meadow, C. T., Boyce, B. R., & Kraft, D. H. (2000). Text searching. In *Text information retrieval systems* (2nd ed., chap. 9). Emerald Group Publishing.

Müllensiefen, D., & Frieler, K. (2004). Cognitive adequacy in the measurement of melodic similarity: Algorithmic vs. human judgments. *Computing in Musicology*, *13*(2003), 147–176.

Müllensiefen, D., & Frieler, K. (2006). The SIMILE algorithms documentation 0.3. *White Paper*.

Myers, C., Rabiner, L., & Rosenberg, A. (1980). An investigation of the use of dynamic time warping for word spotting and connected speech recognition. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'80* (Vol. 5, pp. 173–177).

Newton, E. (1990). *The rocky road from actions to intentions* (Unpublished doctoral dissertation). Stanford University.

Oliveira, J. L., Gouyon, F., Martins, L. G., & Reis, L. P. (2010). IBT: A real-time tempo and beat tracking system.

Organisciak, P., & Downie, J. S. (2015). Improving consistency of crowdsourced multimedia similarity for evaluation. In *Proceedings of the 15th ACM/IEEE-CE on Joint Conference on Digital Libraries* (pp. 115–118).

Parncutt, R. (1994). A perceptual model of pulse salience and metrical accent in musical rhythms. *Music Perception*, 409–464.

Parsons, D. (1975). *The directory of tunes and musical themes*. S. Brown.

Patel, A. D. (2014). The evolutionary biology of musical rhythm: was Darwin wrong? *PLoS Biology*, *12*(5).

Patel, A. D., Iversen, J. R., Bregman, M. R., & Schulz, I. (2009). Experimental evidence for synchronization to a musical beat in a nonhuman animal. *Current Biology*, *19*(10), 827–830.

Phillips-Silver, J., Toiviainen, P., Gosselin, N., Piché, O., Nozaradan, S., Palmer, C., & Peretz, I. (2011). Born to dance but beat deaf: a new form of congenital amusia. *Neuropsychologia*, *49*(5), 961–969.

Pinheiro, J., & Bates, D. (2006). *Mixed-effects models in S and S-PLUS*. Springer

Science & Business Media.

Prince, J. B. (2011). The integration of stimulus dimensions in the perception of music. *The Quarterly Journal of Experimental Psychology*, *64*(11), 2125–2152.

R Core Team. (2015). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*(5), 347–356.

Repp, B. H. (2005). Sensorimotor synchronization: a review of the tapping literature. *Psychonomic Bulletin & Review*, *12*(6), 969–992.

Repp, B. H., & Su, Y.-H. (2013). Sensorimotor synchronization: a review of recent research (2006–2012). *Psychonomic Bulletin & Review*, *20*(3), 403–452.

Robertson, T. (2006). *Proceedings of the 18th Australia Conference on Computer-Human Interaction Design Activities, Artefacts and Environments.* New York, NY: ACM.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization.* Hillsdale, NJ: Erlbaum.

Ross, B. H., & Murphy, G. L. (1999). Food for thought: Cross-classification and category organization in a complex real-world domain. *Cognitive Psychology*, *38*(4), 495–553.

Saracevic, T. (1975). Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, *26*(6), 321–343.

Saracevic, T. (1997). The stratified model of information retrieval interaction: Extension and applications. In *Proceedings of the Annual Meeting-American Society for Information Science* (Vol. 34, pp. 313–327).

Saracevic, T. (2007a). Relevance: A review of the literature and a framework for thinking on the notion in information science. part III: Behavior and effects of relevance. *Journal of the American Society for Information Science and Technology*, *58*(13), 2126–2144. doi: 10.1002/asi.20681

Saracevic, T. (2007b). Relevance: A review of the literature and a framework for thinking on the notion in information science. part II: nature and manifestations of relevance. *Journal of the American Society for Information Science and Technology*, *58*(13), 1915–1933. doi: 10.1002/asi.20682

Schedl, M., Flexer, A., & Urbano, J. (2013). The neglected user in music information retrieval research. *Journal of Intelligent Information Systems*, *41*(3), 523–539.

Schulkind, M., Posner, R., & Rubin, D. (2003). Musical features that facilitate melody identification: How do you know it's "Your" song when they finally play it? *Music Perception*, *21*(2), 217–249.

Schulkind, M. D. (1999). Long-term memory for temporal structure. *Memory & Cognition*, *27*(5), 896–906.

Sheather, S. J., & Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 683–690.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420.

Smith, B. (1995). Psiexp: an environment for psychoacoustic experimentation using the ircam musical workstation. In *Society for Music Perception and Cognition Conference* (Vol. 95).

Snyder, J., & Krumhansl, C. L. (2001). Tapping to ragtime: Cues to pulse finding. *Music Perception*, *18*(4), 455–489.

Steffens, J., & Guastavino, C. (2015). Trends in momentary and retrospective soundscape judgments. *Acta Acustica united with Acustica*, *101*(4), 713-722.

Taheri-Panah, S., & MacFarlane, A. (2004). Music information retrieval systems: why do individuals use them and what are their needs? In *Proceedings of the 5th International Conference on Music Information Retrieval.*

Toiviainen, P., & Snyder, J. S. (2003). Tapping to bach: Resonance-based modeling of pulse. *Music Perception: An Interdisciplinary Journal*, *21*(1), 43–80.

Toussaint, G. (2006). A comparison of rhythmic dissimilarity measures. *FORMA*, *21*(2), 129–149.

Tremblay, A., & Ransijn, J. (2013). *LMERConvenienceFunctions: Model selection and post-hoc analysis for (G)LMER models. [r package] version 2.05.* http://CRAN.R-project.org/package=LMERConvenienceFunctions.

Typke, R., Wiering, F., & Veltkamp, R. C. (2005). A survey of music information retrieval systems. In *Proceedings of the 6th International Conference on Music Information Retrieval* (pp. 153–160).

Tzanetakis, G., Essl, G., & Cook, P. (2002). Human perception and computer extraction of musical beat strength. In *Digital Audio Effects (DAFx)* (Vol. 2).

Uitdenbogerd, A. L., & Yap, Y. W. (2003). Was Parsons right? an experiment in

usability of music representations for melody-based music retrieval. In *Proceedings of the 4th International Conference on Music Information Retrieval* (pp. 75–79).

Van Valkenburg, D., & Kubovy, M. (2003). In defense of the theory of indispensable attributes. *Cognition*, *87*, 225–233.

Voorhees, E. M., & Harman, D. K. (2005). The Text REtrieval Conference. In *TREC: Experiment and evaluation in information retrieval* (chap. 1). MIT Press.

Wang, A. (2006). The Shazam music recognition service. *Communications of the ACM*, *49*(8), 44–48.

Weber, R. (1991). The continuous loudness judgement of temporally variable sounds with an "analog" category procedure. In *5th Oldenburg Symposium on Psychological Acoustics* (pp. 267–289).

Weigl, D. M., & Guastavino, C. (2011). User studies in the music information retrieval literature. In *Proceedings of the 12th International Society for Music Information Retrieval Conference* (p. 335-340).

Weigl, D. M., & Guastavino, C. (2013). Applying the stratified model of relevance interactions to music information retrieval. *Proceedings of the American Society for Information Science and Technology Conference*, *50*(1), 1–4.

Wertheimer, M. (1923). Untersuchungen zur Lehre von der Gestalt. II. *Psychological Research*, *4*(1). (translated from German by D. M. Weigl)

White, B. W. (1960). Recognition of distorted melodies. *The American Journal of Psychology*, *73*(1), 100–107.

Whittall, A. (2011). Melody. In A. Latham (Ed.), *The Oxford companion to music.*

Wilson, T. D. (1981). On user studies and information needs. *Journal of Documentation*, *37*(1), 3–15.

Wilson, T. D. (1997). Information behaviour: An interdisciplinary perspective* 1. *Information Processing & Management*, *33*(4), 551–572.

Xu, Y. (2007). Relevance judgment in epistemic and hedonic information searches. *Journal of the American Society for Information Science and Technology*, *58*(2), 179–189.

Zapata, J. R., Holzapfel, A., Davies, M. E., Oliveira, J. L., & Gouyon, F. (2012). Assigning a confidence threshold on automatic beat annotation in large datasets. In *Proceedings of the 13th International Society for Music Information Retrieval Conference* (pp. 157–162).

Zobel, J., & Moffat, A. (1998). Exploring the similarity space. In *ACM SIGIR Forum* (Vol. 32, pp. 18–34).