Shape, drawing and gesture
cross-modal mappings of sound and music

Kussner, Mats

Awarding institution:
King's College London

SHAPE, DRAWING AND GESTURE: CROSS-MODAL MAPPINGS OF SOUND AND MUSIC

Mats B. Küssner
PhD in Music Research

In liebevoller Erinnerung an Lucie Oma

Abstract

This thesis investigates the notion of shape in music from a psychological perspective. Rooted in the embodied cognition research programme, it seeks to understand what kinds of shapes listeners with varying levels of musical expertise perceive in sound and music by engaging them in overt actions. To that end, two empirical studies have been carried out. In the first experiment, a sample of musically trained and untrained participants was asked to represent visually a series of pure tones varied in pitch, loudness and tempo—as well as two short musical excerpts—by means of an electronic graphics tablet. In the second experiment, a new sample of musically trained and untrained participants was asked to represent gesturally a series of pure tones varied in pitch, loudness and tempo, as well as sixteen short musical excerpts. In one of two experimental conditions, participants' gestures—captured with Microsoft® Kinect™ and Nintendo® Wii™ Remote Controller—created a real-time visualization on a screen in front of them. In order to shed light on cross-modal mappings between drawing/gesturing features (x-, y- and z-coordinates) and sound features (pitch, loudness) correlation analyses, as well as more advanced mathematical tools such as Gaussian processes, were applied. Results revealed that musically trained participants are generally more consistent in representing sound features cross-modally (e.g., pitch–height) but also less diverse in their approaches than untrained participants. Most participants mapped pitch onto the vertical axis and time onto the horizontal axis. Loudness was mostly represented by size in drawings and by various mapping strategies in gestures such as height, size and muscular energy. Representing musical excerpts gesturally led to a wide range of strategies including, dancing, conducting, air instrument playing and tracing of musical features. Findings are discussed in light of embodied music cognition and current theoretical developments within the cognitive sciences.

Table of Contents

Table of Figures

Table of Tables

Acknowledgements

I am hugely indebted to my supervisors Daniel Leech-Wilkinson and Helen Prior for their invaluable support throughout my PhD. Dan was always available when I asked for a meeting, sent him a new draft of a chapter or an article to read, reread and re-reread, required yet another reference letter or, most importantly, needed new ideas on musical shapes and the 'big picture' – even if that meant sacrificing his lunch break or interrupting his sabbatical year. Helen's vast knowledge of the literature—paired with her attention to detail, patience and diligence—is something for every researcher to aspire to. I have learned a great deal from her during my time at King's and am especially grateful to her for continuing my supervision during her maternity leave. I was extremely lucky to have had two very kind and empathetic supervisors who not only provided intellectual stimulation but also fantastic moral support. I am grateful to Dan for organizing regular (cake) seminars and to its core members during my time at King's—Amy, Anna, Eugene, Ed and Nick—for their valuable comments on my work (and for pretending not to have noticed that I was a person in a Chinese room when it came to questions pertaining to recordings). Special thanks go to Eugene for sharing his PhD experience with me and for teaching the 'Music and the Brain' module at King's together, which made the whole endeavour much more enjoyable.

I am very grateful to everyone at the AHRC Research Centre for Musical Performance as Creative Practice who supported me and my work in the past few years in whichever ways: John Rink, Eric Clarke, Mirjam James, Karen Wise, Mark Doffman, David Mawson and my fellow PhD students Michael, Myles and Emily. Many thanks also to Nicolas Gold at UCL who created the software for my drawing experiment, and to Dan Tidhar who developed it further to capture three-dimensional movement shapes. I was lucky enough to be able to collaborate with two mathematicians, Genevieve Noyce and Peter Sollich, who introduced me to Gaussian processes and answered patiently all my naïve questions on maths and statistics: many thanks for that. I am also very grateful to Kristian Nymoen and Baptiste Caramiaux for sharing their enthusiasm for sound and gestures and discussing issues related to data analysis with me, and to George Athanasopoulos for introducing me to cross-cultural representations of musical shapes. It's great to know that there are likeminded folks out there! Moreover, I would like to thank four music students, Charlotte Nohavicka, Freddie Hosken, Jordan Theis and Gregory Gu, who helped me speed up my analysis considerably by cutting, annotating or rating a huge

number of video clips from my gesture experiment. And I am grateful to Daniel Müllensiefen and Aaron Williamon, who agreed to examine this thesis.

I was very lucky to be able to take part in the International Summer School in Systematic Musicology at the University of Jyväskylä in 2010 and attend a number of SysMus conferences in Ghent (2009), Cambridge (2010), Cologne (2011) and Genoa (2013). Besides learning valuable skills for my research and being able to present my findings in a very friendly and constructive environment, I met many new friends, especially Edith Van Dyck, Konstantina Orlandatou, Pieter-Jan Maes, Jon Hargreaves and Kjetil Bøhler, all of whom—to return the compliment—are first-class musicologists and have made my PhD endeavour much more fun. Exploring together various sites in Jyväskylä, London, Ghent and Thessaloniki were experiences I wouldn't want to miss and I hope there are many more to come in the future (in warm places, perhaps). Special thanks are due to other members of the SysMus family, too. Many thanks go out to Stella for lending me the keys to her house and car in Crete, giving me the chance to relax from the PhD life while she was doing hard fieldwork in India; to Manuela for fun times in Jyväskylä, London, Cambridge and Krems, whether dancing ceilidh or reading the bible in French; to Donald for a great workshop on EyesWeb in Jyväskylä, an even greater night out in Würzburg and a marvellous time in Genoa; and to Amos for his inquisitiveness the night before our final group presentation in Jyväskylä and telling me all viola jokes and then some.

I am also very grateful to my friends in London—Bienam [bi:f], Jon D., Marta, Tasos, Filipe, John and Daniel—and back home—Flo, Gerhard and Götz—for heading to the pub together and sharing their stories with me. There's not much more you need in life, really.

I wouldn't have been able to finish this thesis without the support of my family, regardless of how far they are away. Above all, I would like to thank my mum for her unconditional support, her infinite trust in my abilities and for wholeheartedly accepting my choice of living a life far from home. And I would like to thank my grandma who didn't live to see this work completed and to whom I dedicate my thesis. In our weekly phone conversations—sometimes even via video call when she had lunch at my uncle's—she told me enthusiastically all the news about London she had seen on TV or read in a magazine, and I just listened because the findings of my work suddenly seemed irrelevant.

Most importantly, I would like to thank you, Irati: I am endlessly grateful for your love, joy, patience and support, and for keeping me sane during difficult times. When I see you after a long day of work I know that the future will be bright.

London, May 2014

List of Publications

Publications in peer-reviewed journals and book chapters

- Küssner, M. B., & Leech-Wilkinson, D. (2014). Investigating the influence of musical training on cross-modal correspondences and sensorimotor skills in a real-time drawing paradigm. *Psychology of Music, 42*(3), 448-469. doi: 10.1177/0305735613482022
- Küssner, M. B., Tidhar, D., Prior, H. M., & Leech-Wilkinson, D. (2014). Musicians are more consistent: Gestural cross-modal mappings of pitch, loudness and tempo in real-time. *Frontiers in Psychology, 5*, Article 789. doi: 10.3389/fpsyg.2014.00789
- Küssner, M. B. (2013). Music and shape. *Literary and Linguistic Computing, 28*(3), 472-479. doi: 10.1093/llc/fqs071
- Noyce, G. L., Küssner, M. B., & Sollich, P. (2013). Quantifying shapes: Mathematical techniques for analysing visual representations of sound and music. *Empirical Musicology Review, 8(2)*, 128-154.
- Küssner, M. B. (forthcoming-2015). Shape, drawing and gesture: Empirical studies of cross-modality. In D. Leech-Wilkinson & H. M. Prior (Eds.), *Music and shape.* Oxford: Oxford University Press.

Publications in peer-reviewed conference proceedings

- Küssner, M. B. (2013). Shaping music visually. In A. C. Lehmann, A. Jeßulat, & C. Wünsch (Eds.), *Kreativität – Struktur und Emotion* (pp. 203-209). Würzburg: Königshausen & Neumann.
- Küssner, M. B. (2012). Creating shapes: Musicians' and non-musicians' visual representations of sound. In J. Wewers & U. Seifert (Eds.), *Under construction: Trans- and interdisciplinary routes in music research. Proceedings of SysMus11, Cologne, 2011* (pp. 111-122). Osnabrück: epOs-Music.
- Küssner, M. B., Prior, H. M., Gold, N. E., & Leech-Wilkinson, D. (2012). *Getting the shapes "right" at the expense of creativity? How musicians' and non-musicians' visualizations of sound differ.* Paper presented at the 12th International Conference on Music Perception and Cognition / 8th Triennial Conference of the European Society for the Cognitive Sciences of Music, Thessaloniki, Greece.

- Küssner, M. B., Gold, N., Tidhar, D., Prior, H. M., & Leech-Wilkinson, D. (2011). *Synaesthetic traces: Digital acquisition of musical shapes*. Paper presented at the 2nd Supporting Digital Humanities Conference: Answering the unaskable, Copenhagen, Denmark.

Contributions to Chapters

Although I will use “I” throughout this thesis for the purpose of consistency, this work would not have been possible without the help of my colleagues (see also Acknowledgements). I would like to clarify here who contributed what to which chapter, as well as pointing out which parts of my thesis have already appeared, or are about to appear, in publications.

Chapter 1 is entirely my own work. Some early reflections on music and shape that went into this chapter can be found in my conference paper “Shaping music visually” which I wrote in 2010/11. Some parts of Chapter 1—especially the review of children’s drawing—will appear in my book chapter for the ‘Music and shape’ volume (forthcoming in 2015).

Chapter 2—while drawing on a conference paper from 2011 (Synaesthetic traces: Digital acquisition of musical shapes), co-authored by Nicolas Gold, Dan Tidhar, Helen Prior and Daniel Leech-Wilkinson—has been completely rewritten and extended by myself. Descriptions of the software, developed originally by Nicolas Gold, appear as quotations in this chapter to distinguish Nicolas Gold’s work from mine. The section on traditional experimental paradigms will appear in the ‘Music and shape’ volume.

Chapter 3 is based on the journal article “Investigating the influence of musical training on cross-modal correspondences and sensorimotor skills in a real-time drawing paradigm” co-authored with Daniel Leech-Wilkinson. The same study also appeared in two conference proceedings in 2012 (see List of Publications). Helen Prior and Daniel Leech-Wilkinson were involved in the design of this study and Nicolas Gold created the drawing software. I collected the data, analysed the data and wrote up the paper. Daniel Leech-Wilkinson also provided helpful comments for the discussion section.

Chapter 4 is based on the journal article “Quantifying shapes: Mathematical techniques for analysing visual representations of sound and music”, written by Genevieve Noyce, Peter Sollich and myself. I initiated this project by establishing a collaboration with the Department of Mathematics at King’s College London in 2011. In several meetings with me, Peter Sollich and Genevieve Noyce identified adequate analysis techniques for the purposes of my study. Genevieve Noyce then analysed the data and created the figures (including the captions). Genevieve and I interpreted the results and wrote up the journal article together, with helpful input from Peter Sollich. Genevieve Noyce had written up an earlier version of the journal article

for her Master's project in 2012. Since the (theoretical) mathematical parts in Chapter 4 are entirely the work by Genevieve Noyce and Peter Sollich I will use quotes from our journal article to distinguish as clearly as possible their contribution from mine.

Chapters 5 and 6 are based on a project carried out with input from Dan Tidhar, Helen Prior and Daniel Leech-Wilkinson, all of whom worked with me towards the final design of the experiment. Dan Tidhar created the software to capture the hand gestures and assisted with the analysis of Chapter 5. Four student helpers assisted with the analysis of video footage or provided ratings of the perceived shapes for Chapter 6. I analysed the data and wrote up the results. Helen Prior and Daniel Leech-Wilkinson provided helpful comments for the discussion sections. A short version of Chapter 5 has appeared in the journal article "Musicians are more consistent: Gestural cross-modal mappings of pitch, loudness and tempo in real-time" co-authored with Dan Tidhar, Helen Prior and Daniel Leech-Wilkinson.

Chapter 7—the final chapter in which I summarize the main findings of my thesis and discuss them critically in the wider theoretical context—is entirely my own work. Parts of this chapter will be used for my book chapter for the volume on 'Music and shape'.

Preface

In the past three and a half years I inevitably ended up explaining the topic of my thesis to a wide range of people and professionals from various backgrounds: family members, friends, colleagues, musicians, administrative staff, taxi drivers, bankers, waiters, hairdressers, landlords, CEOs and plumbers, to list but a few. When telling these people that my thesis is about 'music and shape' I have often experienced reactions such as "Music and what?" or—in case the acoustics were sufficiently clear and my conversational partners certain they heard the word 'shape' correctly—"What do you mean by shape?"

Although there are several ways of interpreting the meaning of the latter question, it seems to suggest that the concepts of music and shape are stored rather far apart in people's minds: most people do not think of music in terms of shape *at first sight*. There is one exception, however, and that is the case of musicians. For this group of professionals, the intention behind this question might be to ascertain which of the (many) shapes is referred to since it has been revealed that musicians use the notion of shape or shaping very frequently when talking or thinking about music (Prior, 2011b). Almost 90% of the musicians who took part in Prior's online survey reported that they would think about shape when thinking about how to perform music. Interestingly, despite the vast differences in musicians' connotations of shape—phrasing, form, structure, direction, contour, dynamics, line, emotion, gesture, intensity, tension, expression, feeling, colour, pattern, movement, or flow—they are usually able to communicate fluently with fellow musicians or other music professionals such as critics and musicologists when talking about the shape of music. The ubiquitous use of the concept 'shape' among musicians thus suggests that music in performance is perhaps more commonly and more widely experienced as having shape than previously imagined.

I will argue throughout this thesis that listening to music involves all sorts of shapes – even if one is not fully aware of them, or would label them differently perhaps (or not at all if they defy verbal descriptions). For instance, most (Western) people would probably agree that melodies consist of ascending and descending pitches, forming melodic lines that can be visualized and thought of in spatial terms, regardless of whether someone is familiar with Western music notation or not. Therefore, if you have people listen to a simple ascending-descending melodic line and ask them to draw it on a sheet of paper or to indicate the course of this melodic line

with their hand in the air, you can expect that they will readily do so. That is not so say, though, that all their responses will be the same – quite the contrary. One crucial aspect of this thesis is to investigate the differences and commonalities of these shapes, produced in drawing and gesturing experiments, while another is to study the extent to which musical training influences such responses. By studying the perceived shapes of sounds and music, I intend to shed light on facets of music cognition that may be sometimes hard to verbalize and are best expressed through bodily gestures. Yet, these shapes may have the potential to reveal much about how we experience and make sense of music.

This thesis is divided into seven chapters: In Chapter 1, I will outline the theoretical background for my empirical investigations with particular emphasis on embodied (music) cognition. I will review how the concept of 'shape' has been applied in musicology before and discuss relevant work on drawings and gestures in response to music before stating the aims of this thesis. In Chapter 2, I will provide an overview of traditional experimental paradigms and the methods, hardware, software and analytical tools used in this thesis. In Chapter 3, I will report findings from a study in which musically trained and untrained participants were asked to draw the shapes of auditory stimuli consisting of pure tones and two brief musical excerpts. In Chapter 4, I will present the application of some advanced mathematical tools on the dataset from this drawing study, developed in collaboration with the Department of Mathematics at King's College London. In Chapter 5, I will report findings from a study in which musically trained and untrained participants were asked to represent the shape(s) of various pure tones with arm and hand gestures. In Chapter 6, I will present findings from the same gesturing study, reporting and interpreting participants' responses to various short musical excerpts. In Chapter 7, I will summarize and discuss the main findings of my empirical studies, outline some issues and potential improvements, and point to future research.

Chapter 1: Introduction

1.1 Theoretical background

I am approaching the endeavour of investigating the shapes people hear in sound and music from a music-psychological perspective. That is, I am not concerned with how one *could* hear shapes, nor how one *should* hear them, nor how they *were* heard in the past but how they *are* heard by listeners in the present day.[1] Consequently, most theories I draw upon can be situated in the vast realm of the cognitive sciences.

1.1.1 Traditional cognitivist view

Developed in the 1960s as a reaction against behaviourism, the so-called cognitive revolution led to the view of the mind as a computer: it receives incoming information, performs rule-based calculations by manipulating symbols and the result of this computation triggers a response. In other words, cognition starts and ends at the natural boundary of an organism – everything external to it is irrelevant. Many of the experimental paradigms developed in the early days of cognitive psychology have advanced our knowledge considerably and are therefore still in use today. However, the view of the mind as a symbol-manipulating system, detached from the outside world, is not without problems and has therefore been attacked right from the start. Claxton (1980, p. 13) summarizes some of the issues aptly, referring to a very common experimental procedure in which participants are asked to respond to stimuli with button presses:

> "…[cognitive psychology] does not, after all, deal with whole people, but with a very special and bizarre – almost Frankensteinian – preparation, which consists of a brain attached to two eyes, two ears, and two index fingers. This preparation is only to be found inside small, gloomy cubicles, outside which red lights burn to warn ordinary people away. It stares fixedly at a small screen, and its fingers rest lightly and expectantly on two small squares of black plastic – microswitches. It does not feel hungry or tired or inquisitive; it does not think extraneous thoughts or try to understand what is going on. It simply *processes information*. It is, in short, a computer, made in the image of the larger electronic organism that sends it stimuli and records its responses."

[1] This distinction is adapted after Juslin (2013, p. 283) in which he discusses the role of normative and descriptive approaches to musical aesthetics.

The lack of "dealing with whole people", i.e. their bodies as well as their brains, and the neglect of the role of the environment are two of the main reasons that led second-generation cognitive scientists[2] to develop an alternative research programme which is known as embodied cognition (Shapiro, 2007).

1.1.2 Embodied cognition

The beginnings of this research programme can be traced back to phenomenology—most notably Husserl, Heidegger, Merleau-Ponty and Sartre—although the focus has been shifted from mere subjective experiences to explaining broader underlying cognitive mechanisms (Wilson & Foglia, 2011). While it is in many ways opposed to the traditional cognitivist view (for a juxtaposition of both views see Cowart, 2004), the extent to which proponents of the embodied research programme accept or reject features of the traditional view differs according to the refinement of embodied theories in particular disciplines.[3] There is, however, broad agreement that the role of the body and its interaction with the physical environment are crucial for cognition. Rather than receiving incoming information passively, cognition is conceived of as an active process in which goal-directed actions and sensory feedback constitute a continuous dynamic pattern of sensorimotor activation. The external world is not merely a mental representation, as traditional cognitivists would have it, but forms, by interacting with our bodies, an integral part of the cognitive process. The shape of our bodies determines how we may interact with objects in the physical environment and hence how we perceive the world and how we form concepts and categories. According to Wilson and Foglia (2011), the body acts as a constraint, distributor and regulator of cognitive processes. It regulates the coordination of action and cognition across space and time and distributes cognitive activity across both neural and non-neural structures of the body, or even beyond the body, as advocates of the extended mind have argued (Clark & Chalmers, 1998).[4] The notion of the body as a constraint is particularly interesting because it stresses the body's causal role in cognitive processes. It suggests that we are able to grasp objects figuratively *because* we are able to grasp them literally. If we could manipulate objects only with our mouth—but not with our hands—we would have a different concept of them. It should be emphasized that this sort of conceptualisation via

[2] The distinction between first- and second-generation cognitive science is borrowed from Johnson (2007).
[3] Cowart (2004) distinguishes between compatibilist and purist approaches. Proponents of the former suggest that theoretical tools from the traditional view should be kept as long as the embodied view is not able to provide adequate alternatives (e.g., for meta-cognitive processes), whereas proponents of the latter hold that the whole traditional view is flawed and should be substituted, e.g., with the dynamics system theory (Thelen & Smith, 1994).
[4] The idea behind the extended mind theory is that the mind is not restricted to the body. Objects such as a notebooks or smart phones can become part of the mind, as long as information stored within those objects is reliably accessible.

sensorimotor activation is not only limited to concrete objects but is likely to form the basis of more abstract concepts too (Gallese & Lakoff, 2005). The body thus plays a pivotal role in shaping our cognitive processes – from low-level perceptions and actions to higher-level functions and capacities.

Imagine then, for a moment, a parallel universe in which human beings have evolved to form ball-like creatures – with no indication of where the upper and lower ends are located. Would they still hear the sound of an ascending-descending melodic line as going up and down? The seminal work by Lakoff and Johnson (1980) on conceptual metaphors suggests that the answer is 'No'.

1.1.3 Conceptual metaphors

The central thesis put forward by Lakoff and Johnson (1980) is that metaphors are not mere figures of speech—rhetorical means conceived in an abstract world—but grounded in our bodily experiences situated in a cultural environment. What is more, they affect our cognition—how we make sense of the world—and consequently, guide our behaviour. One example often referred to in their book is the metaphor ARGUMENT IS WAR. Our language is interspersed with military expressions—attacking, shooting, defending, targeting etc.—when describing, and hence thinking of, arguments. Our conceptualisation of an argument, they reason, would be quite different if instead of the war metaphor we would have come to think of argument as dance, for instance. The reason why we have developed certain metaphors and not others is because they are all ultimately based on our bodily experiences. Identifying various types of metaphors, Lakoff and Johnson's orientational metaphors are particularly pertinent here. As creatures inhabiting a three-dimensional world, we have developed a sense of what is up and down, not least because bipedalism has structured our bodies into upper and lower parts. Many concepts in our culture are thought of in spatial terms such as HAPPY IS UP and SAD IS DOWN. The experiential basis for this kind of mapping is that a positive state is usually accompanied by upright posture, whereas a negative or depressed state is often associated with droopy posture. Thus, our experience of space is mapped onto the abstract concepts of happiness and sadness – something Lakoff and Johnson refer to as metaphorical mapping from a source domain (usually part of our everyday experience) onto a target domain (usually an abstract concept).

Another example of metaphorical mapping particularly relevant in the context of this thesis is that of MORE IS UP and LESS IS DOWN. In the physical environment we often experience that 'more' of a certain substance entails the level going up. Think of the amount of hay on a haystack, snow on a field, water in a glass or books piling up on a desk – there are innumerable instances in which we encounter 'more' as being higher up (in the air). Even though the experiential bases of HAPPY IS UP and MORE IS UP are not the same, the crucial point is that in both cases the experience of 'upness' gives rise to metaphorical mappings that structure our thinking.

There are often several metaphorical mappings for one concept, e.g., IDEAS ARE FOOD, IDEAS ARE PEOPLE, IDEAS ARE PLANTS, IDEAS ARE PRODUCTS etc., all of which are consistent internally but not necessarily across them. Even though each metaphor structures a concept only partially, in their sum they form a coherent system, enabling us to use different kinds of mappings within one sentence, e.g., "Although she presented a budding theory (IDEAS ARE PLANTS), one of her claims was hard to swallow (IDEAS ARE FOOD)".

Lakoff and Johnson's rethinking of metaphors had far-reaching consequences for many disciplines such as linguistics, philosophy, psychology, literary analysis, music theory, politics, law and even mathematics. At the same time, it also influenced and provided further theoretical support for the embodied cognition research programme, in which the body and its interaction with the physical environment are central. As mentioned above, goal-directed actions play a pivotal role in embodied cognition. Next, I will focus on a theory that was postulated prior to the development of the embodied research programme but can now be seen as being part of it because of its emphasis on the role of action in perception.

1.1.4 Motor theory of speech perception

The central argument of this theory—initiated by Liberman and colleagues in the 1950s (for a review see Galantucci, Fowler, & Turvey, 2006)—is that the object of speech perception, i.e. what we perceive when we perceive speech, is the motor action that produces the speech utterances rather than the auditory information, i.e. the acoustic signal, that is the result of this motor action. According to this view, speech perception is not a passive procedure by which incoming auditory information is processed, as proponents of the traditional cognitivist view would explain it, but rather an active process in which the internal simulation of the motor action

of the vocal tract is responsible for perceiving speech (Liberman & Mattingly, 1985).[5] Being highly controversial at the beginning, there is now increasing evidence in favour of the motor theory of speech perception. Early hints came from studies showing that visual input from speech gestures can influence our interpretation of the auditory information (McGurk & MacDonald, 1976) and enhance speech perception in a noisy environment (Sumby & Pollack, 1954). Liberman's own early research (Liberman, Delattre, & Cooper, 1952; Liberman, Delattre, Cooper, & Gerstman, 1954) revealed the principle of coarticulation, which says that the production of syllables is context-dependent: the motor actions necessary to produce adjacent vowels and consonants overlap temporally. In a clever set of experiments, Whalen (1984) showed that people need longer to identify a vowel preceded by a consonant if the vowel they heard was produced in the context of a different consonant. That is to say, they were tricked by the coarticulatory information of the consonant present in the vowel because they simulated internally the coarticulated motor actions. The fact that the motor system is activated during speech perception has been shown behaviourally (Bell-Berti, Raphael, Pisoni, & Sawusch, 1979; Cooper, 1979), and more recently neuroscientifically (Fadiga, Craighero, Buccino, & Rizzolatti, 2002; Hickok, Buchsbaum, Humphries, & Muftuler, 2003). One conclusion from the review by Galantucci and colleagues is that the motor theory of (speech) perception is not restricted to speech but applies to broader, more general, aspects of perception and action.[6] An important step that led to this progress was the discovery of mirror neurons in the 1990s (Pellegrino, Fadiga, Fogassi, Gallese, & Rizzolatti, 1992).

1.1.5 Mirror neurons

Investigating neural correlates of motor actions in macaque monkeys, Rizzolatti and colleagues discovered that some neurons in the premotor cortex are not only active when monkeys grasp a particular object but also when they observe humans or other monkeys perform the same action (for reviews see Kilner & Lemon, 2013; Rizzolatti & Craighero, 2004; Rizzolatti, Fogassi, & Gallese, 2001). Evidence for the existence of mirror neurons in humans[7] (Fadiga, Fogassi, Pavesi, & Rizzolatti, 1995) has led to an explosion of research, showing that mirror neurons are also active when only the sound of a clearly distinguishable action—e.g., the cracking of a

[5] Liberman and colleagues also claimed that speech processing is carried out in a specialized module (cf. Fodor, 1983). However, given recent findings, this claim is no longer tenable (Galantucci et al., 2006).
[6] This is why the word 'speech' in 'motor theory of (speech) perception' is nowadays often omitted. Interestingly, Liberman's theory has been much more positively received outside his own field of speech perception, possibly because of its wider, non-specific scope.
[7] However, concerns have been raised whether humans' and monkeys' mirror neuron systems play the same functional roles (Heyes, 2010).

peanut's shell for monkeys (Kohler et al., 2002), or typing on a keyboard for humans (Aziz-Zadeh, Iacoboni, Zaidel, Wilson, & Mazziotta, 2004)—is heard. In other words, there is now reason to believe that the brain has evolved to form a system in which actions and their sensory consequences—visual, auditory, tactile etc.—are encoded in the very same neural patterns. A theory which postulated the so-called 'common coding principle' of actions and perceptions has been formalized by Prinz and collaborators (Prinz, 1990; Prinz & Hommel, 2002). Through repeated couplings of actions and their sensory feedback joint cognitive representations are formed in which the neural codes of perception and action are indistinguishable. And this emerging neural pattern is the same as that which theorists of embodied cognition refer to when talking about sensorimotor activation. It is the basis for all our thinking and behaviour, possibly for all higher-order functions such as language, concept formation, memory, problem solving and mental imagery. After outlining the general theoretical framework of this thesis, situated within the embodied cognition approaches, I will now turn to the more specific case of music cognition.

1.2 Embodied music cognition

As mentioned above, the embodied cognition research programme consists of many specialized theories in a vast range of disciplines. Whereas some fields such as linguistics and robotics dealt with and incorporated the embodied cognition approach early on, others, such as musicology, might be described as latecomers. Leman (2007) was arguably the first to formalize a theory of embodied music cognition.[8] The central thesis of his book is that the body is a natural mediator between musical experience and physical environment. More specifically, our bodies act as mediators between physical properties of music, i.e. the acoustic signal, and our musical thoughts, values and intentions. As Leman is also concerned with the role of mediation technologies—for example, devices for accessing music gesturally—he attempts to find a way of describing music that bridges the gap between first-person and third-person descriptions. First-person descriptions are the ones traditionally found in musicology where the subjective interpretation of music—both as text and sound—in a cultural and historical context forms the basis of all argumentation. On the other hand, third-person descriptions deal with all objectively measurable data related to music such as its physical properties (frequency, amplitude, etc.) but

[8] According to Leman (2007, pp. 43-45), the beginnings of embodied music cognition can be traced back to the early 20th century, including work on empathy and expressive movement in music (Lipps, 1903), on dynamic rhythmic shapes of composition styles (Becking, 1928) and on inner shapes and motion in music (Truslit, 1938). For an overview see also Repp (1993b).

also people's quantitatively measured movements in response to music or (neuro-)physiological correlates of music listening. Since Leman aims for a scientific theory—i.e. one that allows researchers to replicate results and falsify hypotheses—yet does not want to get rid of the subjective element either, he proposes a second-person description. Second-person descriptions are subjective responses to music articulated through verbal or non-verbal descriptions of bodily phenomena, forming a continuum from non-verbal bodily gestures to verbal descriptions of bodily states. Unlike first-person descriptions, experience is understood as articulated, not interpreted. In other words, what is essential for a musicology based on the embodied cognition research programme is not the interpretation of subjective thoughts on music, i.e. cerebral intentionality, but the articulation of the personal experience of music through the body, i.e. corporeal intentionality. The possibility of measuring overt corporeal articulations of subjective experiences of music enables researchers to correlate these second-person descriptions with third-person descriptions, ultimately aiming at working out causal relationships between the physical sound and the subjective experience. Leman is cautious to note, though, that the subjective experience can never be completely transferred into physical signals, as second- and third-person descriptions are based on different ontologies. The difference is made by the intentionality, which is present in the corporeal articulations, but not in the physical properties of music.

Leman goes on to specify three types of corporeal articulations—synchronization, attuning and empathy—which I will briefly review here. The underlying assumption of all three types is that corporeal articulations attempt to mimic or imitate aspects of the music in order to make sense of it. Synchronization refers to human beings' tendency to imitate the perceived movement in external stimuli such as sound.[9] In the case of music, people often spontaneously or subconsciously tap along with a beat or nod their head. By doing so, they synchronize with the beat, imitating a structural property of the sound. While tapping along might be seen as a passive activity, Leman also refers to 'inductive resonance' which means that a person is actively experiencing the illusion that they are controlling the action properties of an external stimulus, for example, when conducting an orchestra. However, synchronization is also seen as a low-level sensorimotor activity, contrasting it to embodied attuning, which is seen as a higher-level process.

[9] This idea fits with the ideomotor theory (for a review see Shin, Proctor, & Capaldi, 2010), which says that our actions are represented as the perceived sensory effects, and the presence of a sensory effect is sufficient to trigger its corresponding motor action.

Leman (2007, p. 115) defines embodied attuning as "navigation with or inside music", meaning that a person deliberately attends to some features of the music and tries to imitate them corporeally. Of the three examples provided—the probe tone method (Krumhansl & Kessler, 1982), vocal attuning (Heylen, Moelants, & Leman, 2006) and graphical attuning (De Bruyn, Moelants, & Leman, 2012)—the latter is of particular interest here, as it exemplifies a drawing experiment within the embodied music cognition paradigm. The authors of this study asked a group of participants with Autism Spectrum Disorder (ASD) and a group of controls to draw along with various musical excerpts on an electronic graphics tablet, focussing on either the rhythmic structure or the melodic contour. Results revealed that both groups performed equally well in the rhythm condition, but participants with ASD performed slightly better than controls in the melody condition. Overall, the results are interpreted as evidence that patients with ASD have no difficulty imitating structural aspects of the music at this level of corporeal articulation.[10] The next level described by Leman pertains to empathy.

Empathy generally refers to the ability to identify and experience someone else's emotional state. According to Leman (2007, p. 122), "empathy with music would thus refer to an imitation of the music's emotional intentionality." He distinguishes between three levels of empathic involvement with music. *Observation of affect* is a process by which the listener is able to recognize an emotion expressed in the music by comparing the internally simulated motion of the music with previously experienced sensorimotor patterns of emotional events. If this internal simulation becomes overt—through corporeal articulations such as hand and arm gestures, for example—one may speak of *imitation of affect.* The perceived affective motion pattern in music expressed through overt body movements is supposed to give rise to a deeper understanding of musical phenomena. Finally, if this overt corporeal imitation of the motion in music activates the listener's affective system, the emotion corresponding to the motion patterns is also felt, hence the *feeling of affect*. In all three levels "music can be conceived as a virtual social agent" (p.126), whose behaviour is identified, imitated or felt, entailing various degrees of involvement with music.[11] Having provided a short overview of the main concepts of embodied music cognition I will now turn to the matter of shape in a musical context.

[10] Patient and control groups also performed equally well in a simple synchronization task of tapping along to a musical beat, and in a task assigning emotional labels to musical excerpts, though the latter took participants with ASD five times longer than controls.

[11] Although I am aware that there is a huge philosophical literature on music and agency I will not draw on it as explained at the beginning of this chapter.

1.3 Shape

As my observation from the beginning of this thesis suggests, the majority of people do not readily associate music with shape if no further explanation is given. Thus, in this section, I will outline what kinds of shapes I have in mind when talking about shapes. Historically, the term 'shape' has been used by musicologists and music scholars to denote a variety of concepts; the diversity of connotations of 'shape' nowadays (Prior, 2011a) is therefore no recent development but can already be seen in early writings on music.[12] For example, Stainer and Barrett (1888[1876], p. 174) use it in a structural sense to define 'form': "[t]he shape and order in which musical ideas are presented." Implicit in such a use of the term 'shape' is the understanding of music as text, that is, as a set of hierarchically organized symbols existing in an abstract space. To be sure, these are *not* the shapes I refer to here. Nor am I concerned with the observable shapes, lines, dots, curves and colours of graphic scores, which composers have produced from the 1950s onward, in search of new ways of musical expressiveness. Rather, I am concerned with the shapes of music in performance, i.e. music as sound, which is in fact what many musicians nowadays, as well as in the past—see, for instance, Matthay's (1945[1913]) discussion of musical shape pertaining to the moving or progressing shapes of music in performance—have in mind when talking and thinking about shape. At this point, it is important to note that there are probably many music scholars investigating some concept they decided to call 'shape', which, however, has nothing to do with my endeavour here. On the other hand, there are also (music) scholars investigating the underlying *concept* of shape as I refer to it here, though each of them might have given it a different name. I will solely focus on the work of the latter group and ignore that of the former. Instances where both the label and the concept overlap have been rare in musicology,[13] but there is at least one pertinent exception which I will discuss in more detail here since my empirical investigations (see Chapters 3–6) are based on such an approach.

In an attempt to give new impulses to thinking in music theory, Godøy (1997) proposed a paradigm in which music is not conceptualized as a system of abstract symbols but rather in terms of dynamic musical shapes that exist in a time-space-continuum. The basic idea—based

[12] It is not clear, however, when and in which context 'music' and 'shape' first appeared together in writings, let alone in conversations about music.

[13] The first project investigating the shape of music in performance systematically from various angles is the context in which this thesis was produced. Information about the 'Shaping music in performance' project can be found at http://www.cmpcp.ac.uk/smip.html.

on the seminal work by composer and pioneer of *musique concrète*, Pierre Schaeffer (1966)—is that listeners (and scholars) should focus on the unfolding sonic event with all its emergent musical qualities such as melody, harmony and rhythm – but also texture and timbre, that is, those qualities that music theory and analysis have mostly avoided due to the difficulty of integrating them into an abstract system. By listening to short segments of music over and over again—what Schaeffer called the method of 'sillon fermé'—one is supposed to disengage from the sound source and focus entirely on the sonic qualities (also known as acousmatic listening). Note that, according to Godøy, this view is not necessarily at odds with the motor theory of speech perception according to which perception is a process of internally simulating the sound-producing actions – i.e. something occurring automatically without our being able to control it. To bring Schaeffer's conceptualization of musical objects in line with recent evidence from cognitive science, and more specifically, the embodied cognition research programme, Godøy (1997, 2006) proposes a 'gestural-sonorous object'. According to this view, musical objects are multi-modal, comprised of both the sound-producing actions and the sound itself. In Godøy's reading of Schaeffer, acousmatic listening involves the tracing of various sound features that evolve in time and space; hence it is possible to think of them in terms of gestures. In Godøy's (2006, p. 149) words:

> "This means that from continuous listening and continuous sound-tracing, we actually recode musical sound into multimodal gestural-sonorous images based on biomechanical constraints (what we imagine our bodies can do), hence into images that also have visual (kinematic) and motor (effort, proprioceptive, etc.) components."

He also points out that Schaeffer's system of categorizing sound features—the typology and morphology of musical segments—is largely based on gestural metaphors and that the three broad typologies put forward by Schaeffer—impulsive, sustained and iterative—can be directly linked to the physical gestures necessary to produce such sounds. Godøy (2006) proposes a triangular model consisting of gesture sensations which interact with, and form the binding element between, the continuous sound on the one hand and the multimodal gesture images on the other. A gestural-sonorous object is usually only a few seconds long and can be located at a 'meso level'.[14] In order to access its dynamic shapes and gain knowledge of otherwise hidden or inaccessible features of the music, one is supposed to draw along with a pen on a paper, on

[14] Godøy (2006) distinguishes between micro, meso and macro level. Micro level refers to the low-level acoustic features of the sound, whereas macro level refers to the broader context of the continuous stream of musical objects.

a computer screen, or simply in one's mind. According to Godøy (1997), this analysis-by-synthesis approach to music listening has the potential to open up new strands of research that go beyond approaches based on the acoustic signal only. Specifically, Godøy (2006, p. 156) suggests studying the correlations between our perception of musical segments and its acoustic signals by asking people to trace sound with a pen on a graphics tablet or by gesturing in the air. And this is precisely the approach taken in this thesis, even though I will take Godøy's theorizing merely as a starting point, without following the system of sound categorization introduced by Schaeffer.

For the purposes of this thesis, shape thus denotes not only shapes heard in music but rendered visible through overt hand movements—either in drawings or three-dimensional gestures—during the process of listening. In other words, shape is the overarching term for the perceptions of sound and music made tangible in drawings and gestures. To be sure, I do not aim to provide a new definition of (musical) shape here, since the strength of the concept of shape lies precisely in its flexibility. Rather, I intend to unite existing concepts such as Godøy's 'sound tracings' and Leman's second-person descriptions under the umbrella term of shape to shed light on music cognition. The visual shapes of sound tracings created by the body when acting as a mediator between the physical environment and the musical mind are at the core of my investigation: they are seen as windows onto our musical thinking and behaviour. What is more, since empirical investigations of adults' drawings and gestures in response to sound and music are relatively sparse—as the discussion in the following sections and chapters will show—all studies focussing on human beings' visualizations of sound features and music are embraced in my approach to musical shape. This means I will also draw on a large body of research on auditory-visual correspondences (for a more general review of cross-modal correspondences see Spence, 2011) to provide a more solid foundation for my work on musical shapes. Although my own approach is more concerned with 'shaping' rather than 'shape'—in other words, the active process of representing music visually or gesturally, rather than its resulting product—there should be no doubt that I see both as intrinsically linked, the 'shaping' causing the 'shape' and the 'shape' triggering the 'shaping'. For that reason, evidence from more passive paradigms in which sound features and music are simply matched to *provided* visual shapes—i.e. without the direct involvement of the participant's body causing the visual shapes—will be reviewed in the introductory parts of my empirical studies as well (see Chapters

3 and 5). For the remainder of this chapter, however, I will focus in more detail on active, overt visualizations[15] of sound and music, as obtained in empirical studies asking individuals to draw and gesture in response to auditory stimuli.

1.4 Drawing

As mentioned above, empirical evidence relevant for my own drawing experiment—whether obtained in drawing or other experimental paradigms—will be reviewed at length in the introductory section of Chapter 3. This section is primarily aimed at providing the larger research context for studying music perception and cognition by means of drawings. The first systematic investigation of visual representations of music was carried out in the field of developmental psychology, creating a large empirical body of children's drawings of sound and music. Empirical evidence from these studies will be reviewed here because both their findings and paradigms have been influential for studies with adults.

Children's drawings have played an important role in psychology as it has been argued that they form a window on the child's cognitive development (Hargreaves, 1978; Olson, 1970; Piaget & Inhelder, 1973; Werner, 1980). In a musical context, drawings of simple sound stimuli and musical excerpts might thus be seen as windows on music cognition and the development of musical thinking (Davidson & Scripp, 1988). Even though it is a moot question exactly what these drawings represent—windows on, or perhaps rather reflections of, musical thinking (Barrett, 2000)—they have been studied extensively since the end of the 1970s, owing to two broadly shared assumptions among researchers (Barrett, 2005, p. 125): First, young children may not yet have developed the language to express adequately their musical thinking, and secondly, some musical experiences may defy linguistic descriptions and be better and more revealingly described non-verbally.[16]

In a series of seminal experiments investigating visual representations of simple rhythmic fragments, Bamberger (1980, 1982) paved the way for numerous studies investigating children's, as well as adults', invented notations of music. Based on the visual shapes produced in her experiments in which she asked children aged 4–12 years first to clap a simple rhythm and then to draw it, Bamberger (1982) proposed a developmental trajectory from 'rhythmic

[15] Note that I will use the expressions 'visualization', 'visual representation' and 'visual shape' synonymously hereafter.
[16] In fact, both assumptions—but particularly the second—are not restricted to children and suggest that studies should be carried out with adults as well.

scribbles' mimicking the clapping action with the pen, to figural representations capturing perceptual groupings of the sounds, and then to metric representations displaying the awareness of an underlying metric pulse by assigning each symbol a particular duration. However, this Piagetian view, in which one stage is replaced by the following, has been challenged by evidence showing that children acquire a 'database of strategies' (Barrett, 2005, p. 130), using one or several approaches that seem most appropriate given the nature of the task and the stimuli (Reybrouck, Verschaffel, & Lauwerier, 2009). For example, Upitis (1987), amongst others who extended the work on visual representations of rhythmic sequences (Davidson & Colley, 1987; Davidson & Scripp, 1988; K. C. Smith, Cuddy, & Upitis, 1994), found that, regardless of musical training, children aged 7–12 years are all able to make sense of rhythm by using figural, metrical or a combination of both types of visual representations. Applying various active and passive rhythm tasks, including clapping rhythms, drawing (and recognizing drawn) rhythms, verbal interpretations and tapping along, her findings suggest that children draw on a large pool of representational strategies. Importantly, she also emphasized the role of context, and was able to show in subsequent studies that children are less likely to represent the rhythmic structure if it is embedded within an unknown melody (Upitis, 1992). Only if the pitch structure is fairly simple (e.g., an ascending scale) and the rhythmic structure more complex do children show a more elaborate visual representation of the rhythmic structure (Upitis, 1990). Moreover, Davidson and Scripp (1988, p. 222) call for "increasingly divergent paths of rhythm and pitch in representational development", seeing "rhythm and pitch in a figure-ground relationship, that is, the rhythmic 'figure' in isolation becomes 'ground' when pitch is introduced into the context of the phrase" (p. 226). In a musical culture largely based on pitch, it is perhaps not surprising that children prefer, and find it easier, to draw the pitch rather than the rhythm of a melody. Recent findings support this tendency: Verschaffel and colleagues found that stimuli whose salient feature is the pitch or the melody give rise to more differentiated visualizations than stimuli whose salient feature is either related to rhythm or dynamics (Verschaffel, Reybrouck, Janssens, & Van Dooren, 2010).

These are all examples of individual musical parameters – studied either in isolation or within the context of simple musical fragments. The question of whether findings from such studies can and should be generalized to real musical excerpts is currently debated (Elkoshi, 2002; Reybrouck et al., 2009; Verschaffel et al., 2010). Asking over 100 children aged 7–8.5 years to

draw rhythmic sequences that were either produced (in isolation) by the children or part of a musical excerpt they listened to, Elkoshi (2002) found no correlation between the visualizations of short sound fragments and the musical excerpt, arguing that this gap cannot be closed and that one may not infer from one to the other (see also Reybrouck et al., 2009). On the other hand, more recent evidence suggests that such a correlation may well exist (Verschaffel et al., 2010). Testing a comparably large group of 8–9 and 11–12-year-olds, with and without musical training, it was revealed that the amount of differentiated visualizations[17] in response to short simple sound stimuli, each of which had been designed to highlight one specific musical parameter (pitch, duration and loudness), correlated positively with the amount of differentiated visualizations in response to real musical excerpts, chosen to highlight three corresponding musical features (melody, rhythm and dynamics).

Since the motto of the Gestaltists may well have a kernel of truth,[18] it is important to look at some of the evidence from real musical excerpts. Gromko (1994) asked 60 children aged 4–8 years to sing or play a short folk song provided by the author and then to "write the way the song sounds" (p. 139). In addition, the children's perceptual discrimination was tested in a standardized rhythm and tonal task. Results revealed a positive correlation between the musical understanding score—computed on the basis of the performance in the singing/playing and the perceptual discrimination task—and the depiction of rhythmic and tonal elements in their invented notations, suggesting that representation—alongside the more traditional measures of production and perception—may indeed reflect children's development of musical understanding. Comparing invented notations of familiar and unfamiliar melodies of 50 children aged 6–9 years with no formal musical training outside school, Upitis (1990) found that the most commonly produced visual shapes and symbols are "(a) icons, (b) words, (c) discrete marks for pitches and/or durations, and (d) continuous lines for pitch and/or mood" (p. 94). While there was no apparent age effect, an effect of familiarity showed that words and pictures were more common for familiar songs—according to the children, that is enough to recognize the tune—and discrete symbols for pitch more common for unfamiliar songs. Using the same familiar song

[17] To count as a differentiated visualization the drawing should display any of the following subcategories: (a) a sounding object or action hinting at the temporal unfolding of the sound stimulus, (b) an analogous image, that is, the change in one musical parameter expressed through an image, (c) non-formal notation of the music using abstract shapes, e.g., lines, circles, dots, or (d) formal-conventional notation. Juxtaposed to this category are global visualizations, further subdivided into: (a) depiction of one instrument or (b) several instruments, (c) evocation, that is, any associated pictorial response capturing the sound/music as a whole and (d) music icon, that is, depiction of music notes without referring to specific aspects of pitch, duration or loudness.

[18] Their motto might be succinctly expressed as, 'The whole is different than the sum of its parts'.

as Upitis—"Twinkle, Twinkle, Little Star"—and testing 20 Suzuki-trained[19] children aged 5–10 years (duration of training varying from 7 months to 4 years), Hair (1993/1994) found that apart from the youngest children who used pictures only, the choice of pictures, icons, music symbols and abstract lines and shapes was similarly distributed across levels of age and musical training.

I have already shown that there is no clear developmental trajectory of the strategies for representing music visually, but also evidence pertaining to the influence of musical training is contradictory. Some researchers have found that increased levels of musical training in children lead to more differentiated visualizations of sound and music (Reybrouck et al., 2009; Verschaffel et al., 2010) while others have found no effect of training (Hair, 1993/1994; Upitis, 1987). This suggests that a great deal depends on the nature of the task and the stimuli.

In one case, musical training even appeared to be detrimental to the accuracy of the visual representations (Davidson, Scripp, & Welsh, 1988). The authors asked more than 400 musically trained and untrained children, adolescents and adults to notate the two songs "Row, Row, Row Your Boat" and "Happy Birthday". Over 90% of the trained participants aged 12–18 years were unable to produce a correct conventional notation for the pitches of "Happy Birthday", while their invented notations showed fewer errors. Caused by what the authors called 'concept-driven errors', many trained participants assumed that the first and last note of "Happy Birthday" had to be the same and erroneously 'corrected' their invented notations too. However, a group of trained participants who had focused exclusively on learning to sight-read songs relied more on their perceptual abilities and made no conceptual errors, providing evidence that the kind of musical training children receive significantly affects their musical understanding.

Finally, a study investigating both children's visual and kinaesthetic responses to music is particularly pertinent here (Kerchner, 2000). Asking 12 musically trained and untrained children aged 7–8 years and 10–11 years to listen to the First Movement of Bach's Brandenburg Concerto No. 2 and to describe their listening experience verbally, by creating a 'listening map', and kinaesthetically—by moving their body—it was revealed that the most commonly addressed 'perceptual topics' included "instrument, register, continuous motion, formal sections, repetition,

[19] The Suzuki method has been developed by Japanese violinist Shinichi Suzuki after the Second World War (S. Suzuki, Mills, & Murphy, 1973). Likening musical training to language acquisition, its central tenet is that every child is able to acquire musical skills given the right environment and instruction. Generally, training of the ear and memorizing music is seen as more important than music reading skills.

dynamics, tempo, contour, and pattern" (pp. 36-37). What is more, the type of visualizations was dependent on age: the younger group created less differentiated mappings—drawing pictures, the contour or the instruments—whereas the older group used words and combinations of visual shapes to represent both extra-musical properties (e.g., mood) and musical parameters such as the beat. Regarding the kinaesthetic responses, both groups depicted a broad variety of musical parameters such as "beat, subdivided beat, articulation, melodic rhythm, embellishment, duration, style, phrase, subphrases and motivic fragments, contour, form, and pattern" (p. 42). Perhaps expectedly, both the visual and kinaesthetic responses were more differentiated than the verbal responses.

If the assumption that some musical experiences defy linguistic descriptions is correct, the same should hold for adults. Indeed some of the studies aimed at uncovering aspects of children's musical understanding through visual representations have included adult participants as well. Davidson and colleagues (1988) reported that invented notations of "Happy Birthday" by 7-year-olds are comparable to those of 10-year-olds and untrained adults. Moreover, it was revealed that children older than 9 years, as well as musically untrained adults, show very stable figural representations, while only participants able to read music display fully developed metric representations (Bamberger, 1982). And also Smith and collaborators (1994) found similar drawings of rhythmic sequences across groups of musically untrained children and trained and untrained adults.

To sum up, children use a great variety of visual shapes when asked to represent sound fragments and music in drawings, or—as many authors in the realm of developmental psychology have called them—invented notations. The choice of visual metaphors is probably mediated by age and the amount of musical training, but perhaps even more so by the experimental stimuli, design and wording of the instruction. Importantly, there is no reason to assume that adults' visual representations of sound and music should be any less revealing than those of children. However, only very few researchers, motivated by very different reasons, have attempted to shed more light on adults' drawings of sound and music. Motivations for studies with adults include the influence of type of music and listening situation (Hooper & Powell, 1970), investigating musical understanding (Gromko, 1995), the impact of musical training on visualizations of whole musical compositions (Tan & Kelly, 2004), cross-cultural comparisons of sound drawings (Athanasopoulos & Moran, 2013), sound tracings of sonorous-

gestural objects (Godøy, Haga, & Jensenius, 2006a; Haga, 2008) and embodied attuning to music (De Bruyn et al., 2012). These—except for the study on embodied attuning which I have already introduced in the section on embodied music cognition—will be reviewed in more depth in Chapter 3, as part of the introduction to my drawing study.

Thus far, a more extensive and systematic investigation of adults' drawings of sound and music is lacking – perhaps because drawing is widely regarded as a children's activity or possibly because epistemology in psychology as a social science is mainly based on quantitative approaches, rendering it difficult for open-ended drawing responses to be established and accepted as a valid measure (see also Chapter 2). However, as the review of children's drawings has shown, using the body as a natural mediator between physical sound and musical mind has the potential to provide many fresh insights into music perception and cognition. And this is exactly what I attempt to do when asking adult participants to draw along with sound and music (see Chapters 3 and 4). Before outlining the more general aims of this thesis, however, I will now turn to gestures and music.

1.5 Gesture

Unlike the substantial amount of research on children's drawings of sound and music in the field of developmental psychology—which is, to the best of my knowledge, the only field that has approached drawings of sound and music systematically so far—the context and origin of gestural representations of sound and music are much more diffuse. This has partly to do with the fact that 'gesture' has many connotations—possibly even more than 'shape'—and that gestures have been investigated from many different perspectives in various disciplines using different methodologies and definitions. Compared to drawings, gestures therefore need here an introduction with a wider angle. Gritten and King (2011, p. 1) identify numerous disciplines that have dealt with the concept of gesture, including "musicology, human movement studies, psychobiology, cognitive psychology, cognitive linguistics, anthropology, ethnology, music technology and performance studies." Indeed, it has been suggested that "[a] closer look at the term "gesture" reveals its potential as a core notion that provides access to central issues in action/perception processes and in mind/environment interactions" (Jensenius, Wanderley, Godøy, & Leman, 2010, p. 12).[20] Relating to work by Zhao (2001) and McNeill (2000),

[20] The same authors even go as far as to suggest that "the notion of gesture […] bypasses the Cartesian divide between matter and mind" (Jensenius et al., 2010, p. 13).

Jensenius and colleagues (2010) propose that research on gesture can broadly be viewed from three perspectives: gesture as communication, gesture for control, and gesture as metaphor. For the moment, I will neglect gesture for control, which is mainly based within the realm of Human-Computer Interaction (HCI), and focus on the other two viewpoints.

We often use hand gestures together with speech to emphasize a word, illustrate the shape of an object or point into a particular direction. These kinds of communicative gestures have been studied extensively, providing evidence that hand gestures—but also other types of physical gestures such as facial expressions—not only accompany speech but form an integral part of our thoughts and how we communicate (McNeill, 1992, 2005).[21] In the words of Jensenius and colleagues (2010, p. 15), "gestures and speech are co-expressive, or co-articulatory." This means that the communicative gestures we observe every day when interacting with other people play a crucial role in cognition. They are significant because they are able to carry the meaning of abstract ideas and might represent them sometimes even better than spoken words (Goldin-Meadow, 1999). By studying these hand gestures systematically it should therefore be possible to shed light on some covert cognitive processes (McNeill, 2005) – which is what I intend to apply for the case of music.

But before turning to representational (hand) gestures of music, it is important to consider metaphorical gestures too. The example I have given at the very beginning of this chapter—an ascending-descending melodic line—can be regarded as a metaphorical gesture, i.e. a gesture in sound. Such a gesture is not an observable body movement but rather part of a cognitive process when listening to music. As we have seen earlier on, it is very likely that our cognition of metaphorical gestures involves brain areas for motor action. That is, to be able to hear the melodic line as a musical gesture, it is probable that we simulate internally a physical gesture of the same trajectory. Although our motor output is usually inhibited, it can be easily disinhibited, i.e. when we start moving our bodies to music. Unsurprisingly, there is no consensus among music scholars regarding the nature of metaphorical gestures in music but a great many of them acknowledge, or have come to acknowledge, the role of the body and its interaction with the physical environment (as opposed to the fixed forms of musical symbols in music scores). For

[21] McNeill (1992) identified five functional categories of gestures: iconics, metaphorics, beats, deictics and emblems. He also defined a continuum of gestures—the so-called 'Kendon continuum' based on work by Kendon (1982)—ranging from spontaneous gesticulation over speech-linked gestures, emblems and pantomime to sign language. For a discussion on communicative gestures in a musical context, see also recent work towards a sign language for music (Fulford & Ginsborg, 2013).

Hatten (2006), “[m]usical gesture is biologically grounded, drawing on the close interaction of a range of human perceptual and motor systems that intermodally synthesize the energetic shaping of motion through time into significant events with expressive force” (Gritten & King, 2006, p. xxi). More explicit reference to conceptual metaphors and their grounding in bodily experience is made by Cox (2006) who emphasizes the role of performers’ movements and listeners’ imitations of the heard/seen sound-producing actions in music cognition, and by Johnson and Larson (2003) who, although not explicitly referring to musical gestures, discuss metaphors of musical motion that are based on our everyday experiences. Also Delalande (1988) seems to be in accordance with the embodied cognition research programme, proposing that musical gestures exist at “the intersection of observable actions and mental images” (Jensenius et al., 2010, p. 18). What this very brief overview seems to suggest is that studying gestures always involves some sort of overt or covert bodily gesture – whether these are communicative gestures or metaphorical gestures (in sound). The physical and metaphorical appear to be two sides of the same coin. I have already drawn on musicological research here for the purpose of describing metaphorical gestures within the broader research framework of gestures—communication, control, metaphor—as suggested by Jensenius and colleagues (2010), and will now shift the focus entirely to musical gestures.

In both of their edited volumes on music and gesture, Gritten and King (2006, 2011) showcase the richness of gesture-related research that exists within the field of music. In their more recent volume, they cover topics such as psychobiology, perception and cognition, philosophy and semiotics, conducting, ensemble work and solo piano playing. Note that most of these approaches use different concepts of musical gesture. Unless stated otherwise, I will use the term ‘gesture’ in the sense of a physical body movement hereafter.

One broad but useful distinction between musical gestures has been offered by Leman and Godøy (2010) who distinguish between ‘body-related gestures’ and ‘sound-related gestures’. Body-related gestures denote all the physical gestures that are carried out by humans, while sound-related gestures refer to gestures within the music such as melodic lines or rhythmic figures, i.e. metaphorical gestures in the sense described above. Leman and Godøy also emphasize that body-related gestures should not simply be reduced to movements. They need to be considered in light of the subjective experience and the musical context in which they occur. In that sense, body-related gestures carry meaning and are expressive of something,

unlike other types of movements such as grasping a doorknob (which could become a gesture, though, for example in a theatrical performance in which the act of grasping a doorknob is done in an expressive way, perhaps heralding the beginning of an important line of action.) But the distinction between body-related and sound-related gestures is just one of many ways in which music researchers have more or less consistently categorized and defined gestures.[22] Both types of gestures are relevant for the empirical studies in this thesis, as participants will be asked to represent one type (sound-related) by means of the other (body-related). At this point, it is worthwhile zooming in further into body-related gestures of music.

Jensenius and colleagues (2010, pp. 23-24) define four different functional categories: sound-producing gestures, communicative gestures, sound-facilitating gestures and sound-accompanying gestures. Sound-producing gestures are those necessary to produce sound on an instrument and can further be subdivided into excitatory and modificatory gestures. Communicative gestures are those meant to communicate something – either between musicians or between musician and audience members. Sound-facilitating gestures are those supporting the production of sound and can further be subdivided into support, phrasing and entrained gestures. Sound-accompanying gestures are those not part of the production of the sound itself but rather "follow the music" (p. 24). The latter category is the most pertinent here, as it refers to the sound-tracing gestures outlined in the Shape section above. These are the gestures that are intended to represent features of sound or musical excerpts and that have the potential to reveal what kinds of auditory features people attend to when asked to represent music with arm and hand gestures. Note that the empirical studies in this thesis will exclusively deal with musical gestures of listeners, that is, none of my participants' gestures will produce any musical sounds.

The literature on listeners' gestural representations of sound and music is sparse, and similar to drawing studies, researchers have had all sorts of motivations for their studies.[23] Interestingly, the first studies have been carried out with children as well. Espeland (1987) explored new methods of music listening aiming at making children's inner reactions visible and encouraging them to learn for themselves. As already shown above, Kerchner (2000) studied the effects of

[22] Another simple but effective classification of musical gestures distinguishes between gestures by those who produce sounds, i.e. performers, and those who perceive sounds, i.e. listeners (Jensenius et al., 2010).
[23] Note that there is a recent strand of empirical research concerned with free movements to music (i.e. dance) in adults (Burger, 2013; Thompson, 2012; Van Dyck, 2013) and children (Maes & Leman, 2013). I will not discuss this strand here further and focus instead on studies in which participants were specifically asked to represent some aspects of sounds and music gesturally.

age and musical training on kinaesthetic responses to a musical excerpt, and more recently, Kohn and Eitan (2009) investigated how children, aged 5 and 8 years, represent gesturally sound stimuli varied in pitch, loudness and tempo. Among the studies using samples of adults, music researchers have investigated how professional dancers and laypeople respond gesturally to a diverse set of musical excerpts (Haga, 2008), how participants represent pitched and non-pitched sounds by moving a rod in the air (Nymoen, Caramiaux, Kozak, & Torresen, 2011), how Western participants' gestural representations of traditional Chinese music develop over repeated listening sessions (Leman, Desmet, Styns, Van Noorden, & Moelants, 2009), how smooth and discontinuous hand gestures differ when participants are asked to represent sound stimuli varied in rhythmic complexity, pitch, loudness, brightness and attack envelope (Kozak, Nymoen, & Godøy, 2012), how participants represent gesturally action- and non-action-related sounds (Caramiaux et al., 2014) and how participants' free gestural responses to a classical piece of music match their linguistic descriptions of the expressive qualities of the music (Maes, Van Dyck, Lesaffre, Leman, & Kroonenberg, 2014). As with the relevant drawing studies, I will review all these in more depth in the introductory parts to my own gesture study (see Chapters 5 and 6). To conclude this section, I will point out some important differences between drawing and gesture approaches.

Free hand gestures are quite different from drawing gestures, even though one could see the latter as a subcategory, or special type, of the former. The most obvious difference is that free hand gestures occur in a three-dimensional space, whereas drawing gestures—though also happening in a three-dimensional world—are usually restricted to a two-dimensional plane. This difference is not trivial when studying people's bodily representations of sound and music. For instance, participants in an exploratory drawing study complained that they were unable to trace adequately the sounds on a graphics tablet and would have preferred gesturing in a three-dimensional space (Godøy et al., 2006a). Another important difference is that drawing gestures carried out with a pen usually leave a visible trace, whereas free hand gestures do not. The presence of visual feedback of one's own gesture enables one to keep track more closely of one's own movement, perhaps enhancing the memory for the musical sounds as well. These are issues to keep in mind when comparing findings from drawing and free hand gesturing studies with music. Having discussed drawings and gestures in response to sound and music, I will now outline the aims of this thesis.

1.6 Aims of the thesis

To recapitulate, I have introduced the wider theoretical background of my thesis work, pointing out that the traditional cognitivist view with the mind as a computer needs to be reconsidered in light of the advent of the embodied cognition research programme. From an embodied point of view, the interaction between the body and the physical environment is the core of all cognitive processes. Referring to the theory of conceptual metaphor, I have shown how higher cognitive functions such as language are grounded in bodily experiences and how behavioural and neuroscientific evidence accumulates that emphasizes the role of action in perception. I have discussed how the embodied approach has been formalized within musicology, opening up new opportunities for research. Importantly, I have defined the concept of shape for the purposes of this thesis, and shown how it fits within the embodied music cognition approach. Finally, I have provided an overview of how drawings and gestures have the potential to reveal covert aspects of embodied music cognition, highlighting the fact that systematic research with adult participants is lacking at the moment. Thus, there are two main aims of this thesis.

First, I intend to investigate how people perceive shapes in sound and music. More specifically, my aim is to examine how individuals map sound features and musical excerpts onto the visual and kinaesthetic domain by means of drawings and gestures, respectively. Using the terminology of Leman (2007), I intend to study people's embodied attunings to sound and music. What Leman calls 'graphical attuning' is thus the underlying concept in my drawing experiment. Although he does not specify a type of embodied attuning with free hand gestures, what I will do in my gesture experiment could perhaps be called 'kinaesthetic attuning'. Using the terminology of Godøy (2006), I envisage studying 'sound tracings'. Godøy uses this term without differentiating between drawings and gestures, so it covers both my drawing and gesture experiments.[24] To be sure, the empirical work carried out within this thesis is exploratory work. The hypotheses outlined in the introductory parts of my drawing and gesture experiments are by and large derived from more traditional experimental paradigms investigating cross-modal mappings due to the lack of more research on adults' drawings and gestures in response to sound and music. One important focus of my first aim is to investigate the influence of musical training. Since I have already shown that there exist differences in

[24] In fact, Godøy's exploratory sound-tracing study carried out with an electronic graphics tablet was more concerned with the gestures on the tablet, rather than the visual shapes the participants' gestures created during the experiment (Godøy et al., 2006a).

children's drawings of sound and music, it will be crucial to study the effects of more substantial musical training in adults – both in drawings and gestures. From an embodied viewpoint, learning to play an instrument involves years of highly specific body-environment interactions that leave a visible trace in the sensorimotor patterns of musicians' brains (Zatorre, Chen, & Penhune, 2007).[25] Based on theories such as the common coding principle or the ideomotor theory, I thus predict musicians to have formed different action-perception couplings from musically untrained individuals due to musicians' highly formalized engagement with sound and music, particularly through the interplay of actions (e.g., playing an instrument or conducting), sound (i.e. the sounds produced when playing an instrument or conducting an orchestra) and vision (e.g., musical notation or conductors' gestures). Differences between musically trained and untrained participants in cross-modal mapping tasks will also be highlighted in the introductory parts of my drawing and gesture experiments.

The second aim of this thesis is to explore methods and analysis techniques that take into account the active shaping—i.e. the active process of cross-modal mappings over time—of sound and music. If the body is a natural mediator between the physical environment and the musical mind, then studying how people represent visually or gesturally an ascending-descending melodic line, for example, should be carried out with the appropriate tools to capture goal-directed actions in real-time. In such a paradigm, the shape of a hand gesture is not means to an end but an end in itself. In the following chapter, I will outline my considerations regarding experimental paradigms and methods in some detail.

[25] By now, there is overwhelming evidence that musical training gives rise to both functional and structural changes in the brain (Hyde et al., 2009; Musacchia, Sams, Skoe, & Kraus, 2007). For a review, see Kraus and Chandrasekaran (2010).

Chapter 2: Paradigms, methods and analyses

2.1 Initial considerations

In this chapter, I will discuss my approach to capturing and analysing the perceived shapes of sound and music within an embodied music cognition paradigm. Since there are numerous ways to tackle such an endeavour, I will start by outlining some initial considerations that led me to pursue my thesis work with the methods and analytical tools applied in the following chapters. I will review some more traditional paradigms such as reaction-time protocols, forced-choice matching tasks and stimulus-response paradigms, before introducing the tools and software chosen for my own empirical studies and explaining the rationale of the main analytical technique. Since exploring different analytical techniques is one of the goals of this thesis, some of them will be explained in more detail in Chapter 4.

To begin with, it should be noted that a standardized procedure for measuring cross-modal shapes of sound and music does not yet exist. As I have shown in the previous chapter, researchers have reported a broad variety of motivations for conducting drawing and gesturing experiments – a variety that is mirrored in the breadth of experimental protocols. In a Kuhnian sense, the study of musical shapes by means of drawings and gestures is thus still in its pre-paradigmatic phase (Kuhn, 2012[1962]). Although the studies carried out thus far have (scientific) value, there are several, potentially conflicting theories and paradigms, and no consensus has been reached as to how one should go about addressing research questions systematically. This thesis work will hopefully give new impulses towards a paradigmatic phase, even though it is clear that more work will have to be done before 'normal science' begins.

I clarified at the beginning of Chapter 1 that large parts of the theoretical framework of my thesis can be situated within the cognitive sciences. Since I envisage that my thesis work may be considered in the same realm, it is necessary to point out some characteristic features of scientific theories. In general, a scientific theory should be lawful, accurate, generalizable and predictive. Importantly, what is implied here is that empirical findings should be replicable, i.e. researchers in different labs and at different times should arrive at the same result when following a certain experimental procedure.[26] These features of scientific approaches need to be

[26] It should be noted though that the statistical reality is much more complex: neither replicating an effect nor failing to replicate proves (or disproves) the truth of an empirical finding. For a more detailed discussion of replication within psychology, as well as some fundamental reconsiderations of statistical practice, see Cumming (2014). For a discussion of replicability as a post-publication evaluation see Hartshorne and Schachner (2012).

kept in mind when it comes to choosing appropriate experimental paradigms and analytical tools. However, they are not the only criteria that need to be considered for a systematic investigation of musical shapes. We have seen that subjective experience should not be dismissed but rather be part of a wider range of measurements that can be integrated into a scientific theory – as in Leman's second-person descriptions. Thus, the consideration of participants' verbal reports and subjective experiential ratings are seen as complementary to the objective measures deployed in my experiments. The crucial point is to develop a set of measures that researchers can apply independently by following clearly specified guidelines. This should increase the comparability of results and encourage researchers to attempt replications.

Before introducing some traditionally used tools and methods, I shall discuss an important conceptual distinction that is relevant for the choice of paradigm and analytical techniques for my own experiments. I have pointed out in Chapter 1 that it is vital to capture not only the final shapes but also the process of shaping. This distinction between product and process is not new and appears in various research contexts.

2.2 Product and process

One notable instance of emphasizing the difference between products and processes brings us back to children's drawings. Aiming to establish developmental psychology as a discipline that is vital for our understanding of human cognition—and not just a neat add-on telling us when certain cognitive functions emerge—Karmiloff-Smith (1992) discusses the role of the child as a notator. She observes a crucial difference in the way toddlers approach drawing and writing tasks:[27]

> "[t]he toddler goes about the processes of writing and drawing differently, even though the end products sometimes turn out similar. It is essential to distinguish between product and process, because toddlers' notational products may at times appear domain-general to the observer whereas their notational intentions and hand movements bear witness to a clear differentiation that they have established between the two systems" (Karmiloff-Smith, 1992, p. 144).

[27] Since most toddlers cannot write yet, they are often asked to pretend writing in such experiments.

What is relevant for my purposes is the remark that different processes can lead to the same (category of) product. In terms of drawings of sound and music, this means that studying only the final products may neglect some important features of the cognitive process that can only be disclosed by investigating the act of drawing itself. Note that paying attention to the process of drawing—or gesturing, for that matter—can be achieved with objective measures as well as the involvement of more subjective, experiential feedback. Kerchner (2000, p. 40) chose the latter option when examining children's listening maps:

> "[the children's] mapping *product* was not the only piece that shed light on the details of their perceptions. I relied on reviewing their drawing *process*—the children's verbal descriptions of their maps, and the pointing and listening tasks—to clarify my speculations."

These two examples from developmental psychology have in common that they explicitly point to a distinction between products and processes. In the first case, the motivation is to differentiate the underlying cognitive processes of drawing and writing, respectively, while in the second case—which is particularly relevant in the context of this thesis—the focus is on mapping music onto the visual domain. The tacit assumption of the latter approach is that music itself is a process rather than a product. Given my conceptualisation of music as sound rather than text for the purposes of this thesis (see Chapter 1), it makes sense to think of music as a process unfolding over time. Indeed, this is how sounds are experienced in general – as a temporal event. Although this should (hopefully) seem intuitive from the standpoint of cognitive science, musicology has long dealt with music as a product. Cook (2001) specifically addressed this issue, illuminating the ways in which musicians, musicologists and philosophers have used both notions to conceptualise music – before arriving at the conclusion that music should rather be seen and studied as performance. What is important to note here is that the perception of music viewed as a process should be done justice to by using experimental tools that allow for capturing that very process. As mentioned above, a process is time-dependent, whereas a product is not. That is not to say that the product is irrelevant. Particularly in the case of auditory-visual mappings of sound and music by means of drawings, the resulting product should be part of the analysis. Nevertheless, it seems that traditional accounts of cross-modal mappings of sound (Marks, 2004; Spence, 2011) have relied extensively on the product – at the

expense of the process. In the following section, some of these traditional paradigms will be introduced.

2.3 Traditional experimental paradigms of cross-modal mappings

There is a large number of empirical studies investigating how auditory stimuli are mapped onto the visual and visuo-spatial domains (for reviews see Eitan, 2013a; Marks, 2004; Spence, 2011). In order to be able to illuminate causal relationships, in many of these studies the auditory stimuli comprised static or dynamic pure tones as opposed to musical excerpts, which are harder to control for in rigorous experimental settings. However, cross-modal studies carried out with real musical excerpts usually draw on this vast empirical body of studies when formulating hypotheses regarding the cross-modal perception of musical features such as pitch and loudness. It is thus important to review the experimental paradigms underlying the vast majority of such empirical findings to be able to put embodied approaches using drawings and gestures into context.

To a large extent, our advancement of knowledge of cross-modal correspondences is based on reaction-time paradigms that were developed by Garner in the 1960s around the same time as the cognitive revolution gained momentum, with the underlying metaphor of the human mind as a computer processing incoming information. According to this view, sensory input from different modalities is integrated at various levels of processing ranging from early sensory/perceptual levels to late semantic levels (for a review, see Marks, 2004). The speed with which this processing occurs can be measured in behavioural experiments in which participants respond to features of a dimension of a modality by pressing buttons which have been assigned certain feature values. In the simplest case, there is only one modality involved and features are varied only along one dimension. For instance, participants may be asked to indicate as quickly as possible whether the pitch (i.e. the relevant dimension) of a sound is high or low, while the loudness (i.e. the irrelevant dimension) is kept constant. This task—which has been termed *speeded identification*—often serves as a baseline condition, involving two possible stimuli and two possible responses. If the irrelevant dimension is varied as well (e.g., loudness: soft and loud), we get four possible stimuli (high/soft, high/loud, low/soft, low/loud) while the number of possible responses is still two. In the latter scenario—which has been termed *speeded classification*—participants' task is to ignore the variation in the irrelevant dimension (i.e. loudness) and indicate the feature value (high vs. low) of the relevant dimension

(i.e. pitch). While these examples concern a single modality, there is extensive research combining dimensions from several modalities (Spence, 2011). Whenever there are greater reaction times in comparison to a baseline condition due to the variation of features in an irrelevant dimension or stimulus, this is referred to as *Garner interference.* On the other hand, whenever features from two dimensions—whether within a single modality or across modalities—are aligned congruently (e.g., high pitch – high elevation) such that the pairing gives rise to smaller reaction times in comparison to incongruently aligned features from the same two dimensions (e.g., high pitch – low elevation), this is referred to as *congruence effect*.

In such reaction-time experiments it is important to either balance the position of the response buttons across participants or manipulate it deliberately as a further independent variable. This is due to the well-studied effects of stimulus-response compatibility (Fitts & Seeger, 1953). Approaches using stimulus-response compatibility represent another classic paradigm within which one may study cross-modal correspondences. Crucially, the role of the participants' actions (button presses) in a given environmental setting (the arrangement of the buttons) becomes an integral part of the cross-modal mapping. Such an approach may therefore be called embodied, as it emphasizes the interaction between the body and its physical environment.[28] For instance, in an experimental setting in which the two response buttons for high and low pitch are arranged vertically, a high pitch is classified as 'high' more quickly when the corresponding button is the upper rather than the lower one (Rusconi, Kwan, Giordano, Umiltà, & Butterworth, 2006).

Besides the development and refinement of tasks involving speeded responses, there is an even older type of paradigm concerned with unspeeded responses. In fact, most of the early cross-modal mapping experiments consisted of unspeeded tasks, e.g., asking participants to locate sounds with different discrete pitches in space (e.g., Pratt, 1930; Trimble, 1934). Another commonly observed unspeeded task is forced-choice matching. When employing such a paradigm, individuals are asked to choose from a limited set of responses—there may be several but in some cases as few as two—the one they think fits best with a stimulus presented. For instance, in a classic series of experiments, Walker (1987) asked people to match pure tones varied in frequency, amplitude, waveform and duration with abstract visual figures varied in vertical and horizontal arrangement, size, pattern and shape. But also real musical excerpts

[28] It should be noted though that the limitation of actions to simple button presses also limits, at least to some extent, the "embodiedness" of such an approach.

and prints of paintings have been used in one of the earliest empirical studies in which participants were asked to match music to pictorial representations (Cowles, 1935).

More recently, Eitan and Granot (2006) developed a paradigm in which participants were asked to imagine the movement of a cartoon character in response to musical stimuli varied in pitch, dynamics, tempo and articulation and to indicate the type, direction and speed of the character's movement on rating scales. Although such an approach may at first sight appear opposed to the embodied cognition research programme, as it does not involve any overt body movement at all, a closer look reveals that imagining body movements—probably involving internal simulations of motor actions—might come closer to embodied approaches than any traditional reaction-time tasks. Eitan and colleagues successfully applied their paradigm in various research contexts ranging from studies with children (Eitan & Tubul, 2010), over blind participants (Eitan, Ornoy, & Granot, 2012) to more complex musical stimuli (Eitan & Granot, 2011).

2.4 Towards embodied experimental paradigms of cross-modal mappings

All paradigms described thus far have in common that participants' responses are fairly restricted: they are asked to press one of two buttons as quickly as possible, match a sound excerpt with an (abstract) visualization from a given set of visual metaphors or indicate on a rating scale to what extent certain sound features correspond to the visual and visuo-spatial domains. While this allows researchers to investigate cross-modal mappings of sound and music rigorously by refining their paradigms and manipulations further and building onto an ever-increasing body of evidence, the rigour comes at the cost of richer, qualitative data which provide another fruitful angle on the object of study. What is more, the role of action in perception is—apart from the stimulus-response compatibility approach—mostly neglected. In an embodied approach, the interaction between the body and the physical environment is crucial. In this view, it is this kind of interaction that formed cross-modal correspondences in the first place. Hence, epistemologically, it should make sense to attempt to (re)create scenarios involving overt bodily responses—such as drawings and gestures—for the purpose of studying these cross-modal mappings empirically. Free drawing and three-dimensional bodily gestures should thus enable individuals to engage more naturally with musical sounds, representing an

embodied approach to cross-modal mappings that does not only open up new pathways for inquiry but also enhances the ecological validity of experimental procedures.

Needless to say, the newly acquired freedom comes together with the necessity of adapting existing, as well as developing novel, analytical techniques, which perhaps poses the greatest challenge for researchers opting to capture the perceived shapes of sound and music in an embodied approach. However, before turning to the analytical techniques, it is equally important to consider the tools that capture the data, for these tools determine what kinds of data researchers have to deal with in the analysis stage. As outlined above, second-person descriptions of drawings and gestures should include objectively measurable data, preferably with high precision to satisfy the 'accuracy' criterion of scientific theories. In addition, the experimental tools must allow for time-dependent data acquisition, as I envisage investigating the processes of drawings and gestures in response to sound and music, as well as the resulting products. Whereas in the case of drawings the product is the result of the shaping process, i.e. the result of drawing the perceived cross-modal shapes of sound and music, in the case of free three-dimensional gestures there is no such thing as a product because a gesture usually does not leave a visible trace. Is there a way to study the product of gestures despite this? I will return to this question in the section on software below and focus for now on the experimental tools.

2.5 Experimental tools

2.5.1 Drawing

I have emphasized that both the product and the process of drawings are worth studying. Traditional paper-and-pencil approaches are adequate when one is interested in the resulting product but make it hard to study the process of drawing or the drawing gesture. One possible solution could be to film the act of drawing—perhaps with a camera attached to the ceiling to enable a bird's eye view—and later analyse the video footage. This would require, however, many decisions taken independently by the researchers in the stage of data pre-processing, rendering it difficult to arrive at a standardized procedure. Moreover, researchers interested in motion features of the drawing gestures would need additional tools to extract objective data from the video. While there is a growing community of (music) researchers interested in extracting objective movement data from videos—see, e.g., the EyesWeb platform (Mancini, Glowinski, & Massari, 2012)—for my purposes, it is more efficient to obtain the data directly

from the drawing process itself. Due to technological advances this is now possible by means of electronic graphics tablets. In fact, such a tablet has been used before in an exploratory study of sound tracings (Godøy et al., 2006a; Haga, 2008). Electronic graphics tablets exist in varying sizes and are very similar to analogue types. The main difference is that the pen does not leave a trace on the tablet itself, as is the case when using paper and pencil. However, since the tablet is connected to a computer, the trace of the pen can be made visible with the appropriate software (see below) on a screen in front of participants. For the purposes of my empirical investigation (see Chapters 3 and 4), I have chosen the commercially available electronic graphics tablet Wacom™ Intuos4™ L. With a drawing area of 325 x 203 mm^2, it is the second largest graphics tablet in the Intuos4™ series and should provide participants with sufficient space to draw. This tablet has a high sampling rate—maximum 200 points per second (Wacom™, 2011)—which is important to ensure that the software does not miss a drawn point by virtue of the tablet sampling too slowly. In addition, to capture subtle positional changes, a high resolution is required in both space and pressure sensitivity: this tablet has 5080 lines per inch and 2048 levels of pressure (Wacom™, 2011). The pressure sensitivity was a special feature used in my experiment to change the size of the pen stroke, with more pressure resulting in a thicker line.[29] It was deemed important to be able to change the size of the line instantaneously—without having to go back to already drawn parts—since visual size is a well known feature involved in auditory-visual correspondences (see, e.g., Spence, 2011).

2.5.2 Gesture

The study of human movement is an ancient scholarly topic – for a historic overview of movement studies from Ancient Greece till the present day, see Thompson (2012, pp. 40-43). Fuelled by technological advances, the past fifteen years have seen a steep increase in measuring human beings' body movements for scientific purposes. In music research, there are now various labs (e.g., in Ghent, Jyväskylä and Genoa) carrying out systematic studies on the relationship between music and movement. The most commonly applied motion-capture systems use infrared optical motion tracking. In order to measure the position in a three-dimensional space, individuals are required to wear markers on their body, the positions of which are tracked by several cameras. As explained by Thompson (2012, p. 43),

[29] This tablet also offers the possibility of measuring the tilt, which could have been used to manipulate the colour or saturation of the pen stroke. However, since manipulating the thickness of the stroke with the pressure applied to the pen already introduced a deviation from "normal" drawing, I tried to avoid participants having to control too many features in unfamiliar ways and decided not to use the tilt function.

> "[…] each camera within the network captures two-dimensional coordinates of reflective markers (affixed to actors) by flashing quick pulses of infrared light within the capture volume. When a marker is tracked by at least two cameras, the system is able to triangulate the marker's spatial displacement in three-dimensions."

While such a setup allows researchers to acquire movement data with high temporal and spatial resolution, the disadvantage is that such a system is very expensive and not very flexible.[30] This is why researchers have started to seek cheaper and more flexible alternatives. With a recent boom in applications involving whole-body movements, the gaming industry offers such alternatives. For the purposes of my study, Microsoft® Kinect™ and a Nintendo® Wii™ Remote Controller were identified as appropriate tools to study people's gestural representations of sound and music (see Chapters 5 and 6). The Microsoft® Kinect™ sensor is equipped with an RGB camera, an infrared emitter and an infrared depth sensor.[31] Technical details and functioning of the sensor are explained on the Microsoft® website.[32] According to this source, the RGB camera

> "stores three-channel data in a 1280 x 960 resolution at 12 frames per second, or a 640 x 480 resolution at 30 frames per second. This makes capturing a color image or video possible. […] The emitter emits infrared light beams and the depth sensor reads the [infrared] beams reflected back to the sensor. The reflected beams are converted into depth information measuring the distance between an object and the sensor. This makes capturing a depth image possible."

Apart from the lower price, the Microsoft® Kinect™ system has the advantage that participants do not have to wear markers on their body and that it can be set up straightforwardly in different environments. Additionally, a Nintendo® Wii™ Remote Controller was used to obtain acceleration data, and to be able to record fast hand movements. Both devices were controlled with custom-made software, written specifically for the purposes of the gesture experiment.

Since my thesis work is exploratory I decided to film the experiment with two standard video cameras (Panasonic HDC-SD 700 and 800) to get a richer set of data for the analysis.

[30] It is vital to have motion capture labs, enabling researchers to test hypotheses empirically. At some point, however, it will be necessary to test the laboratory findings in the field, e.g. in a club or at a concert. For such settings, alternative motion capture systems need to be developed.
[31] Although it also contains a multi-array microphone and a three-axis accelerometer, these features were not used in my gesture experiment.
[32] http://www.microsoft.com/en-us/kinectforwindows/discover/features.aspx

Moreover, the video footage enabled me to check the data acquisition with Microsoft® Kinect™ and Nintendo® Wii™ Remote Controller.

2.6 Software

Both the drawing and gesturing experiment were carried out within a programming language called Processing (http://processing.org/). Launched in 2001, Processing is now a professional development environment that is widely used by (visual) artists, designers and researchers. Nicolas Gold developed the program for the drawing experiment. The rationale for choosing Processing is described in Küssner, Gold, Tidhar, Prior and Leech-Wilkinson (2011, p. 2):

> "The ease of constructing graphics-based applications in Processing (Fry & Reas, 2011) led to its selection as a development platform. Two supporting libraries were used: JTablet2 (Bastéa-Forte, 2011) to provide a simple interface with the graphics tablet, and Minim (Di Fede, 2011) to provide multi-threaded audio playback capabilities within the Processing program."

The same authors describe the architecture and functioning—including some prototype features—of the Processing program as follows (Küssner et al., 2011, p. 2):

> "Processing programs (termed "sketches") are typically divided into two primary methods: `setup()` and `draw()`, with the former being executed once and the latter once per frame. In this application, `setup()` deals with basic issues such as window size, drawing initialization, and setting up tablet communication. The operational mode desired by the experimenter is not handled here but using a flag variable in the `draw()` method. If this flag is true, a list of experimental options is displayed. These allow the selection of drawing mode (whether to interpolate between points), drawing sensitivity (whether the pen must press on the tablet to draw), music playback (whether the drawing takes place while music is to be played), and whether replay of a previous trace is required. Information about audio stimuli, replay traces and so forth is stored in a configuration file prepared by the experimenter."

For a screenshot of the Processing program, see Figure 2-1 below. The possibility of interpolating drawn points and drawing along with a replay of a previous drawing were two functions tested in pilot experiments but not implemented in the final experimental procedure.

However, I will report here, and cite at length, the full version of the program to provide an idea of the initial considerations and challenges involved. Küssner and colleagues (2011, p. 2) note:

> "On starting the experiment using a key-press to change the flag variable, music playback is started (if appropriate) and the program alternates between *drawing frames*, where tablet input is polled (this is not an event-driven approach) and a point drawn if appropriate, and *replay frames*, where the next point from a pre-existing trace is drawn in a different colour if the appropriate time has been reached. Processing's frame rate was set to 48 frames/sec thus allowing 24 frames/sec each for drawing and replay. There are advantages and disadvantages to such an approach. 24 frames/sec is sufficient to provide an illusion of continuous motion for both drawing and replay. However, to avoid latency issues it is important that any computation takes place in less than the time taken for one frame. It is possible that a replay point could be drawn up to one frame-duration late if its due-time occurs during the preceding drawing frame. Assuming Processing is able to achieve the desired frame-rate on a given system (which is not guaranteed), this is actually a likely scenario since, for example, a point captured 10 drawing frames into a stimulus must wait to be replayed until at least 10 drawing frames have passed. This approach (plus the multi-threading of the audio library) means that no hard real-time guarantees can be given about the application's response. Empirically, minor decreases in frame-rate have been observed and this is accounted for in subsequent data analysis. If replay-priority were deemed more important than drawing, the frame-ordering could be reversed to allow accurate plotting of replay points at the expense of a single frame delay in drawing capture. The advantage of an equal division between the two types of frame is that in situations where simultaneous drawing and replay occurs, neither one can starve the other of a drawing opportunity. It also assists with debugging in the absence of techniques for easily slicing code on the basis of frames (Gold, 2011). Audio stimuli can be replayed by the experimenter without requiring a new trace. Output is stored as both an image file representing the final state of a trace, and a comma-separated values (CSV) file containing experimental metadata and a time-stamped list of points drawn. This file can be used for subsequent replay."

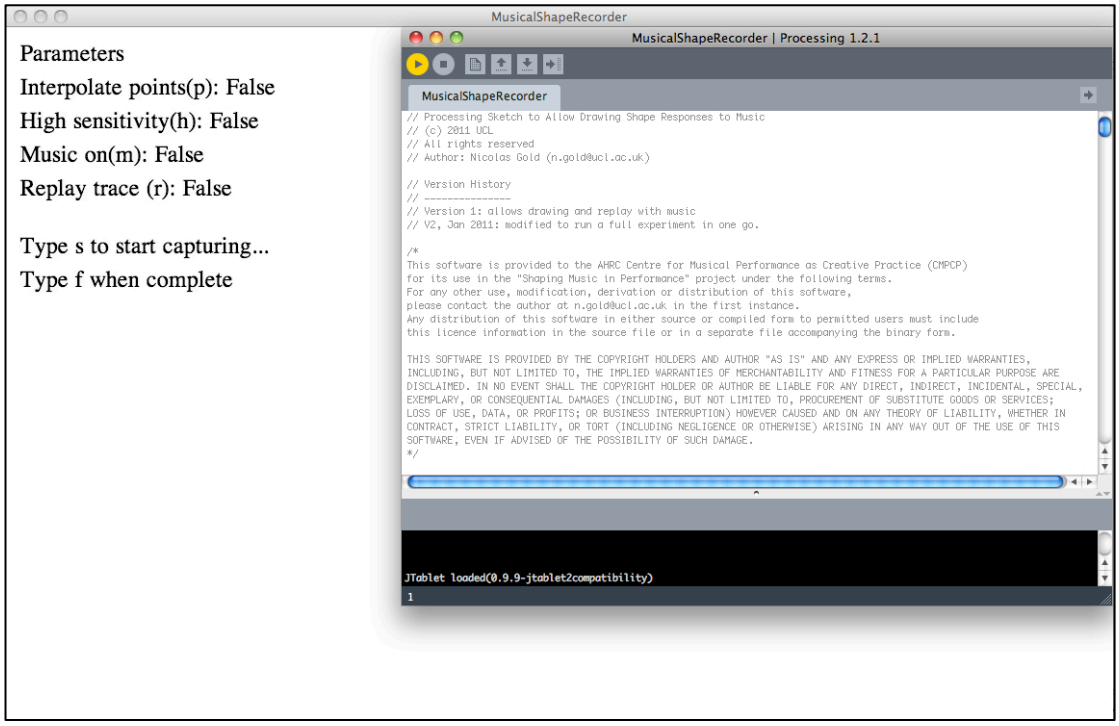


Figure 2-1 Screenshot from shape-capturing program for drawing

Based on the Processing program for drawing, Dan Tidhar developed the Processing program for the gesture experiment. As described above, the main tool for capturing the movement data was Microsoft® Kinect™, which was operated by custom-made software based on OpenNI™ (OpenNI™, 2011), an open source software development kit for natural (bodily) user interfaces.[33] One special feature of the Processing program was the option of visualizing the hand gesture on a screen in front of the participants. The visualization consisted of a black disk with green edge that could be moved across the screen by moving one's hand accordingly (see Figure 5-3). Movement along the z-axis resulted in changing the size of the disk; moving forward increased, and moving backwards decreased, the size of the disk. Moving the disk on the screen left a trace that decayed after ca. 4 seconds (see also Methods of Chapter 5). This option was seen as a way of obtaining a 'shaping product' of gesturing similar to the drawing product of the tablet (though in the latter case, the trace did not decay). Pilot testing also showed that participants were more motivated when they were able to 'draw' a trace with their hand gestures. Hence this option was implemented in the final procedure of the gesture experiment.

[33] More information can be found at http://en.wikipedia.org/wiki/OpenNI.

2.7 Experimental stimuli

Since all experimental stimuli will be described in detail in Chapters 3, 5 and 6, I will only point out some general considerations here. As my review of children's drawings in the previous chapter has shown, there might well exist an ontological gap between simple sound stimuli and real musical excerpts in the sense that a sequence of pure tones increasing in frequency might be processed differently from an ascending melodic line in the context of a symphony. It is thus important to study both kinds of auditory stimuli. However, since there is to date no systematic investigation of cross-modal shapes of sound features by means of drawing and gesturing, my focus will be shifted slightly towards simple sound stimuli, as they are easier to manipulate systematically. Moreover, when introducing novel experimental paradigms such as my drawing and gesturing approaches, it is helpful to keep some familiar elements such as the choice of auditory stimuli. As pure tones have been studied extensively in cross-modal experiments, this will facilitate the assessment of my experimental approaches by comparing them to traditional paradigms. Nevertheless, I will include some real musical excerpts in both the drawing and gesture experiments to ensure more ecologically valid listening situations are covered. The type of stimuli—pure tones or extracts from musical pieces—will also, to some extent, influence the analytical tools. In the remainder of this chapter, I will outline some general considerations regarding analysis procedures in (music) psychology and the analytical tools of this thesis.

2.8 Analysis

2.8.1 Some reflections on data handling and testing in (music) psychology

Quantitative analyses in music psychology—or empirical/systematic musicology, for that matter—are usually driven by prevailing standards in the broader realm of psychology. Müllensiefen (2009) provides an overview of recent statistical tools applied in music psychology and music modelling, and encourages researchers to make use of the full breadth of tools available. In most cases, however, analytical techniques are based on the General Linear Model (GLM), with its various well-known applications such as *t*-tests, Analysis of Variance (ANOVA) or linear regression.[34] Indeed, null-hypothesis significance testing has been the norm in psychology for such a long time—almost a century now—that until very recently it seemed hard to imagine that there will ever be anything else. Even though several attempts have been

[34] I will not explain these statistical tests in detail here. For more information regarding the standard repertoire of statistical tests in psychology see Field (2009).

made in the past few decades to introduce new statistics, it may only be now that the tide is turning (Cumming, 2014), powered by the disclosure of prominent cases of fraud, which—bad enough in their own right—have shifted public attention to some general research (mal)practices in science (Ioannidis, 2005), and psychology in particular (Simmons, Nelson, & Simonsohn, 2011). While there are numerous issues such as under-powered studies (Button et al., 2013) or a publication bias (Ferguson & Brannick, 2012), the core problem is the conventional, and highly arbitrary, probability value of 5%, which his inventor, statistician Ronald Fisher, never intended to be used the way it is used nowadays (Nuzzo, 2014). To put it bluntly, if the significance threshold is reached then everything is fine (and an important step towards publication has been achieved); if the *p* value is just over the threshold however, researchers have been found to engage—deliberately or subconsciously—in what has been coined '*p*-hacking' (Simonsohn, Nelson, & Simmons, 2013).[35] *P*-hacking refers to questionable procedures such as monitoring data during collection, excluding outliers or testing more participants until the significance threshold is reached. Unfortunately, *p*-hacking seems to be rather the rule than the exception.[36] If these data manipulations are not reported in the methods section—and they rarely are—the consequence is that effects get published that at best appear stronger than they really are and at worst do not exist at all. Not surprisingly then, this is one of the reasons why researchers fail to replicate their own or other people's findings – a research activity that is crucial in natural sciences such as physics for an effect to be published in the first place.

For the purposes of this thesis, I will not abolish null-hypothesis significance testing altogether. Although there is some serious progress in psychology now, many people will still expect to see *p* values, and in some cases, there might not yet be suitable alternatives available. But I will try to deemphasize the significance threshold by reporting results that are slightly above it. More importantly, whenever possible I will report effect sizes and confidence intervals, as suggested by Cumming (2014). These have been standard in medical research since the 1980s, as the reported effects in this field—e.g., of a new treatment—are often matters of life and death. If (music) psychology wants to be taken seriously as a science—which is not to say that

[35] Others have called this activity "data-dredging, snooping, fishing, significance-chasing and double-dipping" (Nuzzo, 2014, p. 152).
[36] There is empirical evidence from three leading scientific journals that the frequency of reported *p* values on, or just below, the significance threshold is unproportionally large (Ridley, Kolm, Freckelton, & Gage, 2007), and new methods have been developed to detect such 'fiddling' with empirical results (Gadbury & Allison, 2012). See also the *p*-curve (Simonsohn et al., 2013).

quantitative approaches are any better or more valuable than qualitative ones—the same seriousness needs to be displayed in the analytical tools and procedures, otherwise the use of statistics is pointless.

Having raised the bar for scientific music-psychological research, I should point out once again that I do not regard any intrinsic characteristics of scientific approaches as more valuable than those of other, perhaps more humanities-driven, approaches. In fact, second-person descriptions attempt to combine the best of both approaches, aiming to create a rich picture that is based on subjective experience *as well as* objectively measurable data. What is more, the exploratory nature of my thesis work requires some flexibility regarding methods and analytical tools. In the following section I will introduce and discuss the main analytical tools of this thesis.

2.8.2 Non-parametric correlations

One of the key questions to be addressed in this thesis is how the drawings and gestures relate to, or represent, the features of the sound and musical excerpts. As discussed in Chapter 1, this can be achieved by correlating second-person descriptions (drawings, gestures) with third-person descriptions (auditory features). It is important to note, since this affects the choice of analytical tool, that both types of descriptions consist of time-dependent data. Most parametric correlation tests assume the observation of data points to be independent. In the case of time-dependent data (gestures, sound features), however, this independence is violated because any value $\mathbf{x}_{t+1}$ will usually be dependent on the value $\mathbf{x}_t$. In non-mathematical terms, the position of one's hand at any moment in time will depend on the hand's position just prior to that moment. Similarly, the loudness of a musical excerpt at any moment in time will depend on the loudness just prior to that moment.[37] This presents us with the problem of so-called serial correlation or autocorrelation, where values of the same variable are correlated (i.e. dependent). Another problem with time-dependent data (also sometimes referred to as time series or processes) is the lack of stationarity. Stationarity of a process denotes the constancy of statistical parameters such as mean and standard deviation over time. It is very likely that time-dependent data based on gestures and drawings in response to sound and music are non-stationary: in fact, it is expected that there will be variation of the amount and direction of movement over time. Vines and colleagues (2006, p. 87) summarize the issue of correlating time-dependent data as follows:

[37] There are exceptions, of course, such as the first moment of a *fortissimo* passage after moments of silence.

> "Techniques do not presently exist for accurately calculating the correlation between two functions that are themselves serially correlated and non-stationary (i.e. for which consecutive values are dependent over time in a non-predictable way)."

To explain how people have tackled this problem, it is necessary to provide some more general information about correlations. What researchers usually refer to when talking about correlations is Pearson's product-moment correlation coefficient *r*. Pearson's *r*—which is a parametric statistic—captures the linear, and only the linear, relation between two independent variables *X* and *Y*, which need to be at least on an interval scale.[38] The level of measurement (i.e. interval scale) is crucial to be able to interpret the absolute size of the correlation coefficient. Using lower levels of measurement such as an ordinal scale makes the interpretation of the absolute size impossible because the variance of such scales is not defined.[39] The level of measurement is no trivial statistical detail – even though it is often treated as such. It determines what can and cannot be said about data. For instance, there seems to be a tacit consensus among (music) psychologists that ratings on Likert scales should automatically be classified as interval scales, even though on closer inspection one might arrive at a different conclusion (cf. Field, 2009, p. 8). Take ratings of pleasantness on a scale from 1 (very unpleasant) to 5 (very pleasant) as an example. The tacit assumption is that the difference of the subjective pleasantness between, say, 1 and 2 is half the difference of the subjective pleasantness between 3 and 5. The problem is that there is no way to prove (or disprove) this assumption, which is why researchers—for reasons of convenience of choice of statistical tests—mostly proceed with parametric tests.

To return to the initial problems of autocorrelation and non-stationarity, Vines and colleagues (2006) provide three statistical approaches to overcome this problem: first, if researchers intend to use parametric correlation coefficients, they should at least make sure that the autocorrelation is removed as much as possible (see pp. 89-90 of their article for a practical example). Secondly, instead of using correlations of variables, it might be worthwhile in some research contexts to treat the observed data as functions and apply a technique called Functional Data Analysis (Levitin, Nuzzo, Vines, & Ramsay, 2007). I will explain in Chapter 4

[38] Interval scale means that the ratios of differences—but not the differences themselves—are meaningful. The classic example for an interval scale is temperature (in degree Celsius). It is possible to say that the difference between 10˚C and 20˚C is twice as much as the difference between 5˚C and 10˚C, but it is not possible to say that 20˚C is twice as much/warm as 10˚C.

[39] An example of an ordinal scale is a (sports) table. Given the rank positions alone (i.e. disregarding gained points), it is only possible to say that the 3rd rank is better than the 4th rank, or that the 9th rank is worse than the 8th rank, but nothing meaningful can be said about the size of the differences, which would be necessary to calculate the variance.

why Functional Data Analysis is less suitable for my set of data, and explore alternative analyses (Gaussian processes) that treat the observed data as functions. Thirdly, the authors suggest the use of non-parametric correlation coefficients, such as Spearman's rank correlation coefficient ρ. Since I chose the latter option for my analysis, I will discuss the use of this statistic in some more detail here.

The first of two steps in calculating Spearman's rank correlation coefficient ρ is to rank the data and substitute the original values of the variables with ranks (for a more detailed description, see Field, 2009, p. 542). In case of rank ties—i.e. two or more identical values in the data—the average of these ranks is assigned to each of the tied scores. The second step—the computation of the correlation coefficient—is then the same as for the parametric coefficient Pearson's *r* (see Field, 2009, p. 170). Schubert (2002) was arguably the first—at least in the field of music psychology—to suggest non-parametric correlation coefficients as a means to overcome the issue of autocorrelation and non-stationarity in time-dependent data. As he correctly points out, one should be cautious with interpreting the absolute size, as well as the significance value, of non-parametric correlation coefficients, and with comparing it to non-parametric correlation coefficients from other sources. I have shown above why the absolute size is never interpretable, regardless of whether the underlying data are time-dependent or not. However, it is possible to compare various correlation coefficients *with one another*, i.e. their relative size, when they come from the same source. Moreover—and this is why my digression into levels of measurement was necessary—I will treat the size of non-parametric correlation coefficients as measurements on an interval scale. By doing this, I assume that the difference between, say, ρ = .50 and ρ = .60 is twice the difference between ρ = .30 and ρ = .35. This assumption might be debatable but given the lack of better analysis techniques to date I will proceed with this assumption.

2.8.3 Other analytical techniques

As outlined in Chapter 1, one of the aims of this thesis is to investigate the data from various angles with different analytical tools. To that end, I will explore and evaluate the dataset of drawing responses to sound and music with more advanced mathematical techniques in Chapter 4. This will include modelling techniques, clustering approaches and classification analyses. Apart from these more advanced tools, I will use more conventional statistical tools throughout Chapters 3, 5 and 6. These range from ANOVAs and *t*-tests over chi-squared and

binomial tests to (parametric) correlation analyses, and their use will be explained in the context of the respective analyses. Moreover, I will apply more qualitative approaches, particularly in Chapter 7 when analysing gestural responses to music. Keeping in mind that the study of cross-modal shapes of sound and music is still in its pre-paradigmatic phase, it is important that various analytical techniques are tested at this stage. Even though some of the analytical tools applied here may turn out to be insufficient, I hope this will motivate other researchers to join the exploratory phase (cf. Nymoen, Godøy, Jensenius, & Torresen, 2013) in order to improve existing, as well as develop new, analytical tools with a view to exhausting the full potential of data acquired through drawing and gesturing responses to sound and music.

2.9 Summary and conclusion

In this chapter I have outlined the methods and tools for investigating cross-modal shapes of sound and music. As part of my initial considerations, I have emphasized the notions of product and process and shown how these can be linked to my thesis work in the form of second-person descriptions. I have reviewed traditional experimental paradigms that have advanced our knowledge of cross-modal mappings considerably and highlighted some requirements towards embodied experimental paradigms. I have introduced the specific tools to be used for capturing drawings and gestures, and described the custom-made software as well as the auditory stimuli and their implication for the analysis. I have discussed the current state of data handling and analysis in (music) psychology, followed by a close look at non-parametric correlations and other analysis techniques to be applied in this thesis.

Although the credibility of psychology—and thus that of music psychology as well—has been at stake in the past few years, I should hope that the patient is beyond the critical state. The important take-home message is that methods and statistics matter: in fact, they make all the difference. Hence researchers need to be scrupulous when reporting how they collected and analysed data to enable other researchers to replicate findings. More importantly, methods and statistical tools should be regarded as flexible, and researchers should be able to explore a wide range of analytical techniques – at least in a Kuhnian pre-paradigmatic phase. Though the study of cross-modal musical shapes is arguably a tiny niche—even within music psychology—exploring various analytical tools and putting them into the right context can make other music researchers reflect on their methods and analytical tools, stimulating a wider discussion about the current state of affairs.

In the following chapters (Chapters 3–6), I will show various ways of approaching empirically the study of cross-modal musical shapes—both its products and processes—with the tools and methods outlined in this chapter, before critically evaluating these studies and their approaches in Chapter 7.

Chapter 3: Cross-modal mappings of sound and music in a real-time drawing paradigm

3.1 Introduction

In this chapter, I investigate musically trained and untrained participants' representational strategies, and their performance accuracy, using visualizations of basic sound characteristics (pitch, loudness) and of music, in a real-time drawing paradigm. Literature on cross-modal correspondences of pitch and loudness, on visualizations of music, and on sensorimotor skills in musically trained participants will be reviewed first of all, before bringing these strands of research together to introduce the novelty of this approach.

3.1.1 Cross-modal correspondences of pitch and loudness

Psychophysical research applying experimental matching tasks, speeded identification and speeded classification has shown that dimensions of auditory stimuli such as pitch and loudness often correspond to dimensions of other modalities (for reviews see Eitan, 2013a; Marks, 2004; Spence, 2011). Studies investigating audio-visual and audio-spatial correspondences revealed that humans tend to match higher-pitched sounds with greater brightness (Marks, 1974, 1982; Wicker, 1968), with higher elevation in space (Mudd, 1963; Pratt, 1930; Roffler & Butler, 1968a; S. Wagner, Winner, Cicchetti, & Gardner, 1981; P. Walker et al., 2010; R. Walker, 1987), with smaller objects (Marks, Hammeal, & Bornstein, 1987; Mondloch & Maurer, 2004) and with spikier shapes (P. Walker et al., 2010). Similarly, louder sounds are matched with greater brightness[40] (Bond & Stevens, 1969; Marks, 1974; Stevens & Marks, 1965), with higher contrast (Wicker, 1968) and with larger objects (L. B. Smith & Sera, 1992; R. Walker, 1987).

The extent to which these correspondences are innate or learned is currently debated. Findings indicating that 3- to 4-month-old infants associate higher-pitched sounds with higher spatial positions and with spikier shapes (P. Walker et al., 2010), and that 20–30 day-old infants show some kind of loudness–brightness matching (Lewkowicz & Turkewitz, 1980), suggest that cross-modal matching is innate. However, given the speed of learning of arbitrary cross-modal correspondences (Ernst, 2007), it remains unclear to what extent cross-modal correspondences are primarily the result of statistical learning (Spence, 2011, p. 986). Another experiential factor,

[40] But see Marks (2004, p. 90) for a distinction between brightness and lightness.

the mediation of language, which plays a role for post-perceptual, semantic mappings of pitch (Eitan & Timmers, 2010; Nygaard, Herold, & Namy, 2009) and which may account for pitch–height mappings in congenitally blind children[41] (Welch, 1991), fails to explain, however, evidence from an experiment in which chimpanzees showed mappings between high brightness and high pitch (Ludwig, Adachi, & Matsuzawa, 2011). Although Ludwig and colleagues argue that it is unlikely that chimpanzees acquire this association through statistical learning (but see Spence & Deroy, 2012), the development of humans' cross-modal correspondences is at least shaped by repeated exposure, cultural factors and training (cf. Shayan, Ozturk, Bowerman, & Majid, 2014).

There is good reason to expect effects of musical training on cross-modal correspondences: for instance, Western musical notation is implicated in well established audio-visual mappings; the wider musicological discourse draws heavily on the notion of cross-modal metaphor (Zbikowski, 2002); and perhaps most significantly, musical performance itself engages (the interaction of) auditory, visual, motor and tactile senses. A common way to account for the influence of musical training on cross-modal correspondences is to compare groups of musically trained and untrained participants. Testing over 800 participants differing in age, cultural, environmental and musical backgrounds, R. Walker (1987) showed that, among these four factors, musical training has the largest impact on participants' consistency in matching pitch with vertical space (higher-pitched pure tones referring to higher elevation in space) and loudness with size (louder pure tones referring to larger shapes). Eitan and Granot (2006) similarly concluded that the difference between musically trained and untrained participants in an experimental task whereby participants had to imagine and specify the spatio-temporal motion of a humanoid character in response to changing musical parameters (e.g., pitch contour, pitch interval, loudness) is more a matter of consistency and security than of qualitative differences in cross-modal mappings. Musically trained individuals showed significantly stronger associations than musically untrained individuals between pitch and vertical space (higher pitch referring to higher in space) and pitch and lateral space (higher pitch referring to rightwards motion), whereas no significant differences were found regarding the mapping of loudness. Two further findings of Eitan and Granot's study are pertinent here too. First, the mapping of musical parameters onto

[41] But note that 4- to 5-year-old congenitally blind children in Roffler and Butler's study were unaware of describing pitches verbally with "high" and "low" (Roffler & Butler, 1968b).

spatial or temporal dimensions often occurred asymmetrically. For instance, a falling pitch contour was significantly more strongly associated with downward motion than was a rising pitch contour with upward motion. Secondly, pitch and loudness led to a multidimensionality of cross-modal mappings: changes in loudness were associated with changes in vertical space and direction; in speed; and in the energy level of the imagined character. Changes in pitch were associated with changes in all three spatial axes; in speed; and in the energy level. Similarly, investigating the relationship between musical training and multidimensional cross-modal mappings, Lidji, Kolinsky, Lochy and Morais (2007) found an automatic association of pitch with vertical space irrespective of musical background, and of pitch with horizontal space in musically trained participants only (see also Rusconi et al., 2006; Stewart, Walsh, & Frith, 2004). This corroborates Eitan and Timmers' (2010) finding that musically trained participants (compared to musically untrained) gave a higher-pitched section of a piano sonata compared to a very similar lower-pitched section of the same sonata a significantly higher rating of the verbal metaphor "right" (direction).[42]

Particularly relevant to the present study are two cross-modal drawing experiments. Godøy and colleagues (2010a; 2006a) and Haga (2008) asked 9 participants with varying degrees of musical training to represent short sound fragments (2–6 seconds long) on an electronic graphics tablet. The sound stimuli—produced with traditional and electronic instruments, as well as taken from the environment—were categorised according to a typology proposed by Schaeffer (1966), and comprised impulsive, continuous and iterative sounds. Features of pitch and timbre were classified into stable, unstable/changing and undefined, respectively. Although this study was more concerned with the hand gestures and participants were unable to see the trace they were creating, the analysis was based on the sound drawings. It was revealed that, regardless of their level of expertise, individuals were fairly consistent across participants, e.g., in representing pitch with height and the decay of a percussive sound with a descending line, but differed in respect to sound segments with multiple features such as a constant pitch and changing timbre, which some participants represented with a horizontal line, while others drew curved shapes.

[42] Of a list containing 35 pairs of antonyms (e.g., left – right, loud – soft), only one further significant difference between musically trained and untrained participants was found. Musically untrained participants gave the lower-pitched section of the sonata a significantly higher rating of the verbal metaphor "loud". However, this result might be confounded since the recording of the piano sonata was not normalized for loudness.

The second study worth noting here focused on a cross-cultural comparison of visual representations of sound between the UK, Japan and Papua New Guinea. Using simple sound stimuli varied in pitch contour and asking participants to create marks on a sheet of paper so that other community members could associate them with the sound heard, Athanasopoulos and Moran (2013) found that UK participants and Japanese participants familiar with Western notation used the y-axis for pitch and the x-axis for time, proceeding from left to right. Participants from a traditional Japanese music background depicted time vertically, starting at the top and moving down. While both UK and Japanese participants used symbolic representations, Papua New Guineans showed iconic representations, depicting aspects not deliberately manipulated by the authors such as timbre (e.g., flute sound) or loudness.

To sum up, pitch and loudness are mapped onto a great variety of visuo-spatial dimensions, with musical training influencing mostly the degree, as opposed to the direction or type of mapping. In the light of participants' eclectic visuo-spatial responses to simple auditory stimuli such as pure tones or synthesized instrumental sounds, a question arises: how do musically trained and untrained participants visualize more complex auditory stimuli – in other words, music?

3.1.2 Visual shapes of music

While some audio-visual correspondences operate on automatic, pre-conscious levels, asking participants to depict or match pieces of music visually engages them in a conscious, reflective task. Studies examining adult musically trained and untrained participants' visual responses to pieces of music are sparse, which is why some of them will be discussed here in more detail. Investigating the impact of the type of music, motivation and listening environment on the amount of graphical variation in drawing responses, Hooper and Powell (1970) showed that participants' sketches were more elaborate for (the now somewhat out-dated term) 'absolute music' in comparison with 'program music', when participants rhythmically engaged in the listening, and when the music was presented live as opposed to on record. Tan and Kelly (2004), investigating visualizations of complete musical works, asked their participants to "make any marks" (p. 195) while listening and to explain their drawings verbally. Results revealed that musically trained individuals focused on intra-musical properties (melodic themes, repetition, pitch contour, timbre etc.), making use of abstract representations. Musically untrained individuals, on the other hand, focused on extra-musical properties such as the arousal of (their

own) emotions and sensations, resulting in pictorial representations and short narratives including the listener as an agent or narrator. Similarly, Gromko (1995) compared musically untrained participants' visual representations and verbal descriptions of excerpts of classical music and found that fewer than 50% depicted or described musical elements such as melodic line, rhythm or texture. Those who did represent musical characteristics chose predominantly the rhythm, whereas only 5% depicted several musical elements (see also Dunn, 1997). Applying a matching task, Cowles (1935) asked musically trained and untrained participants to pair musical pieces with paintings depicting landscapes and provide written explanations of their choices. Regardless of musical training, participants matched musical pieces containing a high degree of dynamic changes with paintings rich in dynamic content and 'motor activity'. On the other hand, musical pieces without much dynamic variety were paired with paintings depicting little dynamic content. Whereas the choices for high-dynamic music–painting pairings were reflected in the introspective evidence (i.e. verbal reports were based on intra-musical properties such as rhythm and tempo), the choices for low-dynamic music-painting pairings were often based on extra-musical properties such as mood.

The lack of a more substantial body of empirical evidence notwithstanding, it seems to emerge that musically trained participants focus more on dynamics and shapes of intra-musical properties in their visualizations, whereas musically untrained participants draw more on music-induced associative ideas and emotions. Unquestionably, musical training involves the development of analytic listening but it remains unclear whether, and if so to what extent, musically untrained individuals are simply uninterested in, or genuinely unable to pay attention to the numerous musical parameters interwoven in musical pieces.[43] Another skill acquired through musical training, which undoubtedly lies beyond mere focused attention because it requires years of deliberate practice, is fine-tuned sensorimotor control.

3.1.3 Transferable sensorimotor skills in musically trained and untrained participants

Behavioural and neurocognitive research concerned with musically trained individuals' sensorimotor abilities (for a review see Zatorre et al., 2007) mainly focuses on effects of instrument specificity or other musically relevant tasks such as synchronization. For instance, studies where the participants' task involved 'tapping to the beat' revealed that musically trained

[43] For an extensive discussion of various ways of listening, and more specifically, the distinction between 'musical listening' and 'musicological listening' see Cook (1990, p. 152). For an ecological approach to listening see Clarke (2005).

participants outperform musically untrained participants by showing smaller asynchronies, lower tapping variability and better phase correction after tempo shifts (Repp, 2010; Repp & Doggett, 2007) and that musically trained and untrained participants recruit differing neural networks for synchronization tasks (Chen, Penhune, & Zatorre, 2008). However, the extent to which sensorimotor skills are transferable to non-musical tasks lacks thorough investigation. Both musically trained adults and children have been found to perform better than age-matched, musically untrained groups in simple as well as more complex reaction-time tasks involving visuo-spatial-motor integration (Brochard, Dufour, & Després, 2004; Costa-Giomi, 2005). Children who received two years of musical training showed no enhanced co-ordination of eye-hand movements in a non-musical task compared to children who did not receive any musical training. Thus, the development of sensorimotor skills supposedly requires substantial amounts of practice and might even be dependent on sensitive periods in motor learning (Watanabe, Savion-Lemieux, & Penhune, 2007). Evidence that musically trained individuals acquire enhanced, transferable, fine-tuned visuo-motor skills has been provided by Spilka, Steele and Penhune (2010), who compared musically trained and untrained participants regarding their ability to imitate novel, unfamiliar gestures taken from the American Sign Language. Their results, indicating that musically trained participants' gestures were more accurate on a global (arm, hand and finger movements combined) as well as on a local level (finger only), can be accounted for by theories of experience-based imitation mechanisms such as the associative sequence learning theory (Brass & Heyes, 2005) or the personal action repertoire hypothesis (Spilka et al., 2010). It is therefore likely that musical training equips individuals with many more transferable, fine-tuned sensorimotor skills than previously thought. The present study set out to test this assumption applying a real-time visualization paradigm. To that end, audio-visual correspondences, which are intended to reveal the influence of musical training on mappings of sound and music, serve, at the same time, as a measurement of performance in a fine-tuned sensorimotor task.

3.1.4 Novelty and objectives of the present study

Basic characteristics of sound such as pitch and loudness are mapped, sometimes asymmetrically, onto several visuo-spatial dimensions with varying degrees of consistency depending on the amount of musical training. Whereas most studies in this domain included paradigms such as speeded classification/identification, (unspeeded) matching and stimulus-

response compatibility, to the best of my knowledge no one has systematically investigated how pure tones varying in pitch and loudness, nor musical excerpts, are mapped visually in a real-time drawing paradigm. Here, I use an electronic graphics tablet (Wacom™, 2011) to collect digital data about position on the tablet and pressure applied to the pen, which can then be used, together with the audio data, for various analytical approaches.[44] The advantage of such a paradigm is the possibility of gaining insight into both qualitative and quantitative aspects of sound visualizations, with the latter intended to shed light on the performance accuracy of real-time cross-modal mappings. Hence, the objectives of the present study are two-fold:

1. Comparisons between musically trained and untrained participants' visual representations of sound and music in a free, real-time drawing paradigm by means of an electronic graphics tablet aim to reveal any commonalities or differences that may occur as a result of musical training.
2. The exact measurement of 'drawing performances' by means of an electronic graphics tablet is intended to reveal the extent to which musically trained and untrained participants' visualizations correspond systematically to sound parameters such as pitch and loudness.

3.2 Methods

3.2.1 Participants

Seventy-three participants (42 female) took part in the study (*M* = 28.51 years, SD = 7.71 years, range: 18–54 years). Participants were recruited using college-wide emails sent to undergraduates, postgraduates and staff (both academic and non-academic). From over 300 respondents, who were asked to provide information about their musical training, the most musically trained and untrained were selected, balancing age, sex and instrumental category. Thirty musically untrained participants (17 female, *M* = 28.77 years, SD = 7.14 years) were included[45], none of whom exceeded Grade 1 of the ABRSM examination system (http://www.abrsm.org/); twenty-two had never played an instrument and those who did had stopped playing at least 7 years ago and had not spent time actively making music for more than 6 years (*M* = 3.38 years, SD = 1.60 years). Forty-one musically trained participants (24

[44] Analytical approaches which have been used for similar sets of data include correlation analyses such as Spearman's rank correlation (Vines et al., 2006) or Canonical Correlation Analysis (Nymoen et al., 2011), Functional Data Analysis (Levitin et al., 2007) and time series analysis (Schubert, 2004).
[45] Two participants had to be excluded because they either provided conflicting information regarding their current musical activity or revealed only during the experiment that they had exceeded Grade 1.

female, *M* = 28.63 years, SD = 8.23 years) comprised twelve keyboard players, eight string players, eight wind/brass players, eight composers and five singers. While ten musically trained participants were above grade 8, fifteen at grade 8, and only one at grade 6, not all musically trained participants took formal exams, since there is no such grading system for composers.

3.2.2 Materials

3.2.2.1 Sound

Sound stimuli consisted of eighteen sequences of pure tones (length: 4.5–14.3 seconds; one, five or sixteen different tones per sequence), created with PureData (v0.42.5-extended) and systematically varied in pitch, loudness and tempo, as well as two short musical excerpts (see Table 3-1).[46] An overview of the pitch structure of the pure tones can be seen in Figure 3-1. Pitch range from B2 (123.47 Hz) to D4 (293.67 Hz) was chosen according to the melodic line of the first bar of Chopin's Prelude Op. 28, No. 6, which provided the musical excerpts. The simplest sound stimuli (Nos 1–3) consisted of single pure tones (D4) lasting for five seconds, either decreasing and increasing in loudness; without change in loudness; or increasing and decreasing in loudness. The same loudness patterns were applied to all pure tone sequences. Sound stimuli Nos 4–6 showed a falling-rising pitch contour in semitone steps whereas Nos 7–18 displayed a rising-falling contour. Nos 7–9 and Nos 13–18 used semitone steps, and Nos 10–12 used the notes B2, D3, F#3, B3, D4 in line with the Chopin Prelude. Nos 13–15 showed two decelerations, Nos 16–18 two accelerations, which always ended/started (respectively) at the peak pitch. Two recordings of the first two bars of Chopin's Op. 28, No. 6, one by Martha Argerich from 1975 (No. 19, length: 7.3 s) and one by Alfred Cortot from 1926 (No. 20, length: 8.1 s), were included (see Appendix 3.1).

[46] The sound stimuli can be downloaded at http://tinyurl.com/nqmn3ej.

Table 3-1 Overview of experimental sound stimuli

No.	Length (ms)	Pitch (Note name)	Amplitude	Tempo
1	5000	constant (D4)	decreasing - increasing	N/A (single pure tone)
2	5000	constant (D4)	constant	N/A (single pure tone)
3	5000	constant (D4)	increasing - decreasing	N/A (single pure tone)
4	4900	down - up (D4-B2-D4)	decreasing - increasing	equal, longer notes at bottom and top
5	4900	down - up (D4-B2-D4)	constant	equal, longer notes at bottom and top
6	4900	down - up (D4-B2-D4)	increasing - decreasing	equal, longer notes at bottom and top
7	4900	up - down (B2-D4-B2)	decreasing - increasing	equal, longer notes at top and bottom
8	4900	up - down (B2-D4-B2)	constant	equal, longer notes at top and bottom
9	4900	up - down (B2-D4-B2)	increasing - decreasing	equal, longer notes at top and bottom
10	4500	up - down (B2-D4-B2)	decreasing - increasing	equal, longer notes at top and bottom
11	4500	up - down (B2-D4-B2)	constant	equal, longer notes at top and bottom
12	4500	up - down (B2-D4-B2)	increasing - decreasing	equal, longer notes at top and bottom
13	13600	up - down (B2-D4-B2)	decreasing - increasing	decelerando - decelerando
14	13600	up - down (B2-D4-B2)	constant	decelerando - decelerando
15	13600	up - down (B2-D4-B2)	increasing - decreasing	decelerando - decelerando
16	14300	up - down (B2-D4-B2)	decreasing - increasing	accelerando - accelerando
17	14300	up - down (B2-D4-B2)	constant	accelerando - accelerando
18	14300	up - down (B2-D4-B2)	increasing - decreasing	accelerando - accelerando
19	7300	1st two bars of Chopin's Prelude Op. 28, No. 6 performed by Martha Argerich		
20	8080	1st two bars of Chopin's Prelude Op. 28, No. 6 performed by Alfred Cortot		

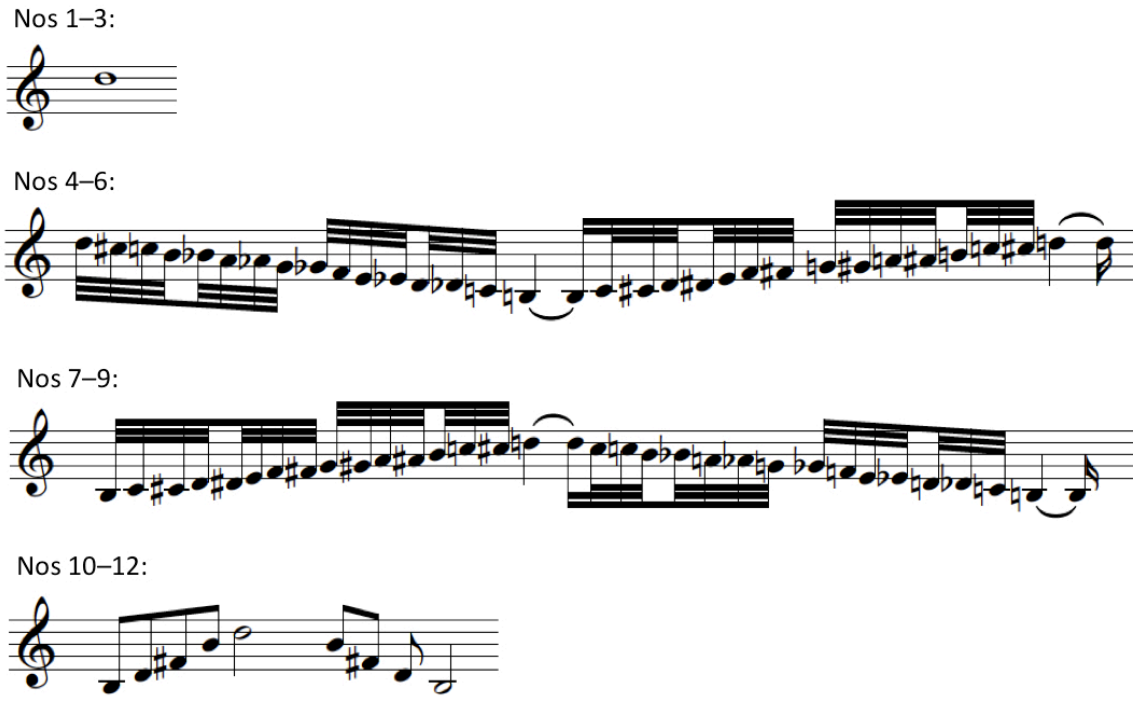


Figure 3-1 Overview of pitch structure of pure tones used in experimental trials. Note that sound stimuli Nos 13–18 (not displayed) have the same pitch structure as Nos 7–9 but different timings (i.e. decelerando and accelerando patterns).

3.2.2.2 Graphics Tablet

A commercial electronic graphics tablet (Wacom™ Intuos4™ L) was used to capture the directional responses in two spatial dimensions and the pressure applied with the pen. Its high sampling rate (maximum 200 points per second (Wacom™, 2011)) and high resolution (5080 lines per inch and 2048 levels of pressure (Wacom™, 2011)) ensured that the software did not miss a drawn point by virtue of the tablet sampling too slowly, and that subtle positional changes were captured. The software was developed in Processing (Fry & Reas, 2011) by Nicolas Gold. For a more detailed account of the capture software see Chapter 2.

3.2.3 Procedure

After signing the consent form, participants were instructed and familiarized with the tablet and pen. At all times, they were able to see their drawings on a screen in front of them. Applying more pressure to the pen resulted in a thicker line. Sound stimuli were presented with a commercially available set of headphones (Sony MDR-7506 Professional Dynamic Stereo Headphones).

During practice trials prior to the experiment, participants were presented with five stimuli very similar to those used in the proper experiment to become familiar with the procedure. Each sound stimulus was played twice. During the first presentation participants were asked to listen carefully without drawing, and during the second presentation they were asked to represent the sound visually by drawing along as it was played.[47] More specifically, they were told to

> 'draw everything in one go, that is, without going back to parts which you have already drawn. You can think of it as a musical performance, whereby it is impossible to go back in time and play the last bar or note again. Also, please try to take into account all sound characteristics you are able to identify and ideally, represent them in your drawing.'

Participants were asked not to use any formal symbols, numbers or letters.

The presentation order was pseudo-randomized by grouping the stimuli into four categories. Stimuli of the first group only varied in loudness, stimuli of the second group varied in pitch and loudness, stimuli of the third group varied in pitch, loudness and tempo, and the fourth group comprised both musical excerpts. To increase the complexity and interaction of musical variables stepwise the order of the four categories was always the same, while the presentation order within each category was randomized. The same procedure was applied during practice trials.

A questionnaire was used to collect demographic data and information about music-listening habits, musical education, language skills, familiarity with graphic scores, and any visual or hearing impairments. Participants were asked whether they used any strategies in representing pitch, loudness and time, and if so, what they were and how consistently they thought they had applied them (on a 5-point scale), how difficult they found the task (on a 5-point scale), whether they liked or disliked anything in particular, and whether they had any additional comments. Upon completion, participants received a small financial compensation for their time. This procedure had been approved by the appropriate ethics committee (REP-H/09/10-15).

[47] Note that the experiment also included a second part, whose results are not reported here, in which participants were asked to draw after the sound was played.

3.2.4 Analysis

3.2.4.1 Selection of participants for correlation analysis (performance accuracy)

A selection of participants was carried out to ensure that the performance accuracy (correlations between drawing features and sound characteristics) corresponded to participants' representational strategies. Thus, based on participants' most common representation strategies (see Results), only those reporting using the vertical axis to represent pitch and the thickness to represent loudness were included. Musically trained participants had to be at least at grade 8 ABRSM for their first or second musical activity, and engage in their main musical activity for at least four hours per week. This resulted in subsamples of twenty-two musically trained and thirteen musically untrained participants.

3.2.4.2 Correlation analysis

Feature extraction of frequency and loudness was carried out in Praat (Boersma & Weenink, 2012). An automatic pitch extraction algorithm was run on both musical excerpts, and adjusted manually to ensure only the melody was included. Pitch was represented on a log-transformed frequency scale in Hertz, and perceived loudness values were measured in sone. All values were standardized (M = 0, SD = 1) per stimulus. Similarly, y values (representing height on the tablet) and pressure values (representing thickness of the line/pressure applied to the pen) were standardized per drawing. Analysis was carried out in MATLAB (R2010a, The MathWorks) and SPSS (Version 19.0.0, IBM SPSS Statistics). Spearman's rank correlation coefficients ρ were calculated due to the nature of the dataset being serially correlated and non-stationary (for a full description see Vines et al., 2006).

3.3 Results

3.3.1 Comparisons between musically trained and untrained participants' visual representations of sound and music

3.3.1.1 Verbal reports: strategies for representing pitch, loudness and time

Differences in frequency between musically trained and untrained participants' representational strategies were assessed with χ^2-tests, or Fisher's exact test where appropriate. Asking participants after the experiment whether they had used any strategies to represent pitch revealed that the majority of both groups used height on the tablet (higher on the tablet referring to higher pitches). Of the musically trained participants, 97.6% reported using this strategy, and

only one musician, a female pianist and singer, reported a different strategy insofar as she had used pressure to represent pitch (the lower the pitch the more pressure applied).

A comparably high percentage of musically untrained participants (80%) chose to represent pitch with height, 3.3% used a mixed strategy of height and pressure (the lower the pitch the heavier the weight on the pen), 6.7% used a completely different approach (e.g., representing feelings) and 10% reported that they did not use any strategy. There was a significant association between musical training ('trained' vs. 'untrained') and the representational strategy used ('height' vs. 'mixed/different/none'), $\chi^2(1) = 6.01$, $p = .037$ (Fisher's exact test). The odds of choosing height to represent pitch were 10 times higher if a participant belonged to the 'musically trained' category than if a participant belonged to the 'musically untrained' category. Three examples of alternative (mixed or different) strategies for pitch representation are shown in Figure 3-2.

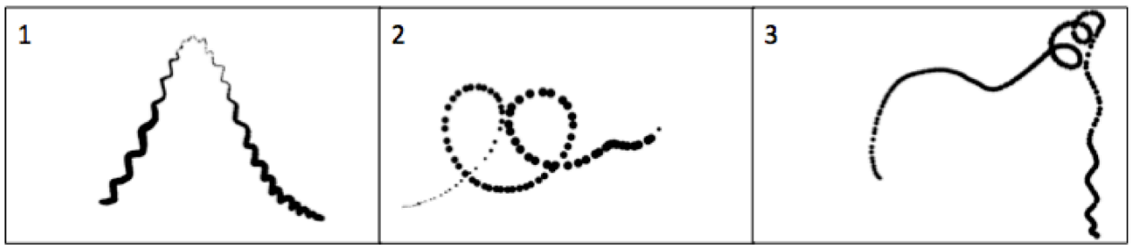


Figure 3-2 Examples of alternative pitch representations. 1. Mixed strategy of height and pressure (musically untrained participant depicting sound No. 16). 2. Different strategy: pressure (musically trained participant depicting sound No. 9). 3. Different strategy: feelings (musically untrained participant depicting sound No. 15).

Regarding musically trained participants' strategies for representing loudness, similar frequency distributions were observed. Of the musically trained participants, 87.8% reported that they used the thickness of the line (equivalent to the pressure applied) to depict loudness (the louder the sound the more pressure applied), while 4.9% applied mixed strategies consisting of thickness and drawing circles or 'wiggles' differing in size. Two musically trained participants chose different approaches, one of them using more pressure for softer sounds and the other using height (higher for louder sounds; when conflicting with pitch, representation of loudness would win). One musician, a female violinist, did not report any strategy.

Displaying more mixed results, 56.7% of the musically untrained participants reported that they used thickness to represent loudness (thicker line for louder sounds), and 16.7% used mixed strategies by applying thickness paired with circles (bigger = louder) and height on the tablet. 23.3% chose to represent loudness with a completely different strategy (height of waves; size of shapes; thickness of line achieved by fast movements up and down; high frequency oscillation), and 3.3% reported using no strategy. There was a significant association between musical training and the representational strategy used ("pressure/thickness" vs. "mixed/different/none"), $\chi^2(1) = 8.88$, $p = .003$. The odds of choosing pressure/thickness to represent loudness were 5.51 times higher if a participant belonged to the "musically trained" category than if a participant belonged to the "musically untrained" category. Six examples of alternative (mixed or different) strategies for loudness representation are shown in Figure 3-3.

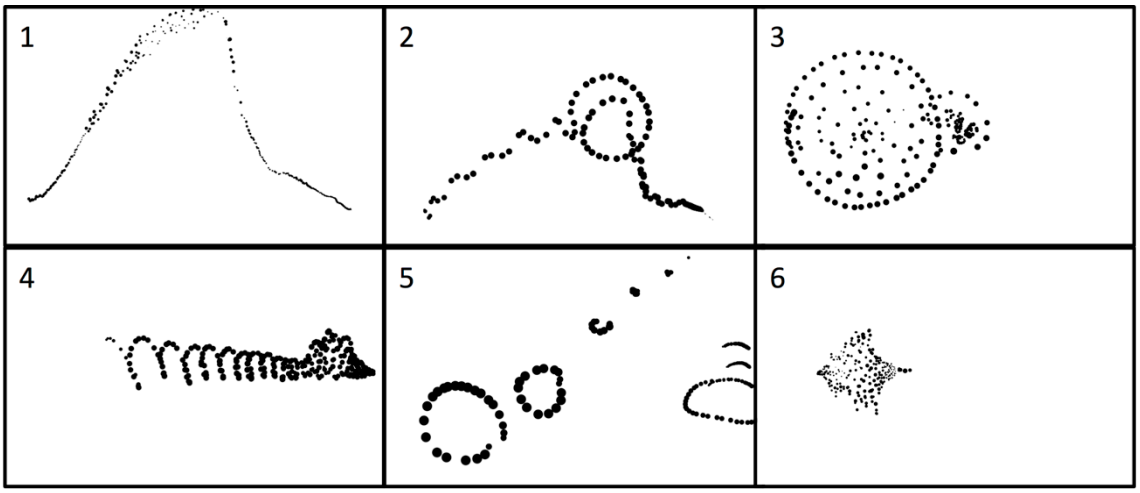


Figure 3-3 Examples of alternative loudness representations. 1. Mixed strategy of thickness and 'wiggles' (musically trained participant depicting sound No. 15). 2. Mixed strategy of thickness and circles (musically trained participant depicting sound No. 9). 3. Mixed strategy of thickness and circles (musically untrained participant depicting sound No. 14). 4. Different strategy: height of waves (musically untrained participant depicting sound No. 16). 5. Different strategy: size of shapes (musically untrained participant depicting sound No. 16). 6. Different strategy: fast movements up and down (musically untrained participant depicting sound No. 15).

Regarding the representation of time, the majority of both groups used the horizontal axis on the tablet (proceeding from left to right). Of the musically trained participants, 82.9% reported using this strategy; 9.8% applied mixed strategies consisting of horizontal axis (left to right) combined with circles, or pressure, or going into the opposite direction when reaching the end of the tablet; and only one musician, the same female pianist and singer who already reported a different strategy for pitch, reported having "started off and ended with a slight curve." Two

musically trained participants reported no strategy, although their drawings revealed that they, as well as the female musician reporting a different strategy, were in fact going from left to right.

Similarly, 83.3% of the musically untrained participants chose to represent time horizontally (left to right); 6.7% used mixed strategies consisting of horizontal axis (left to right) combined with 'swirls', or changing the direction (right to left), or the vertical axis (going down); one participant reported representing his feelings (see Figure 3-2); and 6.7% (two musically untrained participants) reported not using any strategy. Visual inspection of the drawings revealed that one of these musically untrained participants went in fact from left to right. There was no significant association between musical training and the representational strategy used ("horizontal axis" vs. "mixed/different/none"), $\chi^2(1) = .002$, $p > .90$.

3.3.1.2 Exploratory analysis of temporal representation

Visual inspection revealed remarkable differences between musically trained and untrained participants' representations of peak and trough pitches, which lasted longer than all other (ascending and descending) pitches in the stimuli. Some musically trained participants represented them with a longer line, creating a plateau-like shape. Some musically untrained participants seemed to 'pause' their drawing performances at the top (bottom) of the rising (falling) pitch contour, resulting in a somewhat pointed shape of their visual representations. To test the hypothesis that musically untrained participants tend to neglect temporal aspects of pitch, three independent, musically trained raters A (female, 30 years), B (male, 57 years) and C (male, 47 years)—all working in the field of music in Higher Education—were asked to evaluate on a 5-point scale (1 = very poorly, 5 = very well) how well sustained pitches of sound stimuli Nos 4–12 were accounted for in the drawings, resulting in one global rating per participant. Moreover, they were asked to evaluate on a 5-point scale (1 = very poorly, 5 = very well) how well changes in tempo of sound stimuli Nos 13–18 were accounted for in the drawings, resulting in one global rating per participant. All three raters were blind to the participant group. Inter-rater reliability was assessed using a two-way mixed, absolute, average-measures intra-class correlation (Hallgren, 2012). The intra-class correlation (ICC) for 'sustained pitch' was in the excellent range, ICC = .87, and for 'tempo changes' in the fair range, ICC = .56. Ratings were then averaged across all three raters and these averaged 'sustained pitch' and 'tempo change' values were used to compare musically trained and untrained participants' ratings with independent *t*-tests (degrees of freedom adjusted where necessary).

Musically trained participants' visual representations of 'sustained pitch' were rated significantly higher (M = 3.03, SEM = .16) than those of musically untrained (M = 1.69, SEM = .12), $t(68.28) = 6.80$, $p < .001$, $r = .64$ (see Figure 3-4). Similarly, musically trained participants' visual representations of 'tempo change' showed significantly higher ratings (M = 2.81, SEM = .14) than those of musically untrained (M = 2.22, SEM = .10), $t(66.75) = 3.47$, $p = .001$, $r = .39$.

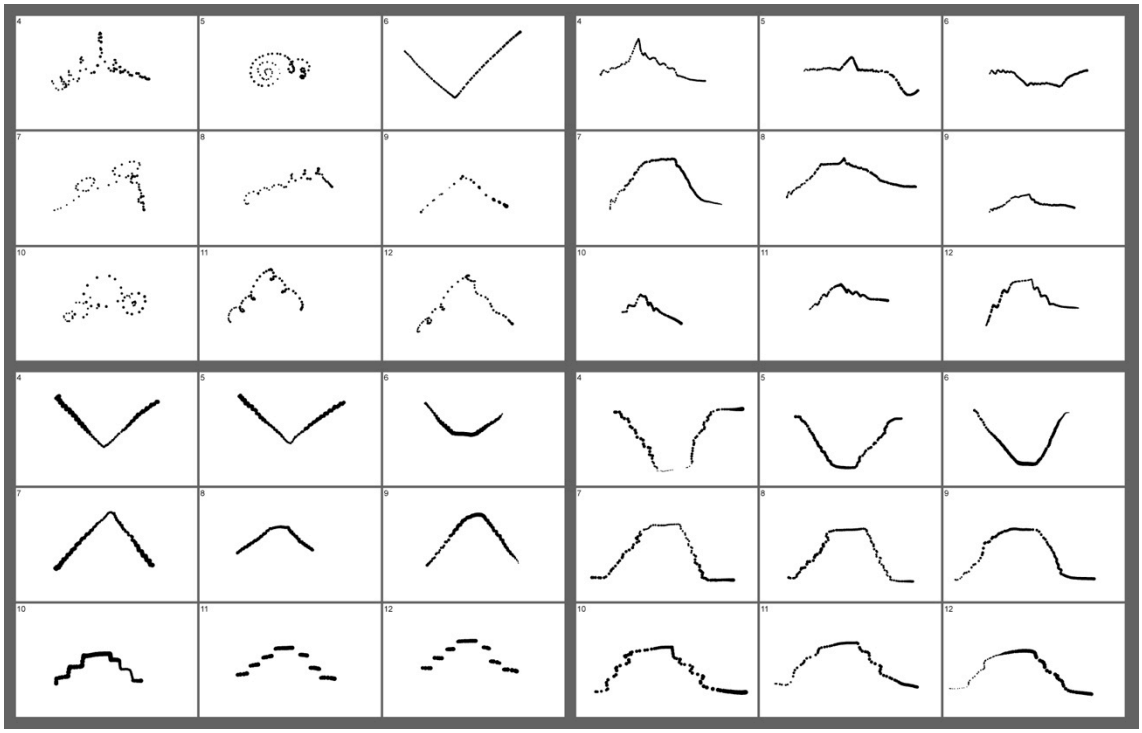


Figure 3-4 Visual representations of sound stimuli Nos 4–12 by four participants who showed low, medium and high ratings of 'sustained pitch' representations. Three independent raters were asked to evaluate on a 5-point scale (1 = very poorly, 5 = very well) how well the sustained pitches were accounted for in the drawings. Top left: musically untrained participant (average score 1.00). Top right: musically untrained participant (average score 2.33). Bottom left: musically trained participant (average score 2.50). Bottom right: musically trained participant (average score 4.67).

Use of the tablet space in absolute vs. relative terms was investigated by comparing whether sound stimuli Nos 16–18 took up more space horizontally (measured by calculating the x range in pixels $x_{max} - x_{min}$) than sound stimuli Nos 7–9. A mixed-design ANOVA was carried out with the between-subjects factor 'musical training' and the two within-subjects factors 'length' (short / long) and 'stimuli type' (amplitude decreasing / amplitude equal / amplitude increasing). All participants except for one musically untrained individual who did not use the x-axis for time were included in the analysis.

Results revealed a significant main effect for 'musical training' ($F(1, 68) = 12.62$, $p = .001$, partial $\eta^2 = .16$) and 'length' ($F(1, 68) = 63.27$, $p < .001$, partial $\eta^2 = .48$), while 'stimuli type' did not reach significance level, $F(2, 136) < 1$, $p > .60$. Regardless of the stimulus length, musically trained participants ($M = 818.33$, SEM = 27.83) used more space along the x-axis on average than musically untrained participants ($M = 664.71$, SEM = 33.09), and longer stimuli ($M = 814.13$, SEM = 21.88) generally led participants to use more space along the x-axis than when presented with shorter stimuli ($M = 668.92$, SEM = 24.95). There was a significant interaction between 'musical training' and 'length' ($F(1, 68) = 17.19$, $p < .001$, partial $\eta^2 = .20$), revealing that musically untrained participants compared to musically trained participants used significantly less space along the x-axis for short sound stimuli ($t(68) = 4.60$, $p < .001$, $r = .49$), while there was no significant difference between groups for long sound stimuli ($t(68) = 1.78$, $p = .079$, $r = .21$). No further significant interaction effects were observed.

3.3.1.3 Exploratory analysis of representational shift between pure tones and music

Visual inspection of drawings in response to the musical excerpts revealed that some participants completely changed their visualization strategy compared to the pure tones. To test the hypothesis that musically untrained individuals (but not musically trained individuals) tend to change their strategy when presented with a musical excerpt, raters A, B and C were asked to compare participants' visualizations in response to musical excerpts with all other visualizations of the same participant and indicate whether a shift in the representational strategy had occurred. Inter-rater reliability was assessed using kappa for each rater pair ($\kappa(AB) = .21$, $p = .087$, $\kappa(AC) = .34$, $p = .009$, and $\kappa(BC) = .13$, $p > .50$), and then averaged to provide a single index (Hallgren, 2012). The averaged kappa indicated fair agreement, $\kappa = .23$ (Landis & Koch, 1977), suggesting that some error variance was introduced by the independent raters, potentially reducing the statistical power of subsequent analyses.

Fisher's exact test ('musical training': trained/untrained; 'shift': yes/no), including the raters' mode values, revealed no significance ($p > .40$). Thus, although more musically untrained (13.3%) than musically trained participants (7.3%) have been rated to change their representational strategy, this difference did not reach significance level.

3.3.2 Comparisons between musically trained and untrained participants' performance accuracy assessed by non-parametric correlation coefficients

3.3.2.1 Consistency and difficulty

There were no significant differences between musically trained and untrained participants selected for the correlation analysis regarding the self-assessed consistency in applying their strategies ($t(33) < 1$, $p > .60$), nor regarding the perceived difficulty of the tasks ($t(33) < 1$, $p > .60$). Both groups reported that they acted fairly consistently in applying their respective strategies ($M_{mt} = 3.84$, $SEM_{mt} = .15$; $M_{mut} = 3.73$, $SEM_{mut} = .20$), and that they found the tasks relatively easy ($M_{mt} = 2.44$, $SEM_{mt} = .17$; $M_{mut} = 2.42$, $SEM_{mut} = .18$).

3.3.2.2 Local correlations of pitch–height

A local correlation denotes a correlation between the drawing characteristics (height or thickness) of a single drawing by a single participant and the perceived sound properties (pitch or loudness), e.g., the correlation between the *y* values (height) and frequency values (pitch) of participant *k*'s visual representation of sound stimulus *x*. To test statistically whether musically trained participants showed larger local correlation coefficients compared to musically untrained participants, four mixed-design ANOVAs with the dependent variable $\rho_{local\_pitch–height}$ and the between-subjects factor 'musical training' were run. The first ANOVA, investigating sound stimuli Nos 4–9, included the within-subjects factors 'pitch' (down-up / up-down) and 'loudness' (decreasing-increasing / increasing-decreasing / equal); the second ANOVA, investigating sound stimuli Nos 10–12, included the within-subjects factor 'loudness' (decreasing-increasing / increasing-decreasing / equal); the third ANOVA, investigating sound stimuli Nos 13–18, included the within-subjects factors 'timing' (decelerando-decelerando / accelerando-accelerando) and 'loudness' (decreasing-increasing / increasing-decreasing / equal); and the fourth ANOVA, investigating music stimuli Nos 19 and 20, included the within-subjects factor 'performer' (Argerich / Cortot). The same four ANOVAs were run including the dependent variable $\rho_{local\_loud-thick}$, as well as another ANOVA, investigating sound stimuli Nos 1 and 3 with the within-subjects factor 'loudness' (decreasing-increasing / increasing-decreasing / equal).

The musical training × pitch × loudness ANOVA (stimuli Nos 4–9) revealed a significant main effect of musical training, $F(1, 33) = 5.35$, $p = .027$, partial $\eta^2 = .14$. Since Levene's test of equality of error variances was significant for 50% of the variables (correlation coefficients of sound stimuli Nos 4, 8 and 9), an independent *t*-test was run using adjusted degrees of

freedom. Musically trained participants (M = .74, SEM = .04) showed higher correlation coefficients than musically untrained participants (M = .55, SEM = .08). This difference was not significant, $t(18.17) = 2.08$, p = .052; however, it revealed a medium-sized effect r = .44.

The musical training × loudness ANOVA (stimuli Nos 10–12) revealed significant main effects of musical training ($F(1, 33) = 5.86$, p = .021, partial η^2 = .15) and loudness ($F(1.71, 56.26) = 3.58$, p = .041, partial η^2 = .10). Musically trained participants (M = .66, SEM = .06) showed higher correlation coefficients than musically untrained participants (M = .41, SEM = .08), and decreasing-increasing loudness contour (M = .62, SEM = .05) led to higher correlation coefficients than increasing-decreasing loudness contour (M = .45, SEM = .06), $F(1, 33) = 7.31$, p = .011, partial η^2 = .18.

The musical training × performer ANOVA (stimuli Nos 19 and 20) revealed significant main effects of musical training ($F(1, 33) = 24.17$, $p < .001$, partial η^2 = .42) and performer ($F(1, 33) = 10.68$, p = .003, partial η^2 = .24). Musically trained participants (M = .67, SEM = .07) showed higher correlation coefficients than musically untrained participants (M = .13, SEM = .09), and the recording by Cortot (M = .52, SEM = .07) led to higher correlation coefficients than the recording by Argerich (M = .28, SEM = .07). The musical training × timing × loudness ANOVA (stimuli Nos 13–18) revealed no significant effects.

3.3.2.3 Local correlations of loudness–thickness

The musical training × loudness ANOVA (stimuli Nos 1 and 3) revealed significant main effects of musical training ($F(1, 32) = 10.68$, p = .003, partial η^2 = .25) and loudness ($F(1, 32) = 6.38$, p = .017, partial η^2 = .17). Musically trained participants (M = .65, SEM = .06) showed higher correlation coefficients than musically untrained participants (M = .32, SEM = .08), and increasing-decreasing loudness contour (M = .57, SEM = .06) led to higher correlation coefficients than decreasing-increasing loudness contour (M = .40, SEM = .06).

The musical training × pitch × loudness ANOVA (stimuli Nos 4–9) revealed significant main effects of musical training ($F(1, 33) = 6.05$, p = .019, partial η^2 = .16) and loudness ($F(2, 66) = 7.32$, p = .001, partial η^2 = .18). Musically trained participants (M = .33, SEM = .05) showed higher correlation coefficients than musically untrained participants (M = .15, SEM = .06). Pairwise comparisons using Sidak correction revealed that increasing-decreasing loudness contour (M = .39, SEM = .05) led to higher correlation coefficients compared to decreasing-

increasing loudness contour (*M* = .12, SEM = .06; *p* = .002) and equal loudness contour (*M* = .20, SEM = .05; *p* = .056), though the latter difference was not significant.

The musical training × loudness ANOVA (stimuli Nos 10–12) revealed a significant main effect of loudness, $F(1.67, 54.99) = 9.08$, $p = .001$, partial $\eta^2 = .22$. Pairwise comparisons using Sidak correction revealed that increasing-decreasing loudness contour (*M* = .35, SEM = .06) led to higher correlation coefficients compared to decreasing-increasing loudness contour (*M* = –.05, SEM = .07; *p* = .001) and equal loudness contour (*M* = .08, SEM = .07; *p* = .002).

The musical training × timing × loudness ANOVA (sound stimuli Nos 13–18) revealed a significant main effects of loudness, $F(2, 66) = 13.07$, $p < .001$, partial $\eta^2 = .29$. Pairwise comparisons using Sidak correction revealed that equal loudness contour (*M* = .17, SEM = .06) led to lower correlation coefficients compared to decreasing-increasing loudness contour (*M* = .42, SEM = .05; *p* = .013) and increasing-decreasing loudness contour (*M* = .54, SEM = .06; *p* < .001). The musical training × performer ANOVA (stimuli Nos 19 and 20) revealed no significant effects.

3.3.2.4 Global correlations of pitch–height and loudness–thickness

A global correlation denotes a correlation between the drawing characteristics (height or thickness) of all drawings of a single participant and the perceived sound properties (pitch or loudness), e.g., the correlation between the *y* values and frequency values of all visual representations of participant *k*.[48] To test statistically whether musically trained participants showed larger global correlation coefficients compared to musically untrained participants, two independent sample *t*-tests were run comparing respectively musically trained and untrained participants' pitch–height and loudness–thickness global correlation coefficients.

Musically trained participants' mean $\rho_{mt\_global\_ph} = .82$ (SEM = .03) was significantly larger compared to musically untrained participants' mean $\rho_{mut\_global\_ph} = .68$ (SEM = .04) for the pitch–height correlations ($t(33) = 2.80$, $p = .009$, $r = .44$). Regarding loudness–thickness correlations, musically trained participants' mean $\rho_{mt\_global\_lt} = .40$ (SEM = .04) was significantly larger compared to musically untrained participants' mean $\rho_{mut\_global\_lt} = .27$ (SEM = .05), $t(33) = 2.36$, $p = .024$, $r = .38$; however, the overall level was lower compared to global pitch–height

[48] Excluding visual representations in response to sound stimuli Nos 1, 2 and 3 for pitch–height correlations, and excluding visual representations in response to sound stimulus No. 2 for loudness–thickness correlations, was necessary because the respective sound stimuli could not be entered into a correlation analysis.

correlations (see Figure 3-5), $t(34) = 16.06$, $p < .001$, $r = .94$, mean $\rho_{global\_ph} = .77$ (SEM = .03), mean $\rho_{global\_lt} = 0.35$ (SEM = .03).

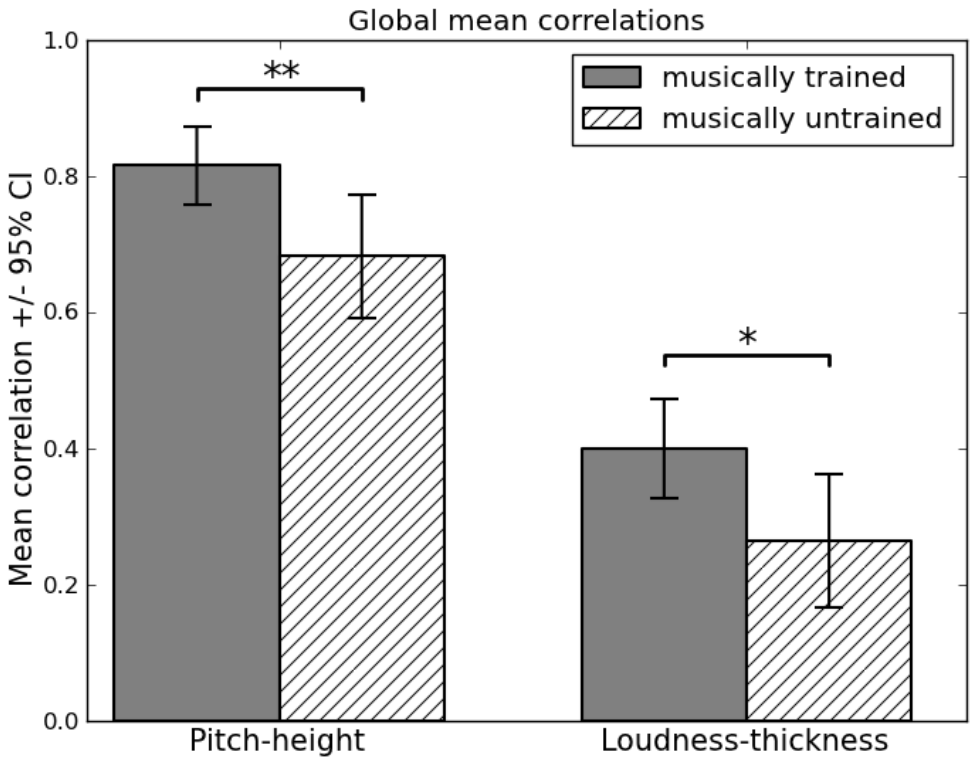


Figure 3-5 Mean global pitch–height and loudness–thickness correlations. Comparisons between musically trained and untrained participants revealed statistically significant differences. * indicates $p < .05$, ** indicates $p < .01$

3.3.2.5 Pitch–height representations of sound stimuli without pitch change

For sound stimuli Nos 1–3 the pitch–height correlations could not be calculated due to the *sine qua non* that variables in a correlation calculation must not be constant. Instead, the mean absolute deviation from the median of *y* was calculated for each individual drawing of sound stimuli Nos 1, 2 and 3. The lower the value of the mean absolute deviation the closer a visual representation was to a perfectly horizontal line. Three independent sample *t*-tests were performed comparing musically trained participants' averaged mean absolute deviation values with those of musically untrained.

For sound stimulus No. 1, it was found that musically trained participants ($M_{mt\_mad\_ph} = .23$, SEM = .05) showed significantly lower scores than musically untrained participants ($M_{mut\_mad\_ph} = .45$, SEM = .07): $t(33) = –2.67$, $p = .012$, $r = .42$. Although the same pattern was observed for sound

stimulus No. 2—musically trained participants' scores ($M_{mt\_mad\_ph}$ = .19, SEM = .05) were lower than musically untrained participants' scores ($M_{mut\_mad\_ph}$ = .33, SEM = .08)—the *t*-test failed to reach statistical significance, $t(33) = -1.73$, $p = .094$, $r = .29$. Sound stimulus No. 3 led again to the same pattern; musically trained participants' scores ($M_{mt\_mad\_ph}$ = .26, SEM = .05) were lower than musically untrained participants' scores ($M_{mut\_mad\_ph}$ = .57, SEM = .07), resulting in a significant difference, $t(33) = -3.80$, $p = .001$, $r = .55$.

3.3.2.6 Loudness–thickness representations of a sound stimulus without loudness change

For sound stimulus No. 2, the loudness–thickness correlation could not be calculated.[49] Thus, the mean absolute deviation from the median of p (pressure applied to pen) was calculated for each individual drawing of sound stimulus No. 2. The lower the values of this measurement, the lower the variation in the thickness of the line. An independent sample *t*-test was performed comparing musically trained participants' averaged mean absolute deviation values with those of musically untrained. There was no significant difference ($t(33) < 1$, $p > .40$) between musically trained ($M_{mt\_mad\_lt}$ = .41, SEM = .04) and musically untrained participants ($M_{mut\_mad\_lt}$ = .48, SEM = .07).

3.4 Discussion

When asked to represent visually pure tones varying in pitch, loudness and tempo, as well as two short musical excerpts, the majority of musically trained and untrained participants reported using the height to represent pitch (higher on the tablet referring to higher pitches), and the thickness of the line to represent loudness (thicker = louder). Various different strategies were observed, and the diversity was generally larger among musically untrained participants (particularly for the loudness representations). Analysis of visualizations of duration revealed that musically trained participants showed greater sensitivity for local timings such as sustained pitches and tempo changes, whereas musically untrained participants were more sensitive to the overall durations of the sound stimuli in relation to one another. Musically trained participants outperformed musically untrained participants regarding the accuracy of their representations (assessed across participants), which was independent of self-assessed

[49] It is worth noting that all loudness–thickness correlations in which the amplitude of the sound stimuli remained equal but the pitch was changed were included in the analysis since perceived loudness changes as a function of pitch (Y. Suzuki & Takeshima, 2004).

consistency and perceived task difficulty. Local and global, non-parametric correlations between sound and drawing characteristics showed that musically trained participants are overall more accurate than musically untrained participants in representing pitch with height and loudness with thickness.

3.4.1 Representational strategies of pitch and loudness

The finding that most participants mapped pitch onto a vertical axis (i.e., height on the tablet) is in accordance with previous research (Mudd, 1963; Pratt, 1930; Roffler & Butler, 1968a; R. Walker, 1987) and increases the external validity of such findings with this real-time drawing paradigm. Similarly, the correspondence shown between loudness and size has been reported before (Lipscomb & Kim, 2004; L. B. Smith & Sera, 1992; R. Walker, 1987) and corroborates previous findings. Free-drawing responses to pure tones revealed a greater tendency among musically trained participants to adhere to pitch–height and loudness–thickness representations in comparison with musically untrained participants. While musically trained participants' more consistent use of pitch–height mappings is in line with previous studies (Eitan & Granot, 2006; R. Walker, 1987), differences in mappings of loudness represent a novel finding which might partly be due to the free drawing paradigm. The greater diversity of representation strategies among musically untrained participants suggests that musical training plays an important role in choosing mapping strategies: it seems that with more musical training the number of potential representations considered decreases.

3.4.2 Representation of time

The correspondence between time and horizontal axis is in line with previous research (R. Walker, 1987). To account for the fact that the majority of participants went from left to right regardless of musical training, literacy seems to be the most likely candidate and has been shown to influence graphic representations of music (Athanasopoulos, Moran, & Frith, 2011).

However, the finding that musically untrained participants tend to neglect temporal aspects of pitch by 'pausing' their visualization performance at peak and trough pitches, as well as visually ignoring the changing durations of pitches in sound stimuli Nos 13–18, was unexpected and deserves closer attention. Note that whether or not participants took into account the varying pitch lengths has no impact on the precision of mere pitch representation. Thus, it is possible to achieve an "exact" representation of pitch while neglecting its temporal aspects. Nevertheless,

the fact that musically untrained participants tended to “pause” their drawings when pitch remained unchanged over time is remarkable not because they are unable to account for temporal aspects of pitch (in fact, they may well be able to do so given a specific instruction) but because they chose not to, or at least, because this aspect of the sound was evidently not salient enough to be worth considering. This effect, which in a different experimental paradigm could have been interpreted as a differential effect of musical training on working memory (George & Coch, 2011), is, due to the real-time nature of the experimental task, better regarded as a predominantly attentional bias. Moreover, since all participants listened to each sound stimulus once before drawing, and then drew along as the sound stimulus was played, the working memory load is very low and thus unlikely to account for the observed differences.

One has to be cautious, though, about drawing premature conclusions regarding musically untrained participants’ attentional bias of time. This becomes evident in the significant interaction between musical training and horizontal expansion of the drawings. Whereas musically trained participants showed no difference between short and long sequences, musically untrained participants used significantly less space horizontally for short sequences. Thus, although musically untrained participants did not consider local pitch timings, they did, unlike musically trained participants, show a sense for the overall length, and captured, if only in a rough manner, the relative global durations of the stimuli. Further research is overdue to shed light on this, as well as other phenomena linking musical training and the perception of time (Phillips & Cross, 2011). It would, for example, be worth testing the hypothesis that while musically untrained participants attribute less informational significance to unchanging pitch in a melodic line, when representing music in space they are less influenced than musically trained participants by the non-proportional nature of the representation of time in standard musical notation (cf. Tan, Wakefield, & Jeffries, 2009). For musically trained participants, then, available horizontal space can represent a whole musical extract, whatever its length, since this is their everyday experience of reading scores; for musically untrained participants space simply maps onto time. At the same time, but for quite different reasons, a continuing note has, for musically trained participants, just as much significance as one that changes.

3.4.3 Representational shift from pure tones to music

Although more musically untrained than trained participants displayed a representational shift from pure tones to music, this effect was non-significant. Note, however, that this result reflects

the outcome of a rating procedure whose overall agreement between raters was only fair. While the raters often disagreed whether a shift had occurred, it is plausible that participants applied such a shift, knowingly or not, more often than is apparent in the rating data. This would be in line with previous findings (Tan & Kelly, 2004) showing that musically untrained participants depict more extra-musical, associative ideas when asked to represent music visually. The possibility that a lifetime of seeing music through notation restricts musically trained participants' options for mapping music cross-modally merits more investigation.

3.4.4 Performance accuracy

Overall, musically trained participants showed higher correlation coefficients, locally and globally, for both pitch–height and loudness–thickness correlations. A closer look at local pitch–height correlations reveals that musically trained participants showed higher correlation coefficients for all short sequences (Nos 4–12) but not the longer ones (Nos 13–18). This suggests that musically untrained participants struggled with shorter sequences, which require immediate synchronization. On the other hand, longer sequences, which, unlike the short ones, were also varied in tempo, allow for more adaptive synchronization behaviour if one has missed the start of a sequence, for instance. Apart from one exception (sound stimuli Nos 10–12) loudness did not interfere with the size of the correlation coefficients. In this exception, the rising-falling pitch contour led to higher pitch–height correlations when paired with a falling-rising loudness contour. However, the same effect was absent in stimuli Nos 7–9 with the same pitch contour, rendering it otiose to ponder possible interpretations. What is more telling is that musically trained participants' correlation coefficients are higher when presented with the musical excerpts. It might be the case that musically untrained participants' response strategies changed when confronted with music, and that musically trained participants are much more familiar with such musical excerpts. Note also that both groups showed higher correlation coefficients for the Cortot (compared to the Argerich) recording. One would have to test the whole Chopin Prelude before drawing conclusions, but this seems to suggest that Cortot's timing was easier to follow with the pen, perhaps because it was slightly longer (780 ms, i.e. 11%) and therefore at a more "comfortable" speed. This finding is particularly interesting, however, because Cortot's extremely flexible timing is much less familiar to listeners today than Argerich's more regular playing. Leech-Wilkinson (2011, 2013) also argues that Cortot's playing

elicits unusually strong embodied associations, which might have facilitated participants' responses.

Loudness–thickness correlation coefficients were generally smaller than those of pitch–height. Arguably, controlling the thickness of the stroke by applying more or less pressure posed a greater challenge to participants because it required fine-tuned motor control. Musically trained participants only showed higher loudness–thickness correlation coefficients for short sound stimuli that were either unchanged in pitch (Nos 1 & 3) or progressed in semitones (Nos 4–9). When the pitch contour resembled the Chopin excerpt (Nos 10–12), or when sequences were longer (including both musical excerpts), there was no difference between groups. Although the trend was always the same (musically trained participants achieving higher values), the overall lower level combined with greater variances precluded significant differences. There was, however, a strong main effect of loudness independent of musical training. Both sequences unchanged in pitch (Nos 1 & 3), as well as all other short sequences (Nos 4–12) led to higher loudness–thickness correlation coefficients if the loudness contour was increasing-decreasing as opposed to decreasing-increasing, or equal. This makes sense since it is an easier motor task to build up pressure on the pen from zero to a maximum (and going back to zero), rather than reducing pressure from a maximum to (nearly) zero (and then building up again). Equally, in musical phrases in many genres, increases in pitch, speed and loudness followed by decreases towards the phrase-end are currently much more common than the reverse, so that participants' expectations of music were being met more readily in the increasing-decreasing sequences. Interestingly, keeping pressure roughly equal during longer sequences (Nos 13–18) was harder than synchronizing with a loudness contour that was either decreasing-increasing or increasing-decreasing. This, too, makes intuitive sense since it is generally harder to keep a force constant over a longer period than to apply gradual changes in whichever direction; and constant intensity is uncharacteristic of musical phrases in everyday musical contexts.

The statistical analysis of sound stimuli Nos 1 and 3 regarding pitch, and of sound stimulus No. 2 regarding loudness, revealed a pattern fitting into the overall picture: while musically trained participants showed higher scores for pitch representations than musically untrained participants, there was no significant difference between groups regarding loudness representation.

Taken together, musically trained participants outperform musically untrained participants in terms of accuracy in a real-time drawing task of pitch and loudness, which adds to the increasing literature of findings suggesting that musically trained participants' (sensori-)motor skills are transferable to other, non-specific domains (Spilka et al., 2010). However, the results are limited insofar as they cannot disentangle musically trained participants' motoric skills from training effects of audio-visual mappings or basic auditory discrimination skills (but note that I attempted to minimize potential differences in the latter by choosing very simple sound stimuli and presenting them twice). Future experiments should address these issues by varying the degree of visual feedback (e.g., include trials in which participants are blind-folded whilst drawing), or by asking participants to draw along an existing shape in synchrony with the sound. Since the results represent an implicit measure of motoric skills it would be valuable to examine whether differences remain when musically trained and untrained participants are instructed to represent pitch and loudness in the context of a motor skill experiment.

3.5 Conclusion

In this chapter, I set out to investigate how musically trained and untrained participants perform in a real-time drawing task that involves tracking pitch, loudness and time of sequences of pure tones, as well as two short musical excerpts, in order to shed light on cross-modal perception, whether conscious or subconscious, and on sensorimotor skills. I was able to demonstrate that auditory-visual correspondences previously reported in the literature also apply in a real-time drawing task, and that musically trained participants' acquired sensorimotor skills are evident in such drawing tasks, and potentially also other motor tasks unrelated to instrument-playing. What is more, I discovered that the visual representation of duration plays a crucial role in differentiating musically trained and untrained participants' ways of listening to and representing sound and music. Importantly, to substantiate the findings from this exploratory study, separate controlled experiments using confirmatory hypothesis testing should be carried out in the future. Further research also needs to be undertaken to show which aspects of musical training are responsible for these varying effects, and what consequences they have for the ways in which musically trained and untrained participants understand music through embodied cross-modal mappings. In the following chapter, I will explore various advanced mathematical techniques to study visualizations of sound and music.

Chapter 4: Exploring advanced mathematical tools to investigate visualizations of sound and music

4.1 Introduction

The previous chapter has revealed insights into how musically trained and untrained participants represent graphically a series of pure tones, as well as two short musical excerpts. Indeed, we have seen that the question of how sound and music are represented visually has been addressed by scholars from various disciplines to shed light on how we perceive motion in music (Godøy et al., 2006a; Repp, 1993a), how auditory information is processed and mapped onto the visual domain (Marks, 2004; Spence, 2011), how children develop an understanding of rhythm (Bamberger, 1982) and represent music graphically (Reybrouck et al., 2009; Verschaffel et al., 2010), and how musical training influences the ways in which we represent short but complete musical compositions visually (Tan & Kelly, 2004). Most of these research strands—which can be grounded in, or related to, theories of multi-modal perception (Stein & Meredith, 1993), embodied music cognition (Leman, 2007) and gestures (Godøy & Leman, 2010; Gritten & King, 2011)—deal with phenomena that happen over time due to the involvement of sound. It is therefore important that the holistic experience over time, formed through integrating auditory, visual and kinaesthetic senses, is accounted for in the analysis as well. Time-dependent analyses are not new in music psychology (Schubert & Dunsmuir, 1999), and have been used, for instance, to model emotional responses to music with non-parametric correlations (Schubert, 2002) or Functional Data Analysis (Vines, Nuzzo, & Levitin, 2005b), or to analyse gestural responses to sound with Canonical Correlation Analysis (Caramiaux, Bevilacqua, & Schnell, 2010; for a critical review of various analytical tools see also Nymoen et al., 2013). While correlational approaches, as applied in Chapter 3, capture only linear trends between features and are only indirectly sensitive to time-dependent data such that delayed—as well as premature—drawing in response to sound gives rise to smaller correlation values, Functional Data Analysis—at least Functional Analysis of Variance (fANOVA)—is less suited for studies including participants' responses to numerous sound/music stimuli, and is rather used to compare multiple responses to few musical excerpts and/or experimental conditions (e.g., Vines et al., 2006). Enabling a more direct investigation of the temporal unfolding of visualizations

concurrent to changes in sound, different advanced analytical approaches are necessary.[50] In this chapter, I intend to show how advanced mathematical tools can aid analysis in an attempt to create a more complete understanding of the way people think about and visualize shape(s) in response to sound and music.

First, I will demonstrate how advanced regression modelling techniques can be used to reduce the dataset, making it more manageable for further analyses while losing only minimal information. Secondly, I will perform clustering analyses on the reduced dataset to investigate the extent to which meaningful groupings emerge from the data. Thirdly, I will run classification analyses to examine the possibility of automatically classifying participants' drawings as belonging to either the 'musically trained' or 'musically untrained' categories and, if successful, what the implications might be for the ways in which musical training shapes cognitive processes. In all stages, the analyses will progress from simplistic to more complex in an attempt to find the best balance between efficiency and output. Finally, I will discuss the outcomes of these mathematical analyses in terms of their applicability to and the interpretability of the drawing dataset (see previous chapter).

4.2 Preparation of the dataset

For a detailed description of the experimental procedure see Chapter 3. As a reminder, the experiment produced a temporally correlated dataset with values for the x- and y-coordinates of the drawings (referred to as X and Y from here on), along with the pressure being applied to the tablet from the pen. Applying more pressure resulted in a thicker line. Data points were spaced approximately 45 milliseconds apart. For these analyses, all drawing data from before the sound stimulus started or after it finished were discarded. Prior to analysis, the frequency data were converted to a natural log scale to compensate for the manner in which humans perceive intervals of pitch.

The attributes of the sound stimuli (log-frequency, intensity, perceived loudness) were sampled every 10 milliseconds. These data were then linearly interpolated between points to produce a dataset with audio data every millisecond, allowing audio data to be matched to each time step of the drawing data. This resulted in between 2227 and 3568 data points per participant. All drawing and audio data were standardized (M = 0, SD = 1) across the entire dataset prior to

[50] However, it may be fruitful to adapt other classical methods such as clustering, classification or principal components analysis to functional data in future studies (Levitin et al., 2007).

being analysed. Time was scaled to between 0 and 1 per stimulus, with 0 being the start of the sound and 1 being the end. This gives each stimulus a slightly different time variable, which was deemed an acceptable compromise since participants listened to each stimulus completely before being asked to draw and thus most participants scaled their drawing in a manner appropriate to the length of the stimulus.

As with the global correlation analysis in Chapter 3, the aim here is to characterise general features of a participant's response to a range of auditory stimuli and find relations that hold across all stimuli presented. Instead of distinguishing between stimuli—which is a very worthwhile endeavour, particularly for real musical excerpts (see Chapter 6), but beyond the scope of this chapter—a single set of parameters will be fitted per participant. More detailed mathematical descriptions can be found in Noyce, Küssner and Sollich (2013). Only the basics necessary to understand the analyses in the context of this thesis will be covered here.

4.3 Regression

The starting point of a regression analysis is the recognition that a consistent visual representation of a sound stimulus requires a systematic dependence between the visual outputs (X, Y and pressure) and characteristics of the sound stimulus. This dependence can be learned from data for each participant, representing the outputs as a deterministic function of the stimulus plus noise. This means that the data from each individual are treated as a regression problem, i.e. as the task of learning a function from noisy data. Learning this function is governed by hyperparameters (defined below), e.g., the relative weight of deterministic and random contributions to an individual's output. These hyperparameters are learnt from the data and essentially indicate what type of function best represents a participant's behaviour, providing a convenient summary of how each individual visually represents sound stimuli. It is proposed that these learnt hyperparameters capture something significant about each participant's response and can thus form the basis for further analysis of trends in individual behaviour.

Initially, only frequency, intensity and perceived loudness were used as the inputs for this regression approach, but it quickly became clear that time needed to be added as a fourth input since most individuals used the x-axis to represent time even for sound stimuli with constant pitch. More generally, the inclusion of time can be viewed as allowing an individual's mapping

from stimulus to visual representation to depend on time, bearing in mind that the temporal evolution of the drawings is one of the foci of this chapter.

The four input variables (time, log-frequency, intensity and perceived loudness) were collected into an input vector $\mathbf{x}$. The three outputs (X, Y and pressure) were treated separately, with one regression function $f(\mathbf{x})$ fitted for each. This analysis was repeated for each participant, resulting in a set of hyperparameters per individual. The specific regression approach used was Gaussian processes (GPs) (Rasmussen & Williams, 2006). GPs are very flexible and can represent functions as simple as linear input-output dependences or as complex as general nonlinear functions requiring an, in principle, infinite number of parameters, which makes them ideal for the purposes of this analysis.

Noyce et al. (2013, p. 131) further note:

> "The complexity depends on the covariance $k(\mathbf{x}, \mathbf{x}')$, which represents the prior correlation of function values for different inputs $\mathbf{x}$ and $\mathbf{x}'$. The mean function $m(\mathbf{x})$ is taken as identically zero throughout, as is commonly done in GP regression. Note that this does not imply that one cannot represent systematic input-output dependencies; [Figure 4-1 to 4-3] below show clearly that such dependences are captured. The effect of assuming a zero mean function can be seen most simply in the context of the linear kernel to be described shortly, where the outputs are effectively modelled as linear combinations of the inputs (plus constants), placing a Gaussian prior distribution over the parameters in each linear combination. The assumption $m(\mathbf{x}) = 0$ then amounts to saying that we have no strong a priori knowledge about the sign of the parameters, and hence use Gaussian priors with zero mean."

4.3.1 Linear regression model

Noyce et al. (2013, p. 131) define the linear regression model as follows:

> "The initial model choice was very simplistic, with a linear kernel chosen as the covariance function:

$$k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T \mathbf{\Lambda^{-2}} \mathbf{x}' + \sigma_f^2.$$

A GP defined by this kernel fits a constant plus a linear function of the inputs to the data. $\mathbf{\Lambda}$ is a diagonal matrix with diagonal entries λ. These can be interpreted as inverse weights of different input components, meaning that a small λ corresponds to an input being an important predictor of the output. This GP has 18 hyperparameters per participant: 4 inverse weights (λ_{time}, $\lambda_{frequency}$, $\lambda_{intensity}$ and $\lambda_{loudness}$), an offset (σ_f) and a noise level (σ) for X, Y and pressure."

4.3.2 Nonlinear regression model: linear plus squared exponential kernel

Noyce et al. (2013, p. 131) go on to define the nonlinear regression model as follows:

"Next, a more complicated model was implemented to capture more of the variability in the data. The same analysis technique was used as for the first GP, with the addition of an automatic relevance detection (ARD) squared-exponential (SE) term to the covariance function. This contributes one length-scale parameter per input direction, with the overall covariance kernel of

$$k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T \mathbf{\Lambda}^{-2} \mathbf{x}' + \sigma_a^2 \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{x}')^T \boldsymbol{\mathcal{L}}^{-2}(\mathbf{x} - \mathbf{x}')\right) + \sigma_f^2.$$

The key property of the additional contribution to the kernel is that it allows the fit to contain not only constant and linear terms, but also arbitrary nonlinear functions that are, in principle, specified by an infinite number of parameters. $\boldsymbol{\mathcal{L}}$ is a diagonal matrix with diagonal entries ℓ. These can be interpreted as determining the distance over which the nonlinear contribution to $f(\mathbf{x})$ varies significantly for different input components, i.e. for distances below ℓ, this contribution is essentially constant. Thus, a large ℓ means that the input component is irrelevant to the variation of the nonlinear part of the regression function, in analogy to a large λ, which has the same meaning for the linear portion. The other additional hyperparameter, σ_a, regulates the amplitude of the nonlinear contribution to $f(\mathbf{x})$; small values mean an essentially linear fit. Overall, the above setup gives 33 hyperparameters per participant: 4 inverse weights (λ_{time}, $\lambda_{frequency}$, $\lambda_{intensity}$ and $\lambda_{loudness}$), an offset (σ_f), 4 lengths (ℓ_{time}, $\ell_{frequency}$, $\ell_{intensity}$ and $\ell_{loudness}$), an amplitude (σ_a) and a noise level (σ) for X, Y, and pressure."

4.3.3 Results and discussion

4.3.3.1 Linear regression models

4.3.3.1.1 Model performance

In general, this set of models performed equally well for trained and untrained participants. Although these models did reasonably well at modelling average outputs, they were not accurately able to predict the extremes (see Figure 4-1, Figure 4-2 and Figure 4-3). The noise hyperparameters (Figure 4-4a, f, k) indicate the goodness of fit for the GPs; a smaller noise term means that more of the variance in the data was explained by the deterministic part of the input-output relation, i.e. $f(x)$, than by random noise. X had the best fit, as is evident from the lower noise hyperparameters shown in Figure 4-4a, and pressure had the worst fit of the three outputs (Figure 4-4k). In particular, these models correctly identified upward trends in Y, but vastly overestimated the resulting peaks (Figure 4-2). Overall, while the model does capture some of the trends, there is clearly room for improvement in the predictions of all outputs.

4.3.3.1.2 Hyperparameter interpretation

As explained above, the size of the optimised inverse weight indicates the relevance of each input to the output being modelled, with small λ indicating high relevance. For X, time was the most important input by far (Figure 4-4b), as is clear from the small inverse weight hyperparameters when compared to the other inputs (Figure 4-4c-e), as expected. Most participants drew from left to right at a relatively constant speed throughout. Time was considerably more relevant for the trained participants (as indicated by their lower λ_{time}values), perhaps indicating that they were better able to keep track of the length of the stimulus. Frequency, intensity and perceived loudness had similar levels of relative unimportance for X with no clear differences between the trained and untrained participants (Figure 4-4c-e).

For Y, frequency, intensity and perceived loudness were all more important inputs than time (Figure 4-4g-j). Here, there is an observable difference between groups: frequency, intensity and perceived loudness were all more relevant for the Y output in the trained group than in the untrained group. For trained participants, frequency appears to be the most important input for Y, as expected (Figure 4-4h). This trend is also apparent in the untrained group, but to a lesser extent. This may indicate that changes in frequency are more likely to affect drawings made by trained individuals, which is consistent with findings of prior studies that trained adults are more

accurate in detecting changes in pitch than people with little or no musical training (e.g., Tervaniemi, Just, Koelsch, Widmann, & Schröger, 2005).

For pressure, perceived loudness was the most relevant input, especially for the trained group (Figure 4-4o). This fits with the 87.8% of trained participants who reported that they represented an increase in loudness of the musical stimulus by pressing down harder on their drawing. Intensity, frequency and time were all of similar importance, though intensity and frequency were generally more important for the trained participants than for the untrained (Figure 4-4m-n).

These results are similar to those of the analysis in Chapter 3, presumably because only a linear covariance kernel was used. Consequently, fitting a GP with only a linear covariance function confirms previous findings from a linear regression, which is an important first step, but does not allow for further insights.

4.3.3.2 Linear plus SE regression models

4.3.3.2.1 Model performance

The addition of a squared-exponential term in the covariance kernel produced a vast improvement in the performance of the GPs. For both trained and untrained participants the new GPs were relatively more accurate at predicting X (Figure 4-1) and pressure (Figure 4-3), with an average increase in log marginal likelihood (per datapoint) of 0.98 and 0.85, respectively. The increased ability of the model to capture the trends in the Y output is especially noticeable, with an average increase in log marginal likelihood of 1.22. The SE version of the model no longer overestimates peaks in the dataset and more accurately captures valleys (Figure 4-2).

The shift in values for the noise hyperparameters confirms these observations. For the linear model the averages of the noise hyperparameters were around 0.7 for Y and pressure and 0.4 for X (Figure 4-4). For the SE GPs, the noise levels dropped considerably, especially for Y. In these models, the highest noise hyperparameter for Y is 0.3 and the average noise value is 0.2 (Figure 4-5). Interestingly, the noise hyperparameters for Y are lower for the trained group, perhaps because they draw in a more predictable fashion. Even though pressure is still the worst-modelled of the three outputs, the overall noise levels are much lower with the addition of the SE kernel (Figure 5). Because of these results, it can be concluded that the improved model

output is worth the extra computing time required. However, while the nonlinear model captures more of the variability inherent in the data, it also produces 33 hyperparameters. These hyperparameters are optimised using 3000 datapoints, so they are still quite well-determined by the data, meaning that adding more hyperparameters is a feasible option. On the other hand, one of the purposes of these analyses is to represent each participant in the hyperparameter-space and, given that there are only 71 participants in the study, having nearly half that many hyperparameters may indicate over-specification. Consequently, results from the simpler linear model will be used for some future analyses as well.

4.3.3.2.2 Hyperparameter interpretation

As for the linear model, the size of the inverse weight and length hyperparameters indicate the relevance of each input to the output being modelled. The interpretation of these hyperparameters is however not as straightforward, perhaps because the addition of the more powerful squared exponential kernel causes many of the linear hyperparameters to be less important in the overall model.

Again, time is the most relevant input for X for both groups (Figure 4-5), which fits with expectations. The trained group had a stronger relationship between time and X than the untrained group, especially in the linear portion of the kernel (Figure 4-5b). Another difference between the two groups occurs in $\ell_{intensity}$, which implies that changes in intensity are more relevant to the prediction of X for the untrained individuals than for the trained (Figure 4-5i). In addition, in the GPs for X the trained group tended towards smaller values of σ_a, the amplitude parameter from the SE kernel (Figure 4-5f). This indicates that the drawings of the trained participants were better captured by the linear portion of the model than those of the untrained.

For Y, the linear hyperparameters have similar levels of relevance in both groups, which is unexpected (Figure 4-5l-o). There is some difference in the SE hyperparameters, however, though again the results are different from the purely linear model. Surprisingly, time comes out as being slightly more relevant to the prediction of Y than frequency, intensity, or loudness, especially in the untrained group (Figure 4-5q-t). This is different from what most participants self-reported and from what has been found previously. After time, frequency and loudness are the most relevant inputs for the trained participants and there are several participants for whom frequency is much more relevant than either intensity or loudness (Figure 4-5r-t). The trained

group also has lower noise levels than the untrained, perhaps again implying that they drew in a more predictable manner (Figure 4-5k).

For the GPs with pressure as an output, a similar trend is observed of minimal differences between any of the optimised hyperparameters corresponding to the linear portion of the covariance function (Figure 4-5v-y). However, the linear hyperparameters are slightly more relevant for the trained than for the untrained group, implying that there is enough of a difference in the responses that different model types are able to capture the variability. In the SE hyperparameters, ℓ_{time} is again the most relevant (Figure 4-5aa), though it is closely followed by frequency, intensity and loudness among the trained participants (Figure 4-5bb-dd). In the untrained group, frequency is the next important input, but there is much more variability in the hyperparameters compared to the trained (Figure 4-5aa-dd).

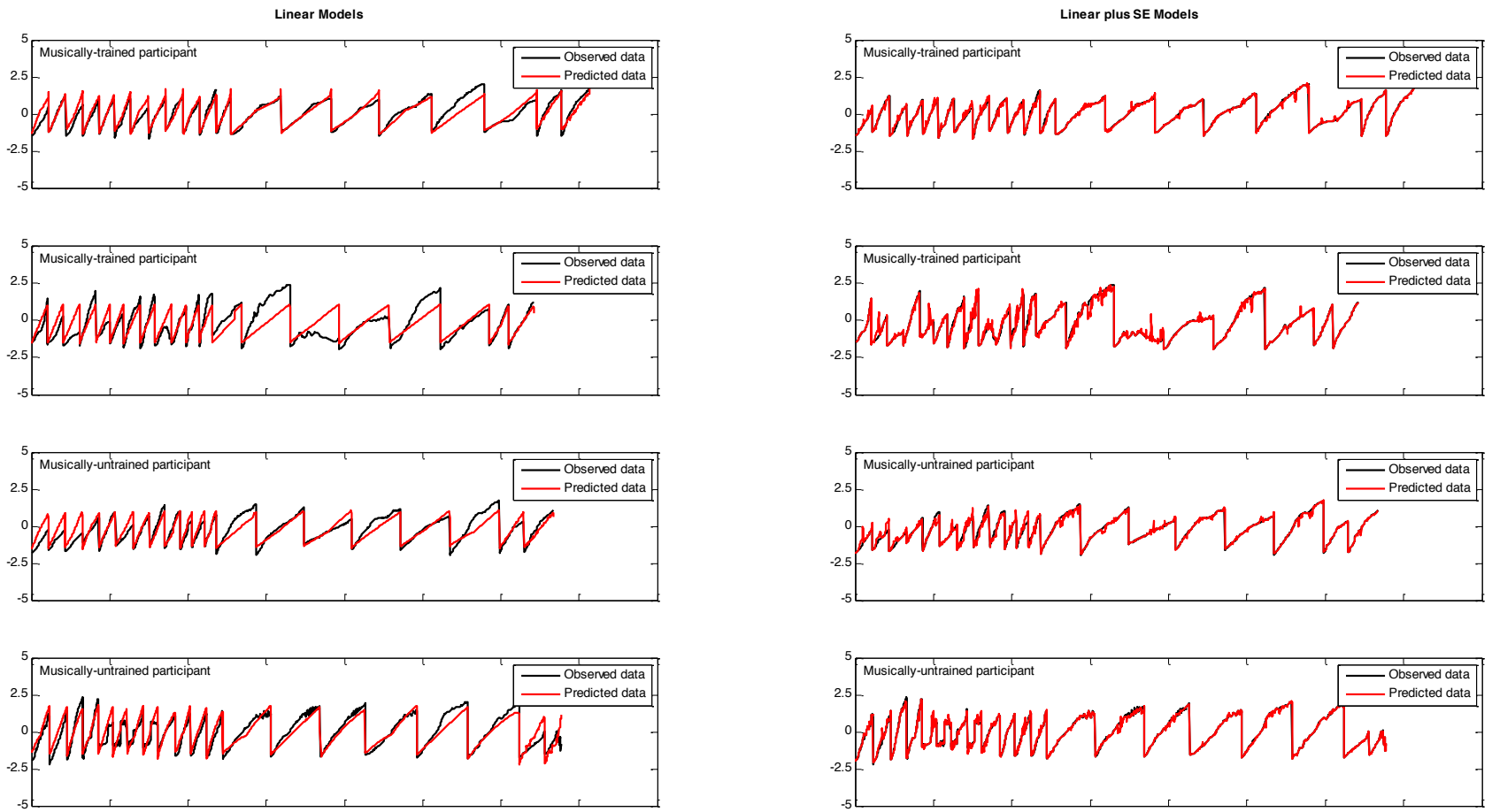


Figure 4-1 Examples of observed X-coordinates (black) compared to predicted X-coordinates (red) using the linear GP regression model (left) and the linear plus SE GP regression model (right) for both musically trained (top rows) and musically untrained participants (bottom rows). The x-axis encompasses all stimuli strung together. The y-axis represents the scaled responses.

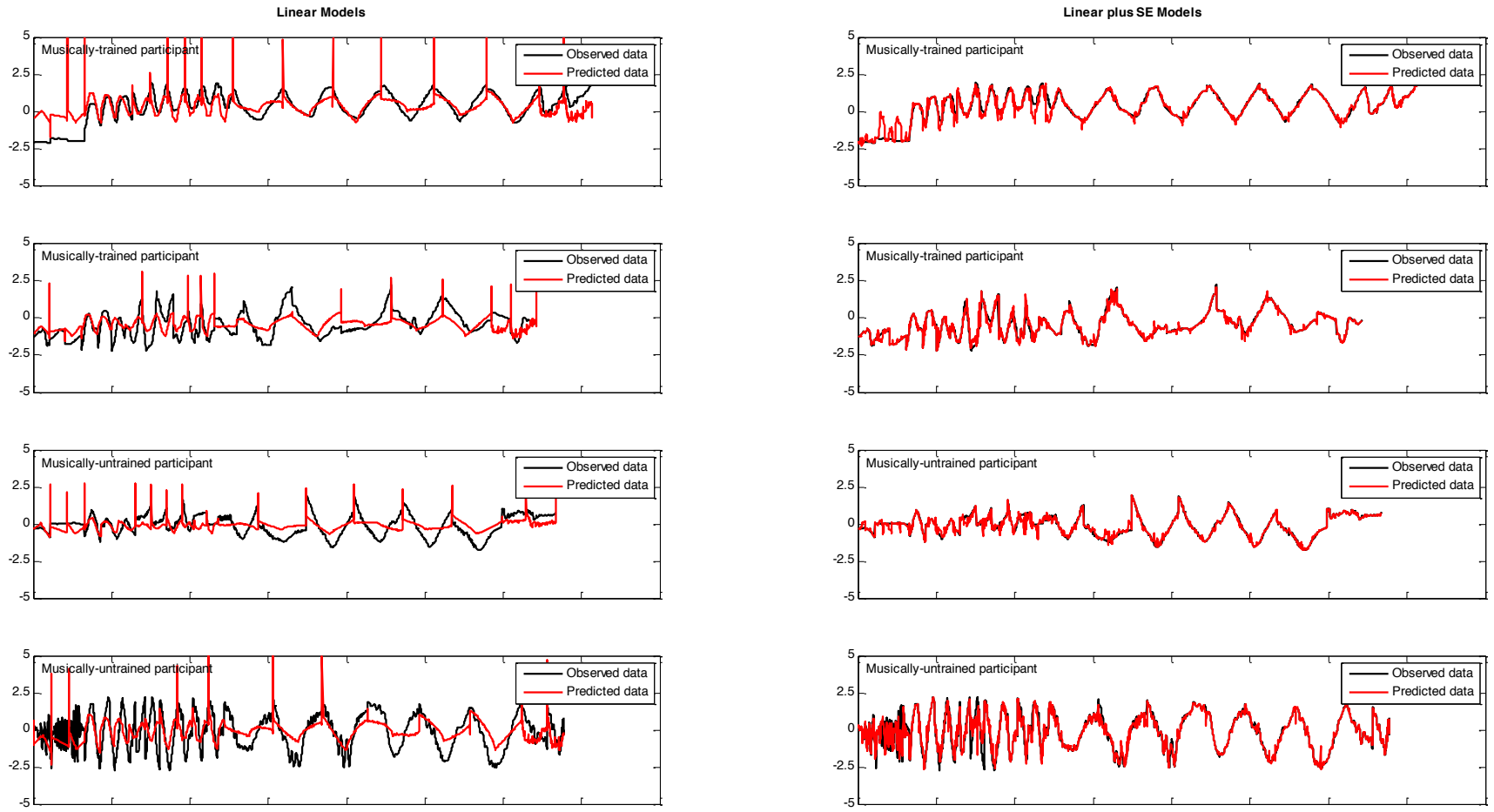


Figure 4-2 Examples of observed Y-coordinates (black) compared to predicted Y-coordinates (red) using the linear GP regression model (left) and the linear plus SE GP regression model (right) for both musically trained (top rows) and musically untrained participants (bottom rows). The x-axis encompasses all stimuli strung together. The y-axis represents the scaled responses.

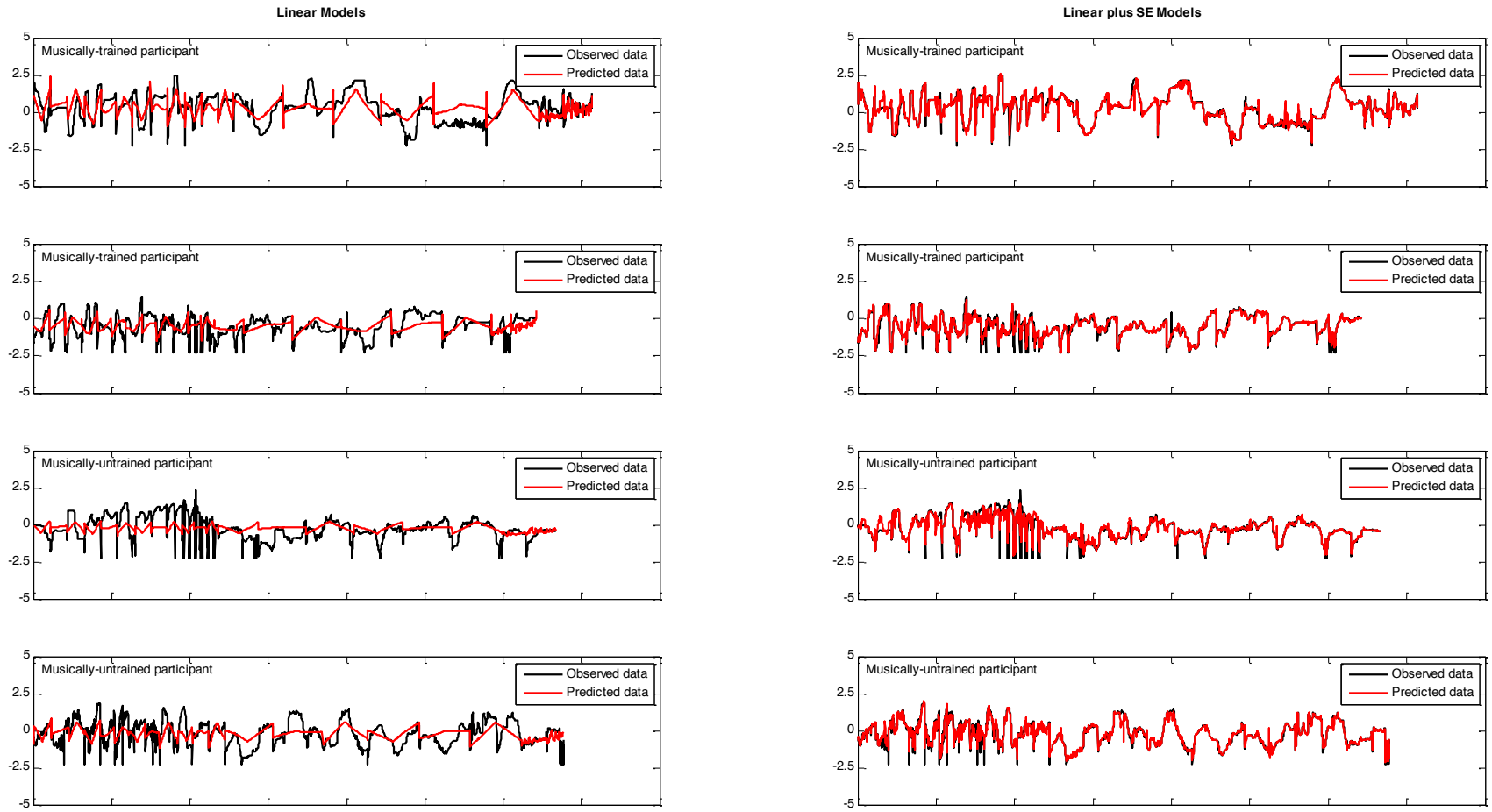


Figure 4-3 Examples of observed pressure values (black) compared to predicted pressure values (red) using the linear GP regression model (left) and the linear plus SE GP regression model (right) for both musically trained (top rows) and musically untrained participants (bottom rows). The x-axis encompasses all stimuli strung together. The y-axis represents the scaled responses.

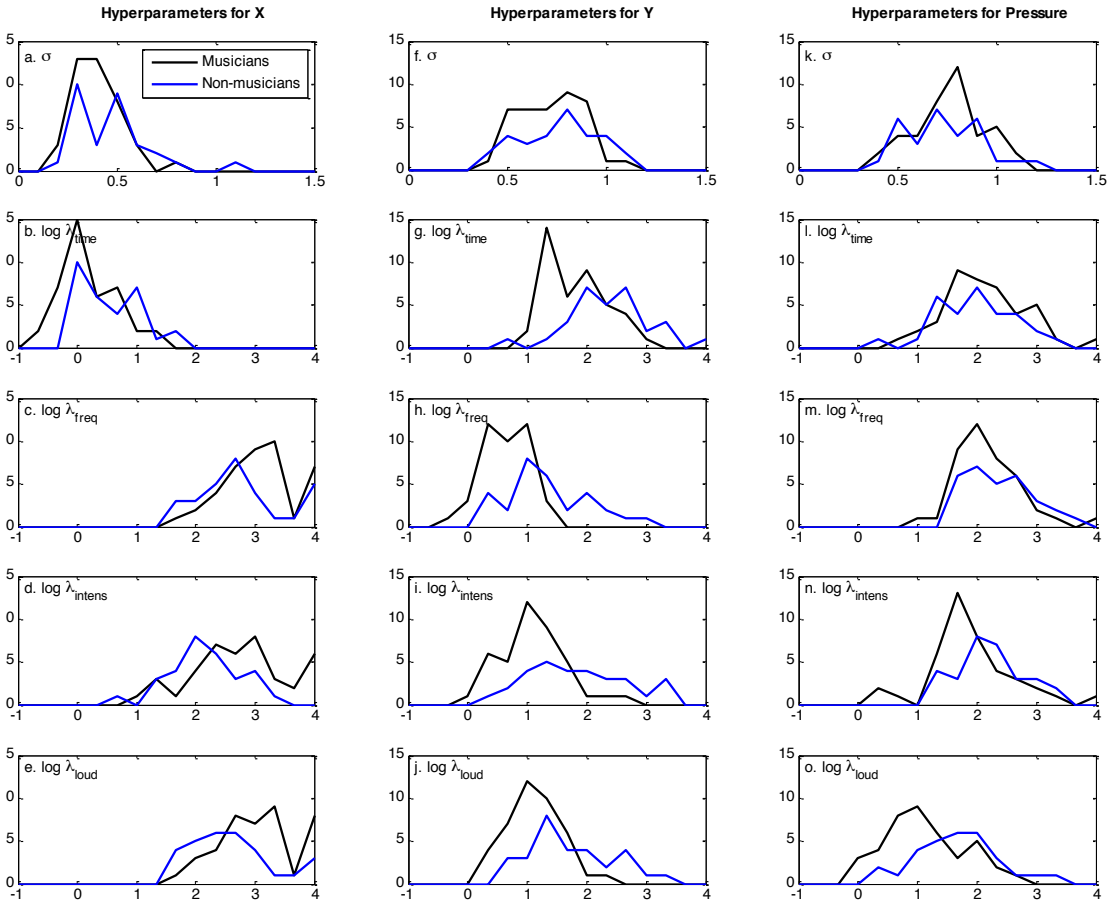


Figure 4-4 Histogram showing the distribution of the optimised hyperparameters from the linear GP regression models with X (a-e), Y (f-j) and pressure (k-o) as outputs. This includes the noise hyperparameter *σ* (first row) and the logged covariance hyperparameters $\log \lambda_{time}$ (second row), $\log \lambda_{frequency}$ (third row), $\log \lambda_{intensity}$ (fourth row) and $\log \lambda_{loudness}$ (fifth row). Note that the covariance hyperparameters are plotted as logs the better to show their distribution, and that the noise and covariance hyperparameters are plotted on different scales. Colours indicate musically trained (black) and musically untrained participants (blue).

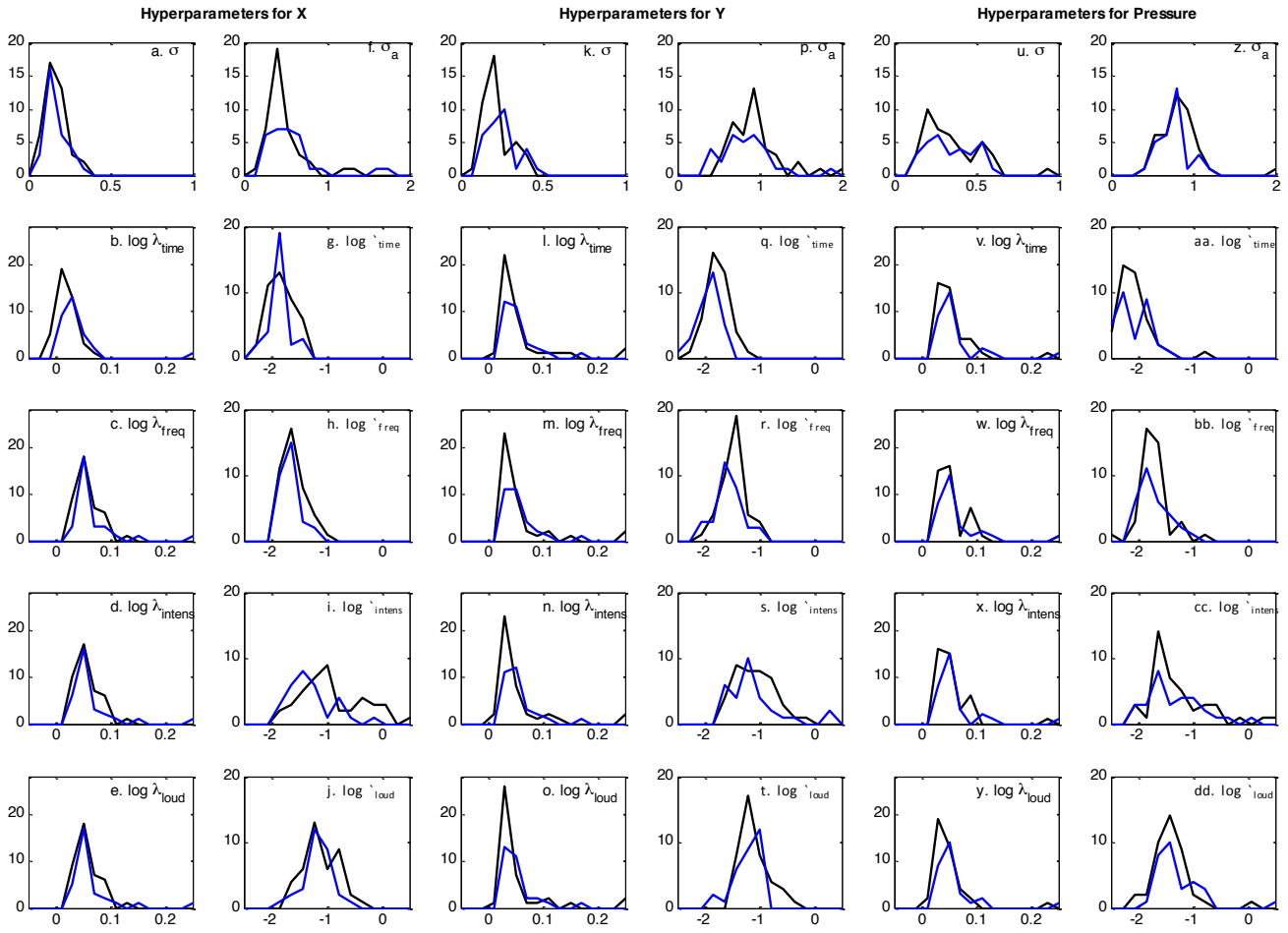


Figure 4-5 Histogram showing the distribution of the optimised hyperparameters from the linear plus SE GP regression models with X (a-j), Y (k-t) and pressure (u-dd) as outputs. This includes the noise hyperparameter σ, the logged covariance hyperparameters $\log \lambda_{time}$ and $\log \ell_{time}$ (second row), $\log \lambda_{frequency}$ and $\log \ell_{frequency}$ (third row), $\log \lambda_{intensity}$ and $\log \ell_{intensity}$ (fourth row), and $\log \lambda_{loudness}$ and $\log \ell_{loudness}$ (fifth row), and the amplitude hyperparameter σ_a. Note that the covariance hyperparameters are plotted as logs the better to show their distribution, and that the noise/amplitude and covariance hyperparameters are plotted on different scales. Colours indicate musically trained (black) and musically untrained participants (blue).

The revealing of time as a particularly relevant input for all three outputs is quite surprising. Time was included to improve the ability to model X and it was not expected to be important for the other outputs. For example, visual inspection of the drawing dataset shows that Y varies considerably with time, sometimes increasing at the beginning of the stimulus, but also sometimes decreasing at the same point in time (or even circling). This encourages the notion that these results may depend on the method of pre-processing the data. Time was the only input that was normalised per stimulus; all other inputs were normalised across the entire dataset. This unique scaling may have caused time to seem more important than if it had been processed like the other inputs. In future work, it might be beneficial to investigate the effect of normalising each input individually, although considerable thought then needs to be put into understanding the effect on the ability to analyse a participant's responses across a variety of stimuli.

Although it is not always simple to pinpoint exactly what all the optimised hyperparameters reveal about the data, their main benefit is in simplifying the dataset. By running this GP analysis, I have created a 33-dimensional space in which each participant can be located, as opposed to the thousands of datapoints that previously corresponded to each individual. These data can then be fed into mathematical algorithms to discern underlying trends in the data, thus allowing the use of powerful analytical tools.

4.4 Clustering

Identifying possible subgroups and trends within the data, either at the level of trained versus untrained individuals or within broader groups, is the focus of attention in the second stage of this set of advanced analyses. There is a variety of mathematical techniques to pull out subgroups from a larger dataset and three of them are discussed here in order of increasing level of complexity. These techniques include exploratory data visualization as well as explicit clustering methods. All analyses were performed on the hyperparameters extracted from the linear plus SE GP regression model.

4.4.1 Principal component analysis

Principal component analysis (PCA) was conducted on the whole set of 33 hyperparameters, as well as separately on the subsets of 11 hyperparameters pertaining to the X, Y or pressure outputs. The main result from the initial PCA was the clear identification of four of the 71

participants as outliers: participants 59, 67, 39 and 73 (figures not shown). Conducting PCA on subsets of the hyperparameters showed that each participant was an outlier for one or two particular inputs. For example, participant 59 is clearly an outlier with regard to the 11 X hyperparameters, but the other three participants are not. Similarly, participants 73 and 67 are outliers for Y, and participants 39 and 67 are outliers for pressure. These participants fall into both the trained and untrained groups.

Because the outliers from this initial run of PCA were skewing the projection such that it was impossible to determine any visual patterns, they were removed from the dataset and PCA was rerun using the remaining participants. While this made the spread of the data clearer, there were no clear patterns in the projected data and the focus is shifted to a more complex analysis.

4.4.2 Spectral clustering analysis

Noyce et al. (2013, p. 140) state that

> "[l]ike PCA, spectral clustering can produce a low-dimensional linear projection of the 33-dimensional hyperparameter space that captures as much of the variance inherent in the data as possible, but it also allows for non-linear projections (von Luxburg, 2007)."

The initial inputs to the spectral clustering analysis were the set of 33 hyperparameters per participant. For subsequent analyses, the 11-hyperparameters subsets were used. Overall, the first step of the spectral clustering analysis effectively reduced the dimensionality of the dataset from 33 (or 11) to 2. A two-dimensional space was chosen for ease of visualization. Note that for PCA, all hyperparameters were standardized to have the same variance across participants, as otherwise large-variance hyperparameters would dominate the principal components. For spectral clustering, raw hyperparameters were used, as the effects of standardization on any nonlinear structures that might be present in the data are less clear.

4.4.2.1 Results and discussion

The outliers from the PCA analysis fit in with the main group in the spectral clustering analysis, which implies that this is a better analytical technique for a dataset of this sort. Nonetheless, there are no clear clusters immediately visible in this projection (Figure 4-6).

One would expect the participants on the extremes of the axis to show some sort of trend, but it was hard to verify this visually. For example, participants 60, 63, 3, 47, 27 and 48 are at the opposite end of the x-axis from 55, 33, 21 and 29, so one would expect their drawings to be relatively different. Most of the participants of the first group drew continuously in a zigzag fashion (see http://www.cmpcp.ac.uk/smip_muvista_slideshow.html) but note that this was not the case for participants 3 and 27. Similarly, most members of the second group produced dotted drawings, but this is not true for participant 55. The extremes on the y-axis range from participants 61, 53, 69 and 35 to participants 57, 9 and 64; visual inspection does however not reveal clear distribution features. Moreover, it does not explain why participants 64 and 57 are also on an extreme of the x-axis but their drawings look nothing like those of participants 42 and 71 (see link to CMPCP website above). Note that the purpose of this cluster analysis is not solely to identify subgroups based on their obvious visual features. To achieve that, it would make more sense to classify the drawings manually according to the various drawing features one is most interested in (e.g., dotted lines). The clustering of these hyperparameters may reveal other meaningful patterns based on features impossible to identify visually, such as small changes in pressure or temporal aspects of the drawings. By applying spectral clustering analysis to only one output at a time, it may be possible to clarify the results from spectral clustering of all the data.

4.4.2.1.1 Pressure hyperparameters

After running the spectral clustering analysis on the subgroups of hyperparameters, some interesting groupings appear. For example, the two groups of participants at opposite ends of the x-axis in the overall spectral clustering analysis are similarly grouped in the pressure analysis (see Figure 4-7a and Figure 4-7d). In addition, the groups of participants found at the extremes of the y-axis in the overall analysis are also separated in the pressure analysis (see Figure 4-7a and Figure 4-7d). Consequently, it is likely that the spread of some of the participants in the original spectral clustering analysis arises from their differing pressure hyperparameters, which is hard to pick up from a visual examination of the drawings.

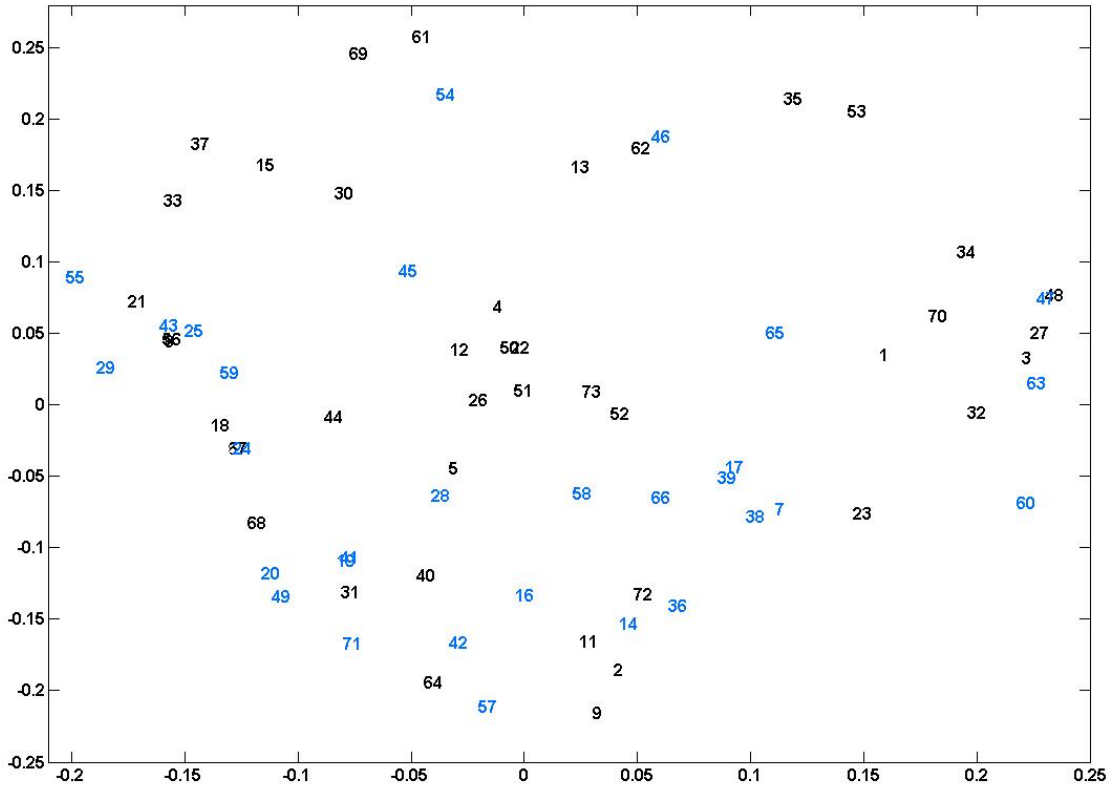


Figure 4-6 Plot of each participant represented in the two-dimensional space created by the second and third eigenvalues of k = 3 spectral clustering using all 33 hyperparameters from the linear plus SE GP regression. The x-axis is the eigenvector associated with the second eigenvalue and the y-axis is the eigenvector associated with the third eigenvalue. The numbers refer to participants—black numbers are musically trained, blue numbers are musically untrained participants.

The main difference between the groups on the x-axis extremes is how well the GP regression was able to model their pressure output. While the left group has some of the smallest noise hyperparameters (ranging from 0.12 to 0.23), the right group has some of the largest among all the participants (between 0.39 and 0.63). The implication is that participants 60, 63, 3, 47, 27 and 48 responded to the sound stimuli in a more predictable manner with respect to the pressure output. Yet, both these groups have a mixture of trained and untrained participants, so the grouping is independent of that division.

4.4.2.1.2 X hyperparameters

Similar trends are visible in the spectral clustering analysis for the X hyperparameters. When using all hyperparameters, participants 53, 35, 69 and 61 are at the opposite end of the y-axis from participants 64, 57 and 9 (Figure 4-7a). In the X spectral clustering, these two groups are

again at opposite extremes (Figure 4-7b). Examining the hyperparameters reveals that the second group has some of the smallest values for $\ell_{intensity}$ while the first group has the largest values of $\ell_{intensity}$ with an average of 0.88 compared to 0.18. While it was hard to pick up a relationship between participants 64, 57 and 9 with purely visual inspection, this analysis implies that these participants may be related in that they drew across the x-axis in response to some part of the stimulus that was not just the passage of time. This is not surprising for participants like 57 (see link to CMPCP website above), who drew circles, but such a relationship is not immediately apparent from the drawings produced by participants 9 or 64 (see link to CMPCP website above).

4.4.2.1.3 Y hyperparameters

For the most part, the groups identified in the overall spectral clustering analysis (Figure 4-7a) are mixed in the spectral clustering output from the Y hyperparameters (Figure 4-7c). The one exception is a slight trend placing participants 9, 64 and 57 towards the top of the y-axis while participants 53, 35, 69 and 61 are towards the bottom, but this distinction is not as clear as in the other spectral clustering results. The spread can be slightly explained by differences in ℓ_{time} for the two groups (averages of 0.14 and 0.18, respectively), but because the overall spread of ℓ_{time} across all 71 participants ranges from 0.07 to 0.29, this difference is not substantial.

These results imply that the distribution in the spectral clustering with all 33 hyperparameters (Figure 4-7a) may be more dependent on the X and pressure hyperparameters than those from the Y analysis. These analyses, however, were conducted for *k* = 3 so that the end result produced two eigenvectors on which to plot. It is possible that if *k* were larger, the resulting distribution might be more related to the distribution of the participants when only the Y hyperparameters are taken into account; at any rate, a larger *k* would be likely to discover different clusters. Overall, it is suggested that spectral clustering analysis may be an improvement over PCA, and that the visual distribution of participants indicates whose hyperparameters are worth investigating more closely. On the other hand, because the embedding produced by spectral clustering is not a simple projection like PCA, the results are more challenging to interpret.

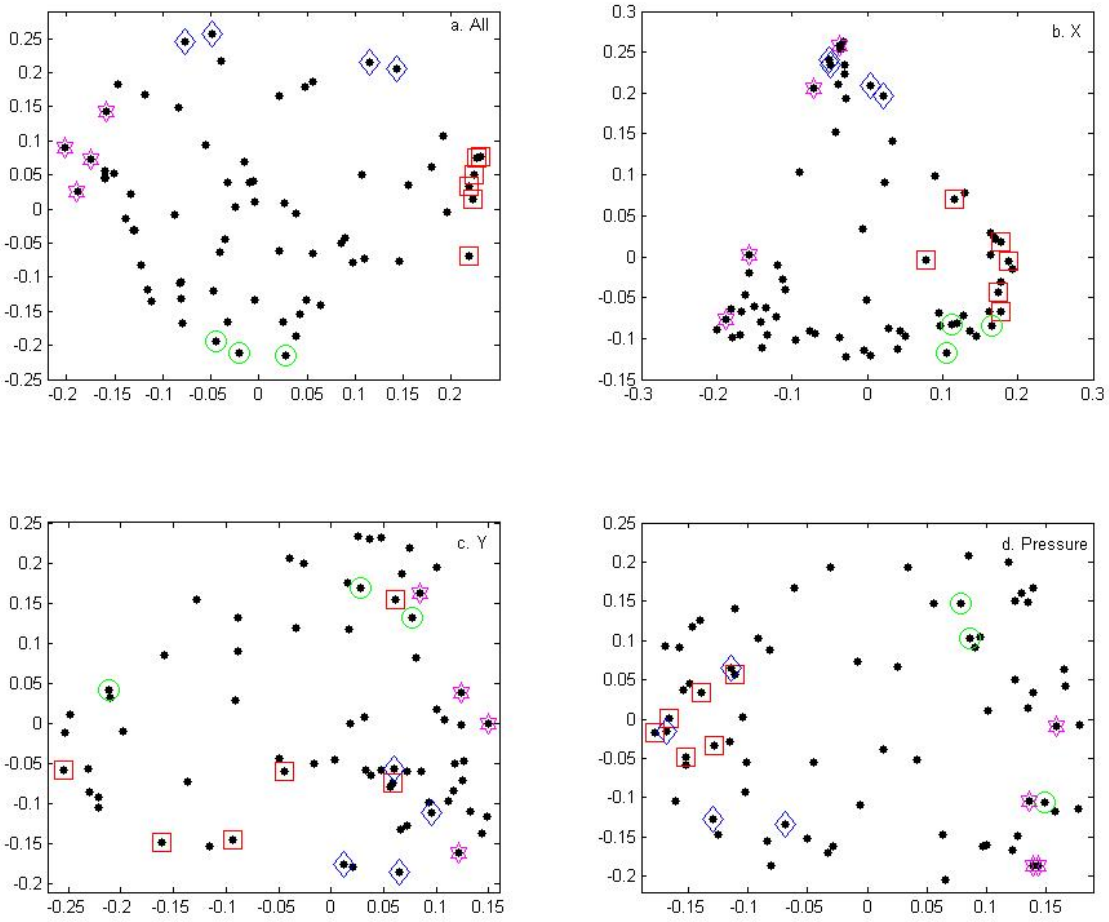


Figure 4-7 Plot of each participant represented in the two-dimensional space created by the second and third eigenvalues of k = 3 spectral clustering using all hyperparameters (a) and using only the 18 hyperparameters from the linear plus SE GP regression for X (b), Y (c), and pressure (d). The x-axis is the eigenvector associated with the second eigenvalue and the y-axis is the eigenvector associated with the third eigenvalue. Each point is a participant. The shapes indicate the groups of participants that fall on the extremes of the overall spectral clustering analysis: blue diamonds are participants 35, 53, 61 and 69; red squares are participants 3, 27, 47, 48, 60 and 63; green circles are participants 9, 57 and 64; and pink stars are participants 21, 29, 33 and 55.

4.4.3 Gaussian mixture models

For the third clustering method, density modelling with Gaussian mixture models (GMMs) was investigated. Noyce et al. (2013, p. 142) note that

> "[a] Gaussian mixture model represents a distribution of datapoints as a superposition of Gaussian distributions with different means (and covariance matrices), each of which can then be associated with a cluster found within the data. To make any statements about whether significant clustering exists one needs a method that determines the number of clusters in the data or, equivalently, the number of Gaussian components.

This is provided by the variational Bayesian (VB) GMM method that was used. The input data were the two-dimensional projections from the spectral clustering analysis described above."

As previously, the analysis procedure was performed four times, once with all the log-hyperparameters and one each with the subsets of hyperparameters for X, Y or pressure – in all four cases after projection to two dimensions by spectral clustering.

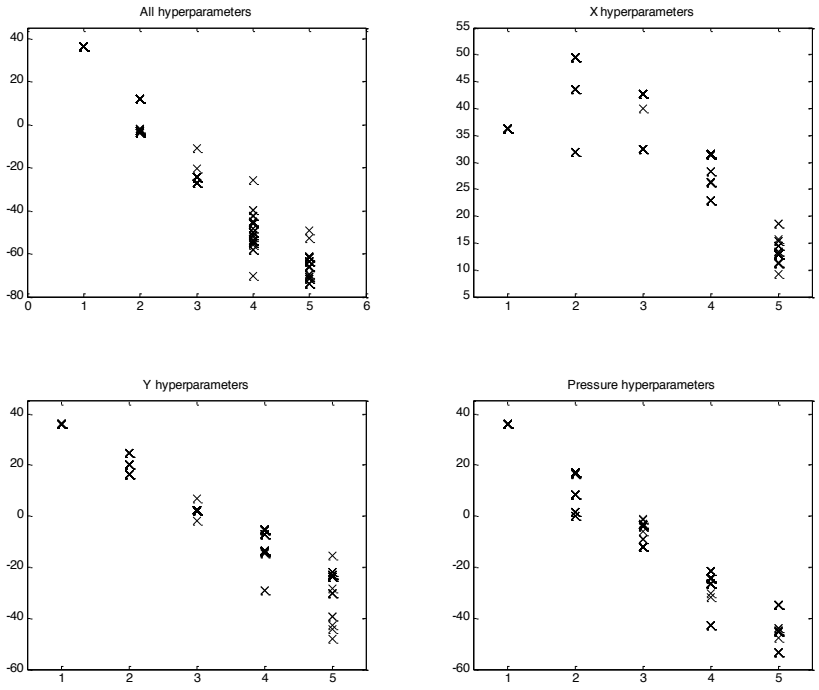
4.4.3.1 Results and discussion

The main benefit of the variational Bayesian approach is that it automatically selects the optimal number (*k*) of Gaussians to fit to the input dataset. In almost all cases, however, the procedure only fitted one Gaussian, which does not identify subgroups (Figure 4-8a). The exception was the X analysis, where the optimal number was two (Figure 4-8a).

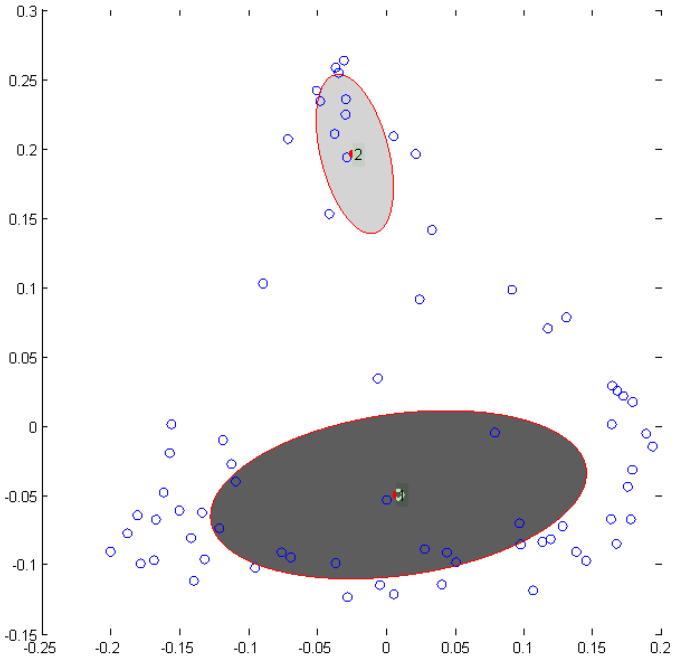
The two Gaussians fit to the X spectral clustering data by the VB analysis are shown in Figure 4-8b – a small cluster at the top and a much bigger cluster that attempts to cover the rest of the data. The smaller cluster is of most interest, especially the participants that fall with the standard deviation density contour, namely participants 15, 30, 37, 62 and 69, all of whom are trained participants. Clearly, there is some sort of consistent response from these participants (with respect to the sound stimulus and the x-coordinate of their drawings) such that they form a separate subgroup. They all have similar values for $\ell_{intensity}$, but so do all the other participants at the top of the y-axis, even those who do not fall within the cluster outlines. Four of them have similar values for λ_{time} (though in the middle of the range, rather than at an extreme), but participant 69 is different enough to prevent this hyperparameter from being the sole reason. For the bottom group (2, 3, 7, 16, 24, 28, 40, 49, 59, 71 and 72, mostly untrained participants), $\ell_{intensity}$ and σ_a are the only hyperparameters that might explain the cluster despite the fact that there is a lot of variability.

Thus, although intensity and time appear to be important predictors for X, this is neither clear from visual inspection nor from the values of the hyperparameters. What is remarkable, however, is the grouping of trained participants at the top and (mostly) untrained participants at the bottom. The next analysis—in line with one of the main hypotheses of this thesis regarding the impact of musical training on music cognition by shapes—will thus examine whether it is

possible to classify trained and untrained participants based on the drawing features represented by the set of hyperparameters.



(a) Plots of the final lower-bound ℓ (y-axis) from 50 random starts of the VB analysis versus the number of Gaussians (k) in the final result (x-axis) when using inputs derived from all hyperparameters and then only the hyperparameters for X, Y, or pressure. The largest ℓ indicates the best fit.



(b) Gaussians fit to the X data using variational Bayes. The centres are the means of the two surviving mixture components and the outlines are the standard deviation density contours.

Figure 4-8 Results from the variational Bayesian analysis

4.5 Classification

GP classifiers were used to find an algorithm for determining whether a given participant is more likely to be trained or untrained. Other classification techniques could have been applied, such as support vector machines, but GP classifiers are a natural choice given the initial use of GP regression; they also provide estimates of uncertainty in the final classification. A GP classifier takes an input vector and predicts the probability with which it belongs to one of two classes. GP classifiers were fitted for each participant by testing both the 18 log-hyperparameters from the linear model and the 33 log-hyperparameters from the linear plus SE model.

4.5.1 Linear classification model

Noyce et al. (2013, p. 145) define the linear classification model as follows:

> "As for the GP regression, the initial GP classifier model choice was simplistic. It was assumed that the probability of the musically trained class was a sigmoidal function (increasing monotonically from 0 to 1) of a 'hidden' function $f(\mathbf{x})$, the latter being modeled by a GP. The simplest case is again to use a linear covariance kernel for $f(\mathbf{x})$:
>
> $$k(\mathbf{x}, \mathbf{x}') = \alpha(\mathbf{x}^T\mathbf{x}' + \sigma_f^2).$$
>
> Because $f(\mathbf{x})$ was fitted as a constant plus linear function, the decision boundary between trained and untrained participants, defined by a threshold value for $f(\mathbf{x})$, will be a hyperplane."

4.5.2 Linear plus SE classification model

Noyce et al. (2013, p. 145) define the linear plus SE classification model as follows:

> "The second model was more complex with an additional isotropic squared-exponential term in the kernel for $f(\mathbf{x})$:
>
> $$k(\mathbf{x}, \mathbf{x}') = \alpha\left(\mathbf{x}^T\mathbf{x}' + \sigma_f^2 \exp\left(-\frac{1}{2\ell^2}(\mathbf{x} - \mathbf{x}')^T(\mathbf{x} - \mathbf{x}')\right) + \sigma_f^2\right).$$
>
> As in the original regression approach, this allows the hidden function $f(\mathbf{x})$ to contain nonlinear contributions. Note that because there are only 71 datapoints, one from each

participant, it would not be viable to fit different length scales for different input directions, and ℓ is therefore taken as common to all directions."

4.5.3 Results and discussion

The GP classifier was set up to predict the probability that each participant belongs to the musically trained class. These probabilities can then be used to create a receiver operating characteristic (ROC) curve, as shown in Figure 4-9. Each point indicates a probability threshold directly related to a threshold for $f(\mathbf{x})$, going from 0 on the bottom left to 1 on the top right. Any participant with a probability above the threshold is assigned to the positive class (i.e. the musically trained class) and then the ROC curve tracks how well the classifier is behaving. The true positive rate is the rate at which the model *correctly* places trained participants in the musically trained class. The false positive rate is the rate at which the model *incorrectly* places untrained participants in the musically trained class. A highly accurate classifier would produce a curve that stays very close to the upper left-hand corner, i.e. has a high rate of accurately classifying trained and a low rate of misclassifying untrained participants.

4.5.3.1 Classification using hyperparameters from the linear GP regression model

From Figure 9a, it is clear that neither of the classifiers trained on the hyperparameters from the linear model accurately predicts whether a given participant is trained or untrained. This holds true for both the linear and SE versions of classifier.

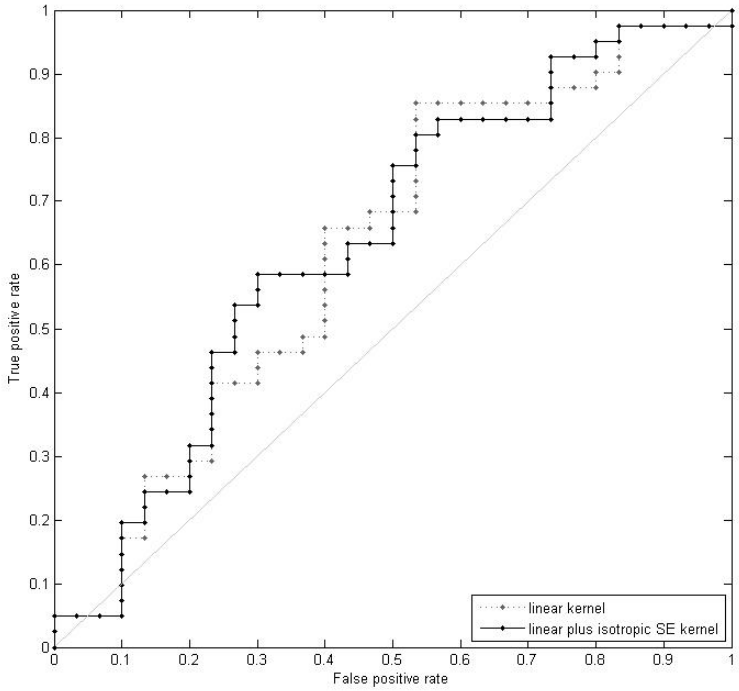
4.5.3.2 Classification using hyperparameters from the linear plus SE GP regression model

In contrast to the first two classifiers, the classifiers trained and tested with the SE hyperparameters are much more accurate at partitioning participants into the correct musically trained or musically untrained class, especially the classifier using a linear plus isotropic SE kernel. As shown in Figure 4-9b, at a probability threshold of 0.51, the true positive rate is 0.80 while the false positive rate is only 0.20. This means that given inputs from 100 trained participants the model would correctly classify 80 as trained, and given inputs from 100 untrained participants the model would correctly classify 80 of those (100 – 20) as belonging to the untrained group.

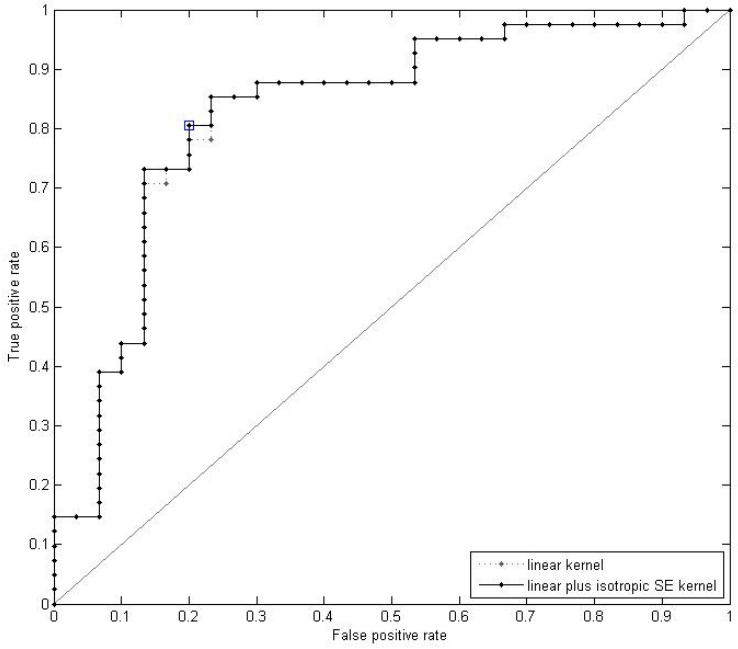
The inaccuracies in this classifier are probably related to the inputs (i.e. the SE hyperparameters) rather than to the design of the classifier itself. To investigate the extent to

which the two groups differ I conducted a 33 × 2 Multivariate Analysis of Variance (MANOVA) with the 33 SE hyperparameters as dependent variables, and 'musical training' as a between-subjects factor. Results showed a highly significant effect (Wilks' λ = .25, $F(33, 37) = 3.34$, $p < .001$, partial $\eta^2 = .75$). Follow-up analysis revealed significant differences between musically trained and untrained participants for the hyperparameters $\ell_{intensity}$ for X ($F(1, 69) = 9.95$, $p = .002$, partial $\eta^2 = .13$) and ℓ_{time} for Y ($F(1, 69) = 9.63$, $p = .003$, partial $\eta^2 = .12$), and marginally significant differences for $\ell_{frequency}$ for Y ($F(1, 69) = 3.48$, $p = .066$, partial $\eta^2 = .05$), the amplitude hyperparameter σ_a for Y ($F(1, 69) = 3.13$, $p = .081$, partial $\eta^2 = .04$) and the noise hyperparameter σ for Y ($F(1, 69) = 2.87$, $p = .095$, partial $\eta^2 = .04$). These results confirm the discussion of the SE hyperparameters (see also 4.3.3.2.2): there was a stronger relationship between X and intensity, time and Y, and frequency and Y in the non-linear portion of the model for musically untrained participants, indicated by their lower hyperparameter values. The lower amplitude hyperparameter of untrained ($M = –.25$, SD = .35) compared to trained participants ($M = –.11$, SD = .31) suggests that drawings along the y-axis of the former group were better captured by the linear portion of the model. However, musically trained participants' lower noise hyperparameters for Y ($M = –1.62$, SD = .38) compared to those of untrained participants ($M = –1.47$, SD = .38) suggests that, overall, musicians' responses along the y-axis were more predictable. As discussed above (and seen in Figure 4-5), there are not always clear differences between the optimised hyperparameters for the trained compared to those for the untrained participants. This most likely indicates a trend inherent in the dataset (i.e. people are not completely predictable!), rather than a failing of the GP regression models.

Nonetheless, this means that any classifier is going to misclassify some of the participants who fall on the border between the two groups. Even though the linear plus SE classifier is not perfect, it is reasonably accurate and can consequently be used to predict whether an unknown drawing was made by a trained or untrained participant.



(a) ROC curve for the two GP classifiers trained using the 18 hyperparameters from the linear GP regression model. The area under the curve is 0.624 for the linear kernel and 0.638 for the linear plus isotropic SE kernel.



(b) ROC curve for the two GP classifiers trained using the 33 hyperparameters from the linear plus SE GP regression model. The square point indicates the optimal probability threshold for classification of 0.51 using a linear plus isotropic SE kernel. The area under the curve is 0.827 for the linear kernel and 0.829 for the linear plus isotropic SE kernel.

Figure 4-9 ROC curves for the GP classifiers

4.6 Summary and conclusion

Overall, even though the techniques used in this study are more complicated than those in the previous analysis of this dataset (see Chapter 3), the additional programming and computing time is decidedly worth it for the richness of the produced output in terms of its ability to explain the data.

It was possible to design a collection of regression GPs that adequately fit the data. Some of the results from the GP regression fit with what was expected, e.g., time as the most relevant input for X or lower noise levels for the trained participants (which implies that they drew in a more predictable manner), but others such as the importance of time for Y and pressure were surprising. At this stage, it remains unclear whether this is due to the pre-processing of the time vector or whether this is a genuine trend in the data. Nevertheless, it illustrates the importance of applying advanced techniques to this dataset, since these trends were not apparent in the previous linear regression analysis.

The main benefit of these analytical techniques is the creation of hyperparameters that give a numerical way to look at responses, similarities, differences and subgroups of participants that one cannot see by visual inspection. The hyperparameters allow us to understand why a participant is different or to find outliers that cannot be identified only by their drawings. To that end, it is worthwhile to fit more complicated GPs to the dataset, both because the richness of the optimised hyperparameters greatly increases the analytical possibilities and because the more complex models fit the data much better.

Though the clustering analysis was inconclusive, that is most likely indicative of trends (or lack thereof) within the dataset itself, rather than deficiencies in the methods used. The spectral clustering analysis provides a visualization of the distribution of participants and allows for the examination of groups or of participants on the extremes that allow us to understand some of the variation and differing responses. Finally, the classification analysis revealed that differentiation between drawings produced by trained and untrained individuals is possible, even if the formal cluster analysis was less successful. There is thus an observable difference, possibly manifested in various characteristics of both the product and process of the visual shaping, in the way in which musically trained and untrained participants approach this task.

By fitting individual GPs and then focusing the rest of the analysis on the optimised hyperparameters, I have shown a way to work with a set of data that is consistent across, and inclusive of, all participants, regardless of their conscious strategies of visualizing sound and music. There are, however, improvements to this method that are worth considering, ranging from altering the methods for pre-processing the data to refining the design of the GPs to account for correlations among the multiple output variables. One could also consider a range of alternative clustering techniques, such as *k*-means applied directly on the spectral clustering embedding, and statistical techniques other than classification for assessing differences between the hyperparameters of the musically trained and untrained groups. As with any analysis, choices for initial parameter values, e.g., in the hyperparameter optimisation, could also affect the results. Future projects could focus on aspects of the data that were neglected in this study, such as subgroups created by sex, age, amount of practice, or musical instrument. In addition, it would be worthwhile to compare participants' verbal reports about how they thought they drew in response to the various input features of this analysis (i.e. frequency and loudness) with the hyperparameters fitted to each participant by the GP model.

With the present set of analyses, I have provided starting points for future studies concerned with time-dependent aspects of cross-modal perception, complementing what is currently reported in the music-psychological literature (e.g., Caramiaux et al., 2010; Nymoen, Torresen, Godøy, & Jensenius, 2012; Schubert, 2004; Vines et al., 2006) and aiming to create a broader variety of analytical tools that can be used to approach data from several angles. Having studied auditory-visual shapes within a real-time drawing paradigm, I will now consider free three-dimensional gestures in the following two chapters.

Chapter 5: Gestural cross-modal mappings of pitch, loudness and tempo in real-time

5.1 Introduction

5.1.1 Origin and shaping of cross-modal correspondences

Research on cross-modal correspondences has shown that people readily map features of auditory stimuli such as pitch and loudness onto the visual or visuo-spatial domain (for reviews see e.g., Eitan, 2013a; Marks, 2004; Spence, 2011). The most extensively studied cross-modal correspondence—that of pitch and spatial height—has produced robust effects revealing that higher (lower) pitch is associated with higher (lower) elevation in space (Ben-Artzi & Marks, 1995; Bernstein & Edelstein, 1971; Bregman & Steiger, 1980; Cabrera & Morimoto, 2007; Casasanto, Phillips, & Boroditsky, 2003; Melara & O'Brien, 1987; A. Miller, Werner, & Wapner, 1958; J. Miller, 1991; Mossbridge, Grabowecky, & Suzuki, 2011; Mudd, 1963; Patching & Quinlan, 2002; Pedley & Harper, 1959; Pratt, 1930; Roffler & Butler, 1968b; Rusconi et al., 2006; Trimble, 1934; R. Walker, 1987; Widmann, Kujala, Tervaniemi, Kujala, & Schröger, 2004). It is unclear, however, what exactly the reason for this cross-modal correspondence is. Different causes of cross-modal mappings have been proposed, e.g., macro-level factors such as development, statistical learning, or culture more generally, and micro-level factors pertaining to experimental paradigms and stimuli selection.

With regard to the impact of culture, there is evidence that the kinds of mappings adults display are influenced by language (Dolscheid, Shayan, Majid, & Casasanto, 2013), emphasizing the importance of conceptual metaphor (Eitan & Timmers, 2010; Johnson & Larson, 2003), which had already been identified by Carl Stumpf as the key mechanism underlying spatial mappings of pitch (Stumpf, 1883, pp. 189-226). Another cultural factor which has been identified as a factor is musical training: trained individuals map auditory features more consistently than untrained individuals, but the kinds of mappings remain consistent across most Western individuals (Eitan & Granot, 2006; Küssner & Leech-Wilkinson, 2014). While culture, and particularly language, thus plays a pivotal role in *shaping* cross-modal correspondences, a growing body of research suggests that their *origin* is to be found elsewhere (but see also Deroy & Auvray, 2013). For instance, studies with infants indicate that 3–4-month-olds show pitch–height and pitch–sharpness associations (P. Walker et al., 2010), 4-month-olds pitch–height

and pitch–thickness associations (Dolscheid, Hunnius, Casasanto, & Majid, 2012), and 3–4-week-olds loudness–brightness associations (Lewkowicz & Turkewitz, 1980). Combined with evidence from audio-visual mappings in non-human mammals (Ludwig et al., 2011), this has led some scholars to conclude that cross-modal correspondences are innate, possibly based on a wide range of neural connections that are gradually lost due to synaptic pruning (Mondloch & Maurer, 2004; K. Wagner & Dobkins, 2011). Others have argued that cross-modal correspondences may be learned rapidly through external, non-linguistic stimulation (Ernst, 2007; as discussed in Spence, 2011) or may be acquired indirectly in cases where the occurrence of cross-modal pairings in the environment seems unlikely (Spence & Deroy, 2012). Further evidence supporting the prelinguistic origin hypothesis comes from studies showing that cross-modal mappings are processed at an early, perceptual level (Evans & Treisman, 2010; Maeda, Kanai, & Shimojo, 2004), unmediated by later, semantic processing (but see also Chiou & Rich, 2012; Martino & Marks, 1999).

5.1.2 Complexity of audio-visuo-spatial correspondences

As implied above, cross-modal correspondences between auditory features and the visuo-spatial domain are manifold, sometimes referred to as one-to-many and many-to-one correspondences (Eitan, 2013a). For instance, pitch has been associated with vertical height (R. Walker, 1987), distance (Eitan & Granot, 2006), speed (P. Walker & Smith, 1986), size (Mondloch & Maurer, 2004) and brightness (Collier & Hubbard, 1998)—i.e. one-to-many—while the same associations have been found for loudness (Eitan, Schupak, & Marks, 2008; Kohn & Eitan, 2009; Lewkowicz & Turkewitz, 1980; Lipscomb & Kim, 2004; Neuhoff, 2001), rendering, for example, pitch/loudness–height a many-to-one correspondence. The full story is, however, more complex than that, as outlined in Eitan (2013a).

First, the type of auditory stimuli, whether static or dynamic, can give rise to opposing results. For instance, static high and low pitches paired with small and large visual disks, respectively, have been shown to enhance performance in a speeded classification paradigm (Gallace & Spence, 2006), providing evidence that high pitch is associated with small objects and low pitch with large objects. On the other hand, Eitan and collaborators (2014), using a similar paradigm, demonstrated that rising pitches paired with an increasing visual object and falling pitches paired with a decreasing visual object yielded significantly faster responses than rising pitches paired with a decreasing visual object and falling pitches with an increasing visual object.

Secondly, manipulating several auditory features concurrently influences participants' cross-modal images of motion (Eitan & Granot, 2011). For instance, an increase in tempo, usually associated with an increase in speed, did not lead to an increase in speed when loudness was concurrently decreasing. Similarly, a rise in pitch, usually associated with an increase in vertical position, led to a *decrease* in vertical position when loudness was concurrently decreasing.

Since environmental sounds, but especially music, are very often varied *dynamically and concurrently* in pitch, loudness, tempo, timbre etc., investigating cross-modal correspondences of these features—which is frequently done by manipulating them in isolation, entailing obvious experimental advantages but also the even more obvious lack of ecological validity—requires approaches taking into consideration the multiple dynamic co-variations of sound features.

5.1.3 Cross-modal mappings of sound involving real or imagined bodily movements

Whereas most experimental paradigms to date have used speeded identification, speeded classification, or forced-choice matching tasks, researchers have recently begun to apply paradigms involving real-time drawings (Küssner & Leech-Wilkinson, 2014), gestures (Kozak et al., 2012) or imagined bodily movements (Eitan & Granot, 2006), in order to delineate a more differentiated picture of cross-modal mappings. Asking participants to imagine the movements of a humanoid character in response to changes in a range of musical parameters, Eitan and Granot (2006) found that pitch is mapped onto all three spatial axes, including asymmetric pitch–height mappings such that decreasing pitch was more strongly associated with descending movements than increasing pitch with ascending movements. Similarly, the authors report two asymmetric mappings of loudness: (1) decreasing loudness was more strongly associated with spatial descent than increasing loudness with spatial ascent, and (2) increasing loudness was more strongly associated with accelerating movements than decreasing loudness with decelerating movements. What is more, results from a study investigating participants' perceptions of the congruency between vertical arm movements and changes in pitch and loudness revealed that concurrent rising-falling movements of one's arm and pitch or loudness gave rise to higher ratings than concurrent falling-rising movements (Kohn & Eitan, 2012). These striking asymmetries might be part of a discrepancy between response time or rating paradigms and those involving more extensive, overt bodily movements.

A few studies only have investigated how changes in auditory stimuli are mapped onto real bodily movements. In an exploratory study, Godøy, Haga and Jensenius (2006a) asked participants to respond with hand gestures—captured with a pen on an electronic graphics tablet—to a set of auditory stimuli that comprised instrumental, electronic and environmental sounds and was classified according to a typology developed by Pierre Schaeffer (e.g., impulsive, continuous and iterative sounds). While the authors report a "fair amount of consistency in some of the responses" such as ascending movements for increasing pitch, they do stress the need for large-scale studies involving the investigation of free movements in three-dimensional space as well as of the influence of musical training. In a subsequent study from the same group, Nymoen, Caramiaux, Kozak and Torresen (2011) found strong associations between pitch and vertical movements, between loudness and speed, and between loudness and horizontal movements, when comparing people's gestural responses to pitched and non-pitched sounds, captured by moving a rod whose movements were supposed to represent sound-producing gestures. While the authors' argument for "a one-dimensional intrinsic relationship between pitch and vertical position" is conceivable in view of their findings, the lack of bidirectional pitch changes (e.g., rising-falling contour) within their auditory stimuli precludes conclusions about potential asymmetric mappings of pitch with bodily movements.

In a similar experiment, Caramiaux and colleagues (2014) compared hand gestures in response to action and non-action related sounds, confirming their hypothesis that the former would entail sound-producing gestures while the latter would result in gestures representing the sound's spectromorphology (Smalley, 1997), i.e. the overall sonic shape. Comparing speed profiles between participants revealed that they were more similar for non-action- than action-related sounds. This shows—and is supported by analysis of interviews carried out with the participants—that once a particular action (e.g., crushing a metallic can) has been identified, the realization of the accompanying gesture is highly idiosyncratic. On the other hand, non-action-related sounds, which are particularly pertinent to the present study, gave rise to more consistent gestural responses.

One study has been carried out investigating free representational movements to sound, in which 5- and 8-year-old children were presented with auditory stimuli separately varied in pitch, loudness and tempo (Kohn & Eitan, 2009). Three independent referees trained in Laban Movement Analysis rated the observed behaviour—the sound being muted—according to the

movement and direction along the x-, y- and z-axes, the muscular energy, and the speed. Pitch was most strongly associated with the vertical axis, loudness with vertical axis and muscular energy, and tempo with speed and muscular energy. In terms of direction, changes in loudness and tempo gave rise to congruent movement patterns, that is, increasing loudness was represented with upward movement and higher muscular activity, whereas decreasing loudness was represented with downward movement and lower muscular activity. The direction of movement along the vertical axis in response to changes in pitch was congruent for increasing-decreasing pitch contours but not for decreasing-increasing contours. This finding is particularly relevant for the present study, as it highlights the asymmetric nature of bodily cross-modal mappings.

5.1.4 The roles of elapsed time and visual feedback

One aspect often neglected when studying mappings of sound is the representation of elapsed time or duration (but see R. Walker, 1987). Although there are no linguistic metaphors suggesting lateral mapping of time (Casasanto & Jasmin, 2012), people do use gestures laterally to refer to the sequence of events, whereby 'left' refers to earlier and 'right' refers to later in time (Cooperrider & Núñez, 2009). Note that the lateral mapping of time is dependent on the direction of reading and writing but can also be reversed experimentally (Casasanto & Bottini, 2010). In the realm of cross-modal mappings of music and sound, progression from left to right has been found in drawing experiments (Küssner & Leech-Wilkinson, 2014; Tan & Kelly, 2004), with some variation cross-culturally (Athanasopoulos & Moran, 2013), but further experiments are needed to investigate how elapsed time of sound and music is represented gesturally.

Another aspect that, to my knowledge, has received little to no attention by researchers is the impact of visual feedback on gestural cross-modal mappings. However, some authors have investigated the more general role of vision in audio-visual correspondences. Various studies have shown that differences of metaphorical cross-modal mappings of sound between sighted and blind participants are negligible (Antović, Bennett, & Turner, 2013; R. Walker, 1985; Welch, 1991). And relating changes in sound features to *bodily* movements has revealed the absence of pitch–height mappings in congenitally blind individuals (Eitan et al., 2012). Thus, to account for the possibility that vision—or conceivably the availability of visual feedback—plays an important role when studying gestural cross-modal correspondences, I introduced a

visualization condition in which participants' gestures created a real-time visualization on a screen in front of them. Rather than comparing sighted and blind participants, I intended to raise the awareness of normally sighted people's gestures by this visualization, as well as enabling them to manipulate the size of a visual object, to investigate whether this affects their cross-modal mappings, for instance in response to changes in loudness.

5.1.5 Aims and novelties

The present study aims to identify how pitch, loudness and tempo are represented gesturally in real-time, and to what extent training and visual feedback influence those cross-modal mappings. To my knowledge, this is the first controlled experiment studying adults' gestural responses to a set of pure tones systematically and concurrently varied in pitch, loudness and tempo. Based on the literature reviewed above, I hypothesise the following outcomes:

1. Elapsed time is represented on the x-axis, proceeding from left to right.
2. Pitch is represented on the y-axis (higher elevation for higher pitches); rising-falling pitch contours (convex shapes) are expected to yield greater pitch–height associations than falling-rising pitch contours.
3. Loudness is represented with forward-backward movements along the z-axis and muscular energy (forward movement / more energy for louder sounds), as well as with spatial height when loudness is the only auditory feature being manipulated (higher elevation for louder sounds). The visual condition is expected to give rise to loudness–size associations, enabled by coupling the movement along the z-axis with size of the visualization on the screen (larger size [= forward movement] for louder sounds).
4. Tempo of pitch change in the auditory stimuli is represented by speed of the hand movements (faster movement for faster tempo) and muscular energy (more energy for faster tempo).
5. Musical training has an impact such that musically trained participants show generally more consistent mappings than musically untrained participants.

5.2 Methods

5.2.1 Participants

Sixty-four participants (32 female) took part in the experiment (age: *M* = 29.63 years, SD = 12.49 years, range: 18–74 years). Thirty-two participants (16 female) were classified as musically trained (age: *M* = 30.09 years, SD = 13.66 years, range: 18–74 years), and 32 (16 female) as musically untrained (age: *M* = 29.16 years, SD = 11.39 years, range: 18–67 years). All participants were required to be 18 years or over, right-handed, and must not have been diagnosed with any vision or hearing impairments (except those corrected to normal vision with glasses or contact lenses). To satisfy the 'musically trained' category, participants must have played either a keyboard instrument, a string instrument, a wind/brass instrument or been a composer,[51] must have had at least Grade 8 of the ABRSM system (http://gb.abrsm.org/en/home) or an equivalent qualification, and must have spent at least four hours per week on average playing their respective main instrument or composing. All musically trained participants were balanced by sex and main musical activity. Musically untrained participants must not have played any musical instrument or composed music for the past six years, must not have played any instrument for more than two years in total, and must not have exceeded Grade 1 ABRSM. Participants were recruited using a college-wide e-mail recruitment system including undergraduates, postgraduates and staff, as well as circulating a call for participants within music conservatoires. All criteria were clearly stated in the recruitment email and checked again with a questionnaire during the experiment. Exceptions included one trained participant who reported playing only two hours on average per week and one untrained participant who reported engaging in musical activities ("electronics, drums, mixing") for 7.5 hours. Another musically untrained participant had played the guitar for four years in total but had stopped playing fourteen years ago, and one untrained participant who had played drums for one year had only stopped five years ago. Since this study is concerned with differences arising from formal training, and none of the musically untrained participants had taken any formal music examination while all musically trained participants were at Grade 8 or above, it was decided to keep all participants for the analysis to ensure a balanced design and sufficient statistical power. None of the participants had taken part in the drawing experiment.

[51] Note that the time participants had been composing varied from 3 years (starting at age 16) to 41 years (starting at age 6). The number of years spent composing is not necessarily an indicator of quality though: for instance, the composer with the fewest number of years of experience had already received the prestigious BBC Young Composer of the Year Award.

5.2.2 Stimuli

Stimuli (see Table 5-1 and Figure 5-1) were synthesized in SuperCollider (Version 3.5.1) and consisted of twenty-one continually sounding pure tones that varied in frequency, amplitude and tempo.[52] All stimuli were eight seconds long. Trough and peak pitches were B2 (123.47 Hz) and D4 (293.67 Hz), respectively, and all but three stimuli (Nos 1–3) had a rising-falling (Nos 4–12) or falling-rising (Nos 13–21) pitch contour. While Nos 1–3 had a constant pitch (D4), No. 1 (as well as Nos 4–6 and Nos 13–15) had a constant amplitude with 50% of the maximum, No. 2 (as well as Nos 7–9 and Nos 16–18) was linearly decreasing and increasing in amplitude (90% – 10% – 90%; reaching 10% after four seconds), and No. 3 (as well as Nos 10–12 and Nos 19–21) was linearly increasing and decreasing in amplitude (10% – 90% – 10%; reaching 90% after four seconds). Nos 4, 7, 10, 13, 16 and 19 changed pitch linearly, reaching the top (bottom) after three seconds, before going into the opposite direction after one second and reaching the bottom (top) after three seconds and staying there for another second. The tempo of Nos 5, 8, 11, 14, 17 and 20 was decreasing with factor –0.5 twice, resulting in two decelerandi: the first four seconds until the top (bottom) and the second four seconds when returning to the bottom (top). By contrast, the tempo of Nos 6, 9, 12, 15, 18 and 21 was increasing with factor 0.5 twice, resulting in two accelerandi.

[52] The sound stimuli can be downloaded at http://tinyurl.com/nqmn3ej.

Table 5-1 Overview of experimental sound stimuli

No.	Frequency (Note name)	Amplitude	Tempo
1	constant (D4)	constant	N/A
2	constant (D4)	decreasing - increasing	N/A
3	constant (D4)	increasing - decreasing	N/A
4	rising - falling (B2–D4–B2)	constant	equal
5	rising - falling (B2–D4–B2)	constant	decelerando - decelerando
6	rising - falling (B2–D4–B2)	constant	accelerando - accelerando
7	rising - falling (B2–D4–B2)	decreasing - increasing	equal
8	rising - falling (B2–D4–B2)	decreasing - increasing	decelerando - decelerando
9	rising - falling (B2–D4–B2)	decreasing - increasing	accelerando - accelerando
10	rising - falling (B2–D4–B2)	increasing - decreasing	equal
11	rising - falling (B2–D4–B2)	increasing - decreasing	decelerando - decelerando
12	rising - falling (B2–D4–B2)	increasing - decreasing	accelerando - accelerando
13	falling - rising (D4–B2–D4)	constant	equal
14	falling - rising (D4–B2–D4)	constant	decelerando - decelerando
15	falling - rising (D4–B2–D4)	constant	accelerando - accelerando
16	falling - rising (D4–B2–D4)	decreasing - increasing	equal
17	falling - rising (D4–B2–D4)	decreasing - increasing	decelerando - decelerando
18	falling - rising (D4–B2–D4)	decreasing - increasing	accelerando - accelerando
19	falling - rising (D4–B2–D4)	increasing - decreasing	equal
20	falling - rising (D4–B2–D4)	increasing - decreasing	decelerando - decelerando
21	falling - rising (D4–B2–D4)	increasing - decreasing	accelerando - accelerando

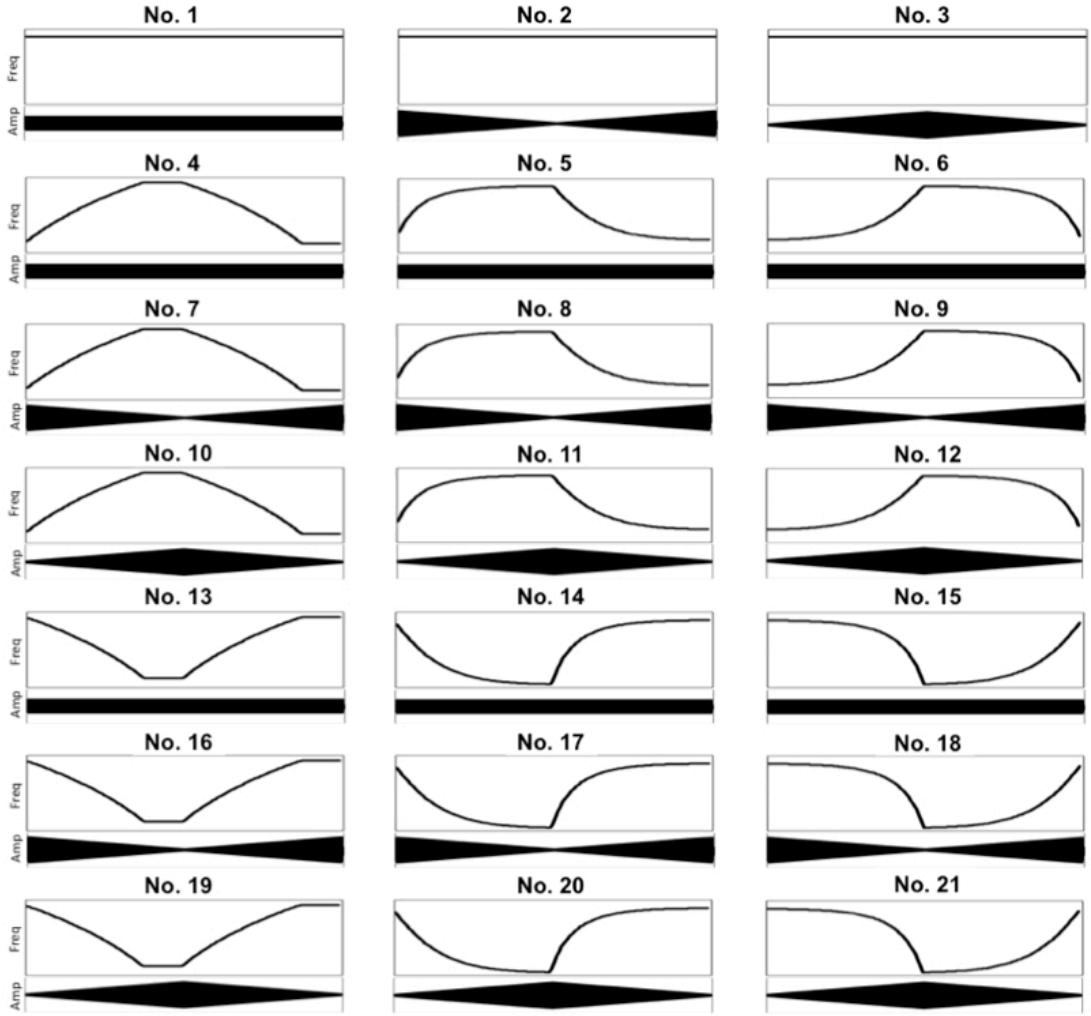


Figure 5-1 Overview of frequency and amplitude contours of experimental sound stimuli. All x-axes represent time (length of stimuli: eight seconds). Highest/lowest frequency: 123.47 Hz / 293.67 Hz. Equal amplitude means 50% of the maximum, decreasing amplitude means 90% to 10% of the maximum and increasing amplitude means 10% to 90% of the maximum. Freq: log frequency (Hz); Amp: amplitude.

5.2.3 Motion capture

To capture participants' hand movements, a Microsoft® Kinect™ and a Nintendo® Wii™ Remote Controller were used. The bespoke software for the purposes of this experiment was developed in Processing v1.2.1 (Fry & Reas, 2011). The whole experimental session was also recorded with two video cameras (Panasonic HDC-SD 700/800), one filming the participant frontally (whole body) and the other capturing the visualizations on a screen in front of the participants (see Procedure). The visualization was based on data from the Kinect™, recording the position of the hand. Additionally, participants held the Wii™ Remote Controller in the same hand that was performing the gestures. If the latter was shaken strongly enough for the

acceleration threshold of 10 m/s^2 to be exceeded, then input data for the visualization included Wii^{TM} data as well to reflect the intensity of the shaking hand movements.

5.2.4 Procedure

A schematic overview of the procedure can be seen in Figure 5-2 below. After signing the consent form, participants read detailed instructions and any remaining uncertainties were discussed with the experimenter. Participants were introduced to the Kinect^{TM} and Wii^{TM} Remote Controller technologies, made aware of the experimental space, and familiarized with the noise-cancelling headphones to be worn during the experiment (Bose QuietComfort® 15 Acoustic Noise Cancelling®). The participants' task was to represent the sound stimuli with their right hand in which they held the Wii^{TM} Remote Controller; it was stressed that (a) there were no 'right' or 'wrong' responses, (b) participants' responses should be consistent such that, if the same sound occurred twice, they should make the same movement, and (c) they should try to represent gesturally all sound characteristics they are able to identify.

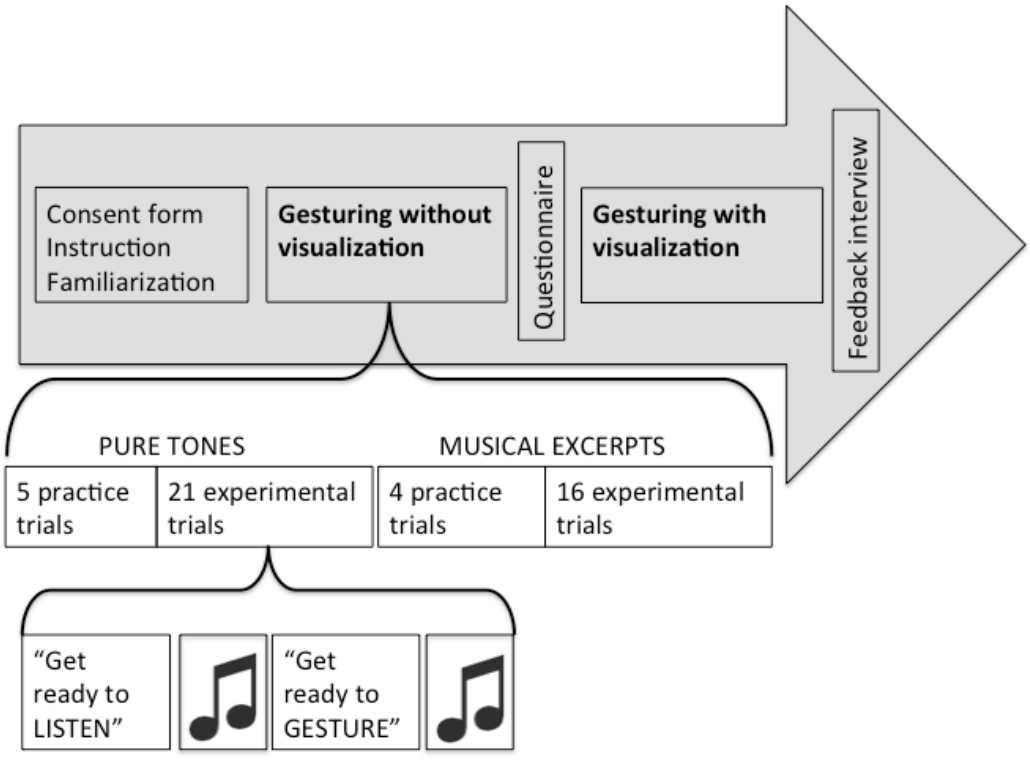


Figure 5-2 Overview of experimental procedure

The experiment itself consisted of two parts. In the first part, participants gestured without seeing a visualization, whereas in the second part, their gestures created a real-time visualization on a screen in front of them (see also Chapter 2). The visualization consisted of a black disk with a thin green edge (see Figure 5-3). The size of the disk could be increased by moving one's hand towards the screen, or decreased by moving backwards. Movement towards the right (left) from the participant's perspective resulted in a rightward (leftward) motion on the screen. Similarly, raising (lowering) one's hand resulted in the disk going up (down). To achieve a three-dimensional effect (cf. 'Performance Worm' by Dixon, Goebl, & Widmer, 2002), the decay rate of the disk at any position was 60 frames (ca. 4 seconds); during the decay, the disk gradually lightened before disappearing completely. There was a trade-off between the lag of the visualization and its smoothness, which was adjusted in various test sessions, resulting in a lag in the range of 100–150 ms.

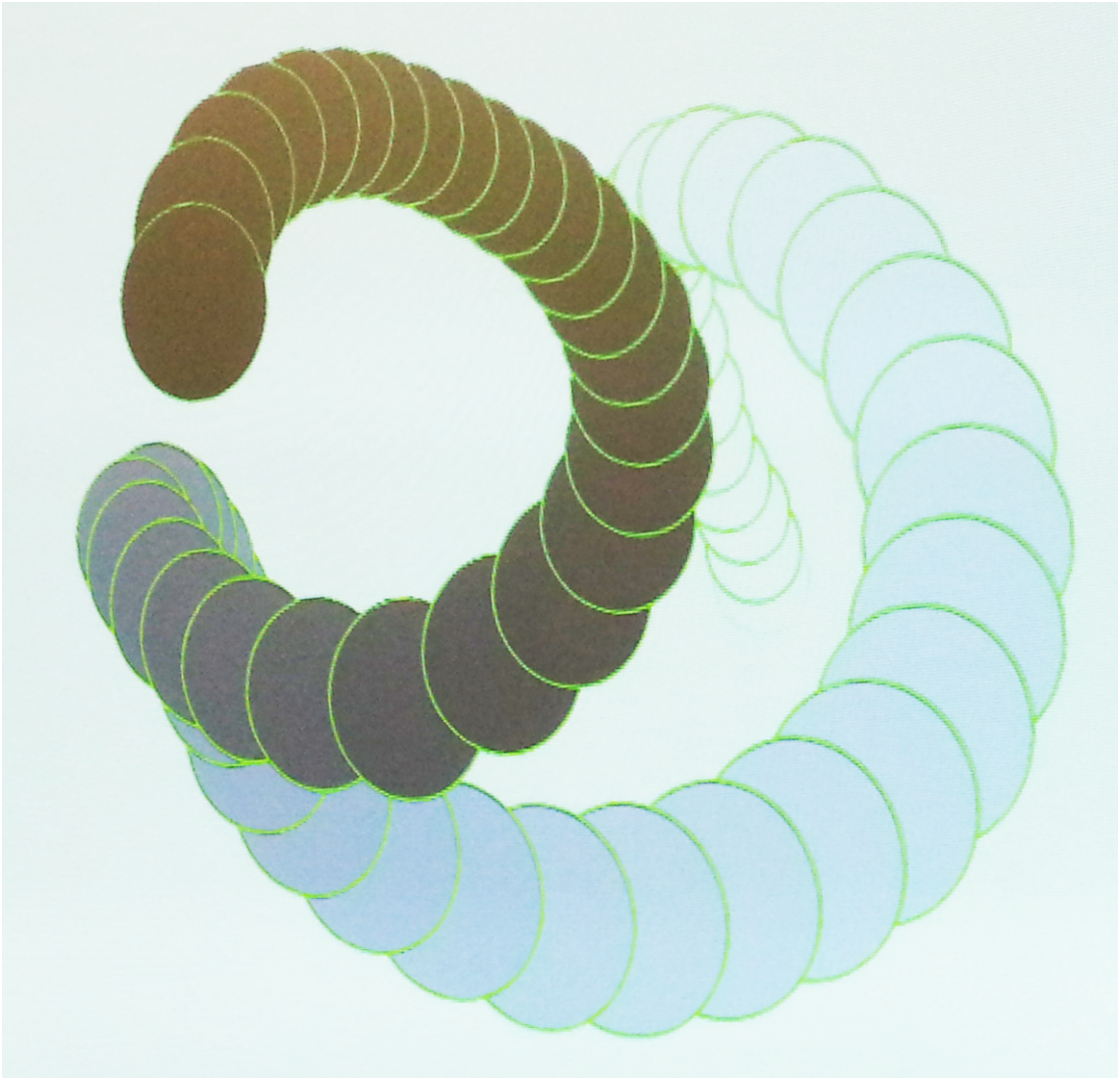


Figure 5-3 Real-time visualization on screen in front of participants

Each part consisted of the same blocks in the same order, while the presentation order of stimuli within the blocks was randomized. After a short calibration procedure with the Kinect™ to identify and track the participants' right hand, a summary of the instructions appeared on the screen. Once participants were ready, they informed the experimenter, who was seated behind another screen and was not able to see their movements, and the first block—practice trials consisting of five pure tones—was started. If participants did not have any further questions after the practice trials (they could repeat the practice trials as often as they wished), the second block consisting of all twenty-one pure tones was started. Participants were presented with each stimulus twice consecutively. The first time, they were supposed to listen only: two seconds prior to the stimulus onset the instruction "Get ready to LISTEN. X stimuli left. [countdown]" appeared in the upper left corner of the screen, informing participants about the number of

stimuli left in this block (X) and starting a short countdown. The second time, participants were supposed to represent the sound stimulus gesturally while it was played. The instruction "Get ready to GESTURE. [countdown]" appeared and participants were again prepared for the onset of the stimulus with a countdown. The third and fourth block consisted of practice and experimental trials with short musical excerpts, which will be analysed and discussed in Chapter 6. In the second part, the calibration process was followed by a free mode during which participants were encouraged to "enjoy and explore" the visualization without accompanying sound prior to the start of the practice trials with pure tones. This procedure had been approved by the College Research Ethics Committee (REP-H/10/11-13).

5.2.5 Data analysis

The sound features—frequency in Hz and loudness in sone, sampled at 20 Hz each—were extracted with Praat version 5.3.15 (Boersma & Weenink, 2012). Frequency values were log-transformed to account for human perception of pitch, as is common practice in psychophysical experiments (e.g., see Micheyl, Delhommeau, Perrot, & Oxenham, 2006, p. 39). Both log-transformed frequency values and loudness values were then standardized (M = 0, SD = 1) per sound stimulus. The Kinect™ data—X, Y and Z coordinates sampled at ca. 15 Hz[53]—were extracted together with their timestamps. All three spatial coordinates were then standardized (M = 0, SD = 1) per sound stimulus. Next, sound features were linearly interpolated to realign them with the movement data at the timestamps of the Kinect™ data, creating a matrix with six columns (timestamp, frequency, loudness, X, Y, Z) per stimulus.

It would have been possible to use GPs for the current dataset, as the nature of data is very similar to the one described in Chapter 4. In fact, I have run the GP regression model on the current dataset and achieved similarly (good) results for a linear plus SE regression model in comparison with the drawing data from Chapter 4 (results not reported here). However, since the focus of the analysis in this chapter is slightly different—namely on how interactions of sound features affect gestural responses—I decided to proceed with a more traditional analytical approach using the General Linear Model as outlined below.

[53] The software specifies a frame rate of 20.83 Hz or 48 frames per second. However, due to the heavy computational load (getting data from the sensors, various mathematical operations, drawing on the screen and writing data to disk) the actual sample rate is reduced and inevitably varied from frame to frame. Mean frame length was 66.24 ms (SD = 4.95 ms) and median 68 ms. These statistics were extracted post-hoc from the recorded data.

As an indicator of the degree of the association between sound and movement features Spearman's rank correlation coefficient ρ—a non-parametric correlation coefficient—was calculated. This measure has been suggested for time-dependent data by Schubert (2002), and has been used by various scholars for similar datasets (e.g., Küssner & Leech-Wilkinson, 2014; Nymoen et al., 2013; Vines et al., 2006). It has been argued that one needs to be cautious when interpreting the size of correlation coefficients derived from time-dependent data. For Spearman's ρ, this is even more straightforward: regardless of time-dependence, the absolute size of this coefficient is never interpretable because its variance is not defined. Though the significance of a single Spearman's ρ derived from a time-dependent dataset might not be meaningful, it can be valuable to compare several correlation coefficients (see Chapter 2).

For the purpose of this analysis, global and local Spearman's rank correlation coefficients ρ were computed. The number of data points for a local correlation was N = 119, and for a global correlation N = 2142. Only sound stimuli Nos 4–21 were entered into the analysis (unless stated otherwise) since stimuli Nos 1–3 contain constant features that cannot be entered into a correlation analysis.[54] 'Global' denotes the correlation between sound features of all stimuli of a single participant and their accompanying hand movements (e.g., global frequency–Y correlation coefficient of participant *k*). 'Local' denotes the correlation between sound features of a particular stimulus of a single participant and their accompanying hand movements (e.g., local frequency–Y correlation coefficient of sound stimulus *s* of participant *k*). These correlation coefficients were then used as data in subsequent statistical analyses.

The following analytical steps were applied to investigate gestural representations of elapsed time, pitch and loudness and carried out in IBM SPSS Statistics (Version 20). First, the absolute global correlation coefficients between elapsed time, frequency and loudness and, respectively, the three spatial axes X, Y and Z, were entered into three ANOVAs with the within-subjects factor 'space' (X / Y / Z) to identify the three strongest correlations between each variable and each of the three axes (e.g., elapsed time and X), which was then examined further in the subsequent steps of the analysis. Secondly, the original (rather than absolute) global correlation coefficients were examined to identify the direction of movement (see Hypotheses 1–3) and to illuminate the role of visual feedback. Thirdly, the effects of interactions between musical

[54] Note that the loudness was only genuinely equal in stimulus no. 1. Due to the equal-loudness-level contour of pure tones (Y. Suzuki & Takeshima, 2004) and the use of loudness measured in sone, stimuli Nos 4–6 and Nos 13–15, whose amplitude was constant, could be entered into the analysis because their perceived loudness varied marginally according to the pitch contour.

parameters (pitch contour, loudness contour and tempo) on the size of the correlations were investigated by means of local correlation coefficients, resulting in ANOVAs with the between-subjects factors 'training' (musically trained / musically untrained; see Hypothesis 5) and 'sex' (male / female),[55] and the within-subjects factors 'vision' (without visualization / with visualization), 'pitch' (rising-falling / falling-rising; see Hypothesis 2), 'loudness' (equal amplitude / decreasing-increasing / increasing-decreasing) and 'tempo' (equal / decelerando-decelerando / accelerando-accelerando). All post-hoc pairwise comparisons were Sidak-corrected. Fourthly, to investigate whether muscular energy of the hand was associated with loudness and tempo variations in the stimuli (see Hypothesis 3), data from the Wii™ Remote Controller were collected when the difference in acceleration between the current and previous frame exceeded 10 m/s^2. That is, when participants shook the Controller strongly enough (henceforth 'shaking event'), the software recorded a shaking event with a timestamp. The absolute number of shaking events was inserted into bins of 200 ms and plotted separately for each stimulus across all participants, musically trained participants, musically untrained participants, the visual condition and the non-visual condition, respectively (see Appendix 5.1–5.5). To test whether the distribution across two quarters of the sound stimuli was significantly different from an equal distribution, multiple binomial tests were run (*p* values Bonferroni-corrected). The length of each quarter was two seconds. The beginning of the third quarter coincided with the start of a pitch ascent/descent in stimuli changing in pitch, and with the start of a decelerando/accelerando in stimuli changing in tempo. Fifthly, to investigate whether the speed of the hand movement was associated with tempo variations in the stimuli (see Hypothesis 4), the mean velocity in response to each quarter of a sound stimulus was the dependent variable of an ANOVA with the within-subjects factors 'half' (1^{st} / 2^{nd} half of a stimulus), 'quarter' (1^{st} / 2^{nd} quarter of each half), 'vision' (without / with visualization), 'pitch' (up / down), 'loudness' (equal / decreasing / increasing) and 'tempo' (equal / decelerando / accelerando), and the between-subjects factors 'training' and 'sex'. Also, median velocity[56] (including lower and upper quartile) was plotted separately for each stimulus across all participants, musically trained participants, musically untrained participants, the visual condition and the non-visual condition, respectively (see

[55] Exploring potential sex differences of gestural responses to sound and music was an idea developed based on feedback from a reviewer of the drawing experiment (Küssner & Leech-Wilkinson, 2014).

[56] Per participant, each pair of consecutive data points defines a segment of tracked motion, in which the average velocity is calculated as the Euclidean distance between the two hand positions, divided by the duration of the segment. Statistics across participants were calculated by locating the appropriate segments (and hence the velocities) for a fixed grid of sampling points along the *t* axis.

Appendix 5.1–5.5). Whenever the assumption of sphericity was violated in repeated-measures ANOVAs, the degrees of freedom were adjusted using the Greenhouse-Geisser correction.

5.3 Results

5.3.1 Absolute global correlation analysis

5.3.1.1 Elapsed time

There was a significant main effect of 'space' ($F(1.33, 82.63) = 57.13$, $p < .001$, partial $\eta^2 = .48$), and all three pairwise comparisons revealed significant differences (see Figure 5-4A). Correlations with X ($M = .49$, SEM = .04) were greater than with Y ($M = .13$, SEM = .01) and with Z ($M = .20$, SEM = .02; both $p < .001$), and correlations with Z were greater than with Y ($p = .001$).

5.3.1.2 Pitch

There was a main effect of 'space' ($F(1.82, 112.97) = 243.76$, $p < .001$, partial $\eta^2 = .80$), and all three Sidak-corrected pairwise comparisons revealed significant differences (all $p < .001$; see Figure 5-4B). Correlations with Y ($M = .65$, SEM = .02) were greater than with X ($M = .13$, SEM = .02) and with Z ($M = .24$, SEM = .02), and correlations with Z were greater than with X.

5.3.1.3 Loudness

There was a main effect of 'space' ($F(2, 124) = 128.03$, $p < .001$, partial $\eta^2 = .67$), and all three pairwise comparisons revealed significant differences (see Figure 5-4C). Correlations with Y ($M = .31$, SEM = .01) were greater than with X ($M = .10$, SEM = .01) and with Z ($M = .15$, SEM = .01; both $p < .001$), and correlations with Z were greater than with X ($p = .001$).

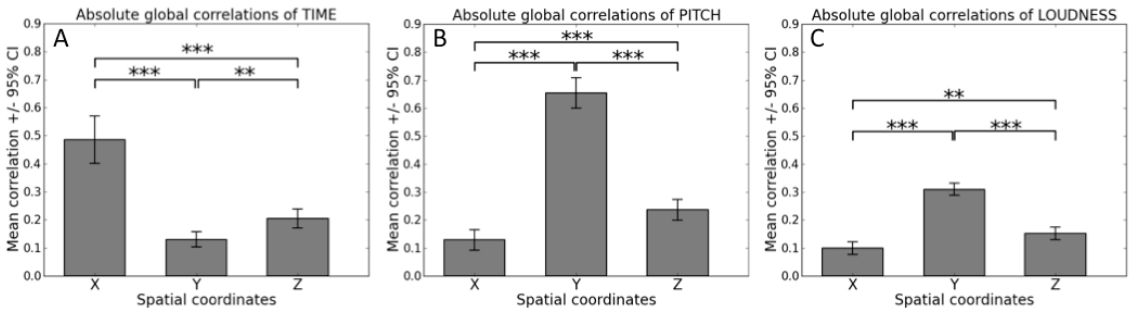


Figure 5-4 Absolute global correlations of elapsed time (A), pitch (B) and loudness (C) with all three spatial axes. ** indicates $p < .01$, *** indicates $p < .001$

5.3.2 Gestural representation of elapsed time

5.3.2.1 Global correlations: spatial direction of elapsed time

Having identified the x-axis as largest correlate of elapsed time, the next step is to examine the direction of movement. Table 5-2 shows the distribution of musically trained and untrained participants who showed (a) negative correlations in both conditions, suggesting that they moved from right to left as time elapsed regardless of visual feedback, (b) a negative correlation in the non-visual condition and a positive correlation in the visual condition, suggesting that they moved from right to left in the absence of visual feedback and from left to right with visual feedback, (c) a positive correlation in the non-visual condition and a negative correlation in the visual condition, suggesting that they moved from left to right in the absence of visual feedback and from right to left with visual feedback, and (d) positive correlations in both conditions, suggesting that they moved from left to right regardless of visual feedback.

For the musically trained, there were marginally significant, and for the musically untrained, highly significant associations between the sign of the correlation coefficient in the non-visual condition and the sign of the correlation coefficient in the visual condition (trained: $\chi^2(1) = 4.73$, $p = .063$; untrained: $\chi^2(1) = 14.11$, $p < .001$; both with Fisher's exact test). The odds of trained participants moving from left to right as time elapsed in the visual condition were 7.67 times higher if they moved from left to right (rather than right to left) in the non-visual condition. The odds of untrained participants moving from left to right as time elapsed in the visual condition were even 38 times higher if they moved from left to right (rather then right to left) in the non-visual condition. However, the sign of the correlation coefficient should always be interpreted together with the magnitude of the correlation; thus, Table 5-2 also shows the mean and standard deviation values of the correlation coefficients of both conditions. All six musically trained participants in the visual/right-to-left condition showed absolute correlation coefficients smaller than, or equal to, .10. In other words, those participants displayed a very weak association, if any, between elapsed time and movement towards the left or the right. The three trained participants showing negative values in the non-visual condition and positive values in the visual condition, showed a radical shift from slightly negative (non-visual mean $\rho = -.30$) to strongly positive (visual mean $\rho = .78$). The majority of musically trained participants showed a fairly strong association between elapsed time and movement towards the right regardless of condition (see statistical analysis below). With regard to the musically untrained participants,

eight moved from right to left with slight consistency in both conditions (note, however, the large standard deviations); four moved fairly consistently from left to right in the non-visual condition (note again the large standard deviations) and showed practically no association in the visual condition; and one musically untrained participant displayed a rather weak association between elapsed time and movement towards the left in the non-visual condition and no association at all in the visual condition. Similar to the trained group, the majority of untrained participants showed a fairly strong association between time and rightward movement regardless of condition.

Comparing the two majority groups of trained (n = 23) and untrained participants (n = 19), a 2 × 2 ANOVA with between-subjects factor 'training' and within-subjects factor 'vision' was run (dependent variable: correlations between elapsed time and x-axis). Although the data seem to suggest an interaction between 'training' and 'vision' such that musically trained participants' values decrease in the visual condition whereas those of untrained participants increase, this interaction was not significant, $F(1, 40) = 1.51$, $p > .20$. Main effects were not significant either.

Table 5-2 Number of participants classified by the sign of their global correlation coefficients between elapsed time and movement along the x-axis in the non-visual and visual condition

			non-visual	
			negative (right to left)	positive (left to right)
Musically trained participants	visual	negative (right to left)	**3** nv: –.07 (.04) v: –.04 (.03)	**3** nv: .10 (.05) v: –.09 (.06)
		positive (left to right)	**3** nv: –.30 (.09) v: .78 (.11)	**23** nv: .63 (.37) v: .60 (.39)
Musically untrained participants	visual	negative (right to left)	**8** nv: –.31 (.28) v: –.29 (.21)	**4** nv: .37 (.36) v: –.08 (.07)
		positive (left to right)	**1** *nv: –.17 *v: .01	**19** nv: .59 (.34) v: .63 (.30)

Note. Bold integers indicate number of participants per cell, followed by their mean correlation coefficients with standard deviation in brackets. Negative refers to negative associations between time and movement along the x-axis, suggesting a leftward movement. Positive refers to positive associations between time and movement along the x-axis, suggesting a rightward movement. nv/non-visual: non-visual condition, v/visual: visual condition, asterisk denotes single value instead of mean.

5.3.2.2 Local correlations: interactions between musical parameters

There was a main effect of 'tempo' ($F(1.62, 97.13) = 4.91$, $p = .014$, partial $\eta^2 = .08$) with one significant pairwise comparison ($p = .042$) revealing that 'accelerando-accelerando' ($M = .42$, SEM = .06) resulted in higher time–X correlation coefficients than 'decelerando-decelerando' ($M = .36$, SEM = .06). A significant interaction effect between 'loudness' and 'sex' ($F(2, 120) = 3.10$, $p = .049$, partial $\eta^2 = .05$) was observed. Contrasts revealed a significant interaction when comparing male and female time–X correlation coefficients to decreasing-increasing loudness (male: $M = .36$, SEM = .08; female: $M = .38$, SEM = .08) compared to constant amplitude (male: $M = .33$, SEM = .08; female: $M = .43$, SEM = .08), $F(1, 60) = 7.01$, $p = .010$, partial $\eta^2 = .11$, but not to increasing-decreasing loudness compared to constant amplitude ($F(1, 60) = 2.30$, $p > .10$), nor to decreasing-increasing loudness compared to increasing-decreasing loudness ($F(1, 60) < 1$). When discussing sex effects more generally I will attempt to provide an explanation of why this might be important.

5.3.3 Gestural representation of pitch

5.3.3.1 Global correlations: spatial direction of pitch

Investigating the direction in which participants moved their hands along the y-axis when presented with changes in pitch revealed a very clear result: All sixty-four participants showed positive correlation coefficients in both conditions, suggesting that they moved their hand upwards with increasing pitch and downwards with decreasing pitch. Primary response data—gestural trajectories along the y-axis in response to sound stimulus No. 4 (rising-falling pitch) in the non-visual condition—are shown for a subsample of sixteen randomly chosen musically trained participants (Figure 5-5 left) and sixteen randomly chosen musically untrained participants (Figure 5-5 right).

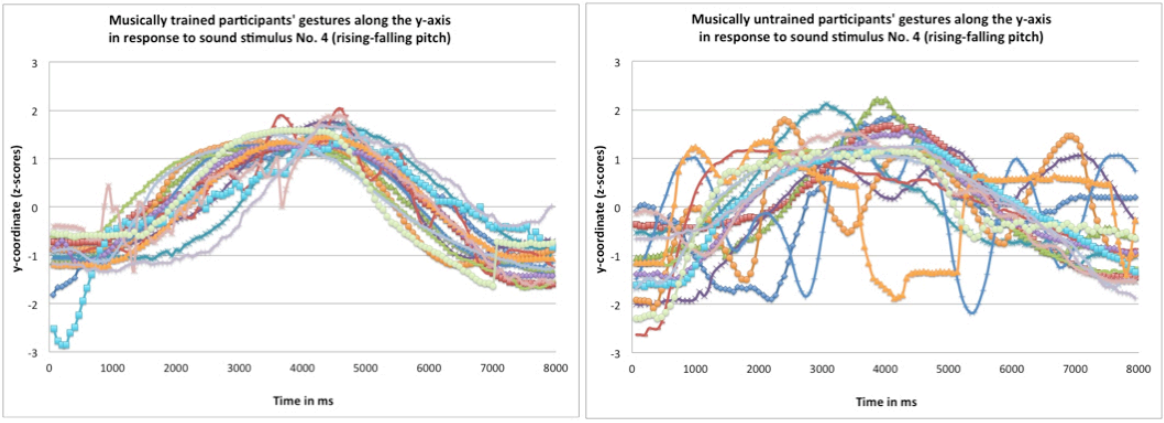


Figure 5-5 Gestural trajectories along the y-axis in response to sound stimulus rising and falling in pitch (No. 4) in the non-visual condition by a subsample of sixteen randomly chosen musically trained participants (left) and sixteen randomly chosen musically untrained participants (right).

5.3.3.2 Local correlations: interactions between musical parameters

Results revealed a significant main effect of 'vision' ($F(1, 60) = 10.28$, $p = .002$, partial $\eta^2 = .15$) and of 'training' ($F(1, 60) = 21.37$, $p < .001$, partial $\eta^2 = .26$). The positive association between pitch and height was larger in the non-visual ($M = .68$, SEM = .02) compared to the visual condition ($M = .63$, SEM = .03), and larger for musically trained ($M = .77$, SEM = .03) compared to untrained participants ($M = .55$, SEM = .03). There was a main effect of 'pitch' ($F(1, 60) = 12.15$, $p = .001$, partial $\eta^2 = .17$), showing that rising-falling pitch contours ($M = .69$, SEM = .02) gave rise to higher frequency–Y correlation coefficients than falling-rising pitch contours ($M = .62$, SEM = .03).

There was also a main effect of 'loudness' ($F(2, 120) = 6.51$, $p = .002$, partial $\eta^2 = .10$), revealing that constant amplitude ($M = .68$, SEM = .02) gave rise to higher frequency–Y correlation coefficients than both decreasing-increasing ($M = .65$, SEM = .03, $p = .042$) and increasing-decreasing ($M = .64$, SEM = .03, $p = .001$) loudness contours. And there was a main effect of 'tempo' ($F(2, 120) = 7.24$, $p = .001$, partial $\eta^2 = .11$), revealing that equal tempo ($M = .69$, SEM = .03) compared to 'accelerando-accelerando' ($M = .62$, SEM = .03)—but not to 'decelerando-decelerando' ($M = .66$, SEM = .02)—resulted in higher frequency–Y correlation coefficients ($p = .004$).

Moreover, several two- and three-way interactions were observed. There was a significant interaction effect between 'pitch' and 'training' ($F(1, 60) = 6.86$, $p = .011$, partial $\eta^2 = .10$),

revealing that the observed main effect of 'pitch' is chiefly due to musically untrained participants' lower frequency–Y correlation coefficients when presented with falling-rising pitch contours (*M* = .49, SEM = .04) compared to rising-falling pitch contours (*M* = .61, SEM = .03), *t*(31) = 3.49, *p* = .001, *r* = .53. In comparison, musically trained participants' frequency–Y correlation coefficients did not differ significantly (rising-falling pitch contours: *M* = .77, SEM = .03; falling-rising pitch contours: *M* = .76, SEM = .04), *t*(31) = .86, *p* > .30, *r* = .15, as shown in Figure 5-6A below.

Furthermore, there was a significant interaction between 'tempo' and 'training' (*F*(2, 120) = 14.58, *p* < .001, partial η^2 = .20). Contrasts revealed significant interactions when comparing musically trained and untrained participants' frequency–Y correlation coefficients to equal tempo compared to 'accelerando-accelerando' (*F*(1, 60) = 10.54, *p* = .002, partial η^2 = .15), to 'decelerando-decelerando' compared to 'accelerando-accelerando' (*F*(1, 60) = 28.53, *p* < .001, partial η^2 = .32), but not to equal tempo compared to 'decelerando-decelerando' (*F*(1, 60) = 3.51, *p* = .066). Inspecting the interaction graph (see Figure 5-6B below), this suggests that musically untrained participants' frequency–Y correlation coefficients decrease when accelerando-accelerando tempo profiles (compared to both equal tempo and 'decelerando-decelerando') are present, whereas musically trained participants' frequency–Y correlation coefficients decrease when decelerando-decelerando tempo profiles (compared to both equal tempo and 'accelerando-accelerando') are present.

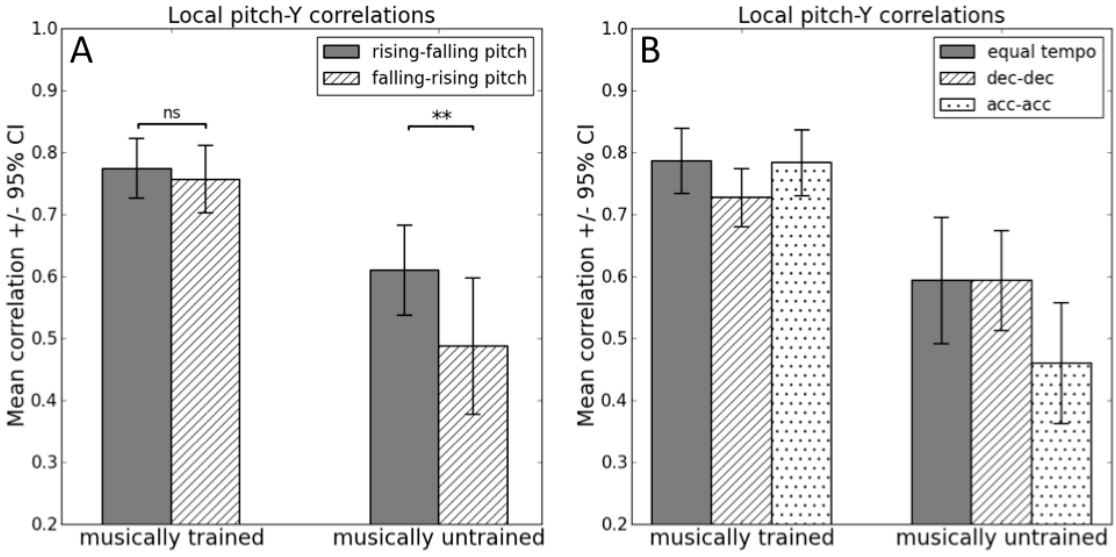


Figure 5-6 Influence of interactions between musical training and pitch contour (A), and between musical training and tempo profiles (B) on local pitch–Y correlations. ** indicates *p* < .01, ns: not significant

There were also significant interaction effects between 'pitch' and 'tempo' (F(1.71, 102.32) = 14.63, p < .001, partial η^2 = .20), between 'pitch', 'tempo' and 'training' (F(1.71, 102.32) = 3.71, p = .034, partial η^2 = .06), and between 'pitch', 'tempo' and 'vision' (F(1.73, 103.96) = 3.95, p = .027, partial η^2 = .06). Since both three-way interactions include 'pitch' and 'tempo', only those—but not the two-way interaction—will be examined more closely here.

The first set of contrasts compared musically trained and untrained participants at each level of 'pitch' across each level of 'tempo'. No significant differences were found between musically trained and untrained participants' frequency–Y correlation coefficients when comparing rising-falling pitch contours to falling-rising pitch contours when the tempo was equal compared to a 'decelerando-decelerando' pattern (F(1, 60) = 2.38, p > .10), nor between musically trained and untrained participants' frequency–Y correlation coefficients when comparing rising-falling pitch contours to falling-rising pitch contours when the tempo was equal compared to an 'accelerando-accelerando' pattern (F(1, 60) = 1.90, p > .10). There was however a significant difference between musically trained and untrained participants' frequency–Y correlation coefficients when comparing rising-falling pitch contours to falling-rising pitch contours when the tempo pattern was 'decelerando-decelerando' compared to 'accelerando-accelerando' (F(1, 60) = 6.26, p = .015, partial η^2 = .09), see Figure 5-7 below.

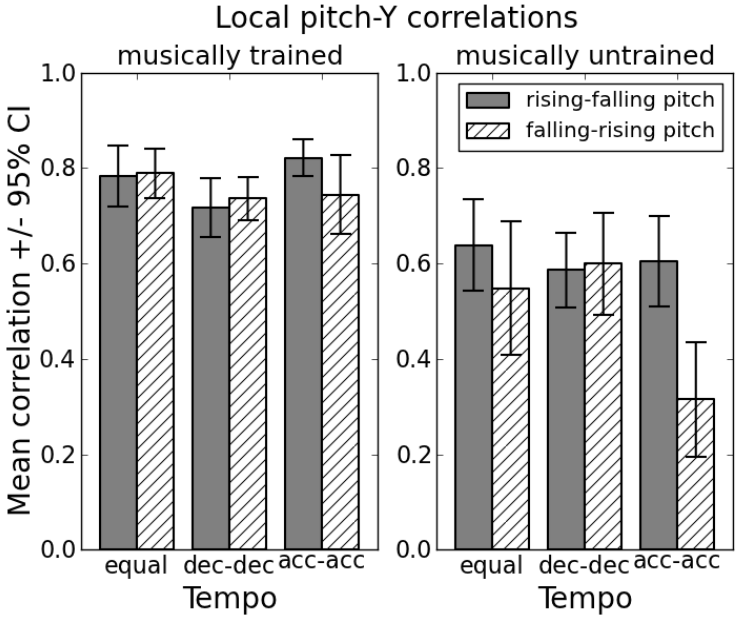


Figure 5-7 Influence of interaction between musical training, pitch contour and tempo profile on local pitch–Y correlations

The second set of contrasts compared the conditions 'without visualization' and 'with visualization' at each level of 'pitch' across each level of 'tempo'. No significant differences were found between 'without visualization' and 'with visualization' when comparing rising-falling pitch contours to falling-rising pitch contours when the tempo was equal compared to a 'decelerando-decelerando' pattern ($F(1, 60) = 1.45$, $p > .20$), nor between 'without visualization' and 'with visualization' when comparing rising-falling pitch contours to falling-rising pitch contours when the tempo pattern was 'decelerando-decelerando' compared to 'accelerando-accelerando' ($F(1, 60) = 2.90$, $p = .094$). However, there was a significant difference between 'without visualization' and 'with visualization' when comparing rising-falling pitch contours to falling-rising pitch contours when the tempo was equal compared to an 'accelerando-accelerando' pattern ($F(1, 60) = 6.14$, $p = .016$, partial $\eta^2 = .09$), see Figure 5-8 below. I shall return to the role of the visualization in shaping participants' responses in the Discussion.

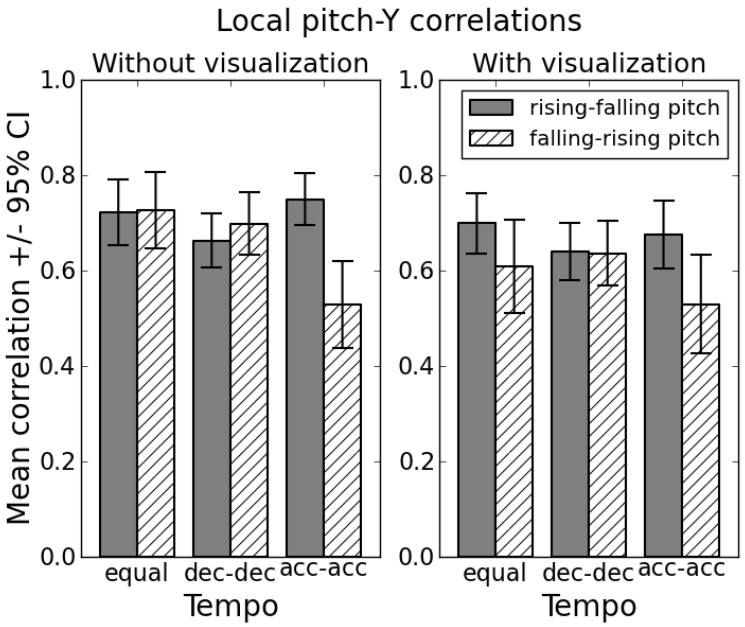


Figure 5-8 Influence of interaction between visualization, pitch contour and tempo profile on local pitch–Y correlations

5.3.4 Gestural representation of loudness

5.3.4.1 Global correlations: spatial direction of loudness–Y associations

Apart from two musically untrained participants, all participants showed positive correlations between loudness and height regardless of condition, suggesting that they moved their arm upwards with increasing loudness and downwards with decreasing loudness. The question

arises, however, whether participants indeed chose to represent loudness with the y-axis, or whether this is a spurious effect, caused by interactions between pitch and loudness in the stimuli. Recall that stimuli Nos 10–12 and Nos 16–18 consist, respectively, of concurrently increasing-decreasing and decreasing-increasing pitch and loudness contours, whereas stimuli Nos 7–9 and Nos 19–21 consist of opposing pitch and loudness contours. Thus, it is vital to consider the local correlations to identify whether the positive loudness–Y correlations values are in fact a side effect of frequency–Y correlation coefficients. If so, there should be a significant interaction effect between 'pitch' and 'loudness', resulting in negative loudness–Y correlations for stimuli when the pitch contour is rising-falling (falling-rising) and the loudness contour is concurrently decreasing-increasing (increasing-decreasing).

5.3.4.2 Interactions between musical parameters – local correlations of loudness–Y

Although there are significant main effects of 'training', 'vision', 'loudness' and 'tempo', as well as significant interaction effects between 'pitch' and 'tempo', and 'loudness' and 'tempo', the main focus here is on a highly significant interaction effect between 'pitch' and 'loudness' ($F(1.27, 75.94) = 664.39$, $p < .001$, partial $\eta^2 = .92$). Contrasts revealed significant interactions when comparing loudness–Y correlation coefficients of rising-falling and falling-rising pitch contours to equal amplitude compared to decreasing-increasing loudness patterns ($F(1, 60) = 734.74$, $p < .001$, partial $\eta^2 = .93$), to equal amplitude compared to increasing-decreasing loudness patterns ($F(1, 60) = 382.95$, $p < .001$, partial $\eta^2 = .87$), and to decreasing-increasing loudness patterns compared to increasing-decreasing loudness patterns ($F(1, 60) = 753.49$, $p < .001$, partial $\eta^2 = .93$). Inspecting the interaction graph (see Figure 5-9 below), it becomes obvious that participants map pitch, not loudness, onto the y-axis.[57] When rising-falling pitch contours are paired with decreasing-increasing loudness contours the loudness–Y correlation coefficients are negative ($M = –.46$, SEM = .02), and when paired with increasing-decreasing loudness contours they are positive ($M = .69$, SEM = .02). Similarly, when falling-rising pitch contours are paired with increasing-decreasing loudness contours the loudness–Y correlation coefficients are negative ($M = –.35$, SEM = .03), and when paired with decreasing-increasing loudness contours they are positive ($M = .61$, SEM = .04). Also the slight decrease of loudness–Y correlation coefficients from rising-falling ($M = .73$, SEM = .02) to falling-rising pitch contours

[57] Of course, one might interject that by looking at the gesture data only it cannot be ruled out that participants were in fact tracing the loudness (moving the arm upwards with decreasing loudness), rendering the pitch–height associations a spurious effect! However, not only would the direction of this loudness–height association be very unusual but post-experiment interviews clearly revealed that this demur is unfounded.

(*M* = .64, SEM = .03) when the amplitude is equal fits into the picture, as it reflects the main effect of 'pitch' for frequency–Y correlation coefficients.

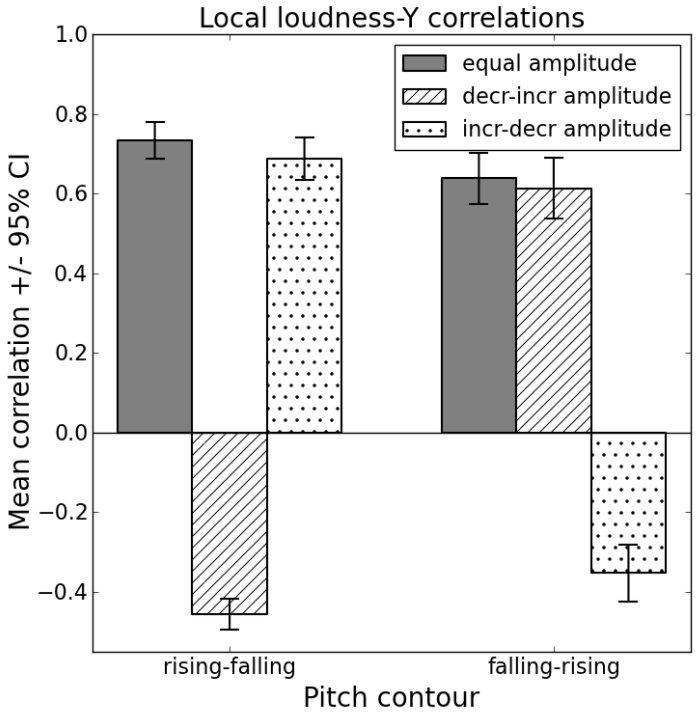


Figure 5-9 Spurious loudness–height association: influence of interaction between pitch and loudness contour on local loudness–Y correlations

The interaction between 'pitch' and 'loudness' was further qualified by three significant three-way interaction effects between these two within-subjects factors and 'musician', 'sex' and 'tempo', respectively. Due to the clear results obtained from the interaction between 'pitch' and 'loudness', the three-way interactions are, however, not deemed worth pursuing further, and the focus is shifted to stimuli without change in pitch to investigate whether there exist associations between loudness and height when loudness is the only auditory feature being manipulated.

Running a repeated-measures ANOVA on loudness–Y correlation coefficients of stimuli Nos 2 and 3 with the within-subjects factors 'vision' and 'loudness', and with the between-subjects factors 'training' and 'sex', significant main effects of 'loudness' ($F(1, 60) = 12.00$, $p = .001$, partial $\eta^2 = .17$) and 'training' ($F(1, 60) = 5.54$, $p = .022$, partial $\eta^2 = .09$) were observed. The increasing-decreasing loudness contour (*M* = .36, SEM = .04) gave rise to higher loudness–Y correlation coefficients compared to the decreasing-increasing loudness contour (*M* = .13, SEM

= .05), and musically trained participants (*M* = .32, SEM = .05) showed higher loudness–Y correlation coefficients than untrained participants (*M* = .17, SEM = .05). There was also a significant interaction effect between 'training' and 'sex' ($F(1, 60) = 5.16$, $p = .027$, partial $\eta^2 = .08$), revealing that female musically untrained participants' loudness–Y correlation coefficients are lower compared to those of male musically untrained participants, while there is no sex difference between musically trained participants (see Figure 5-10 below).

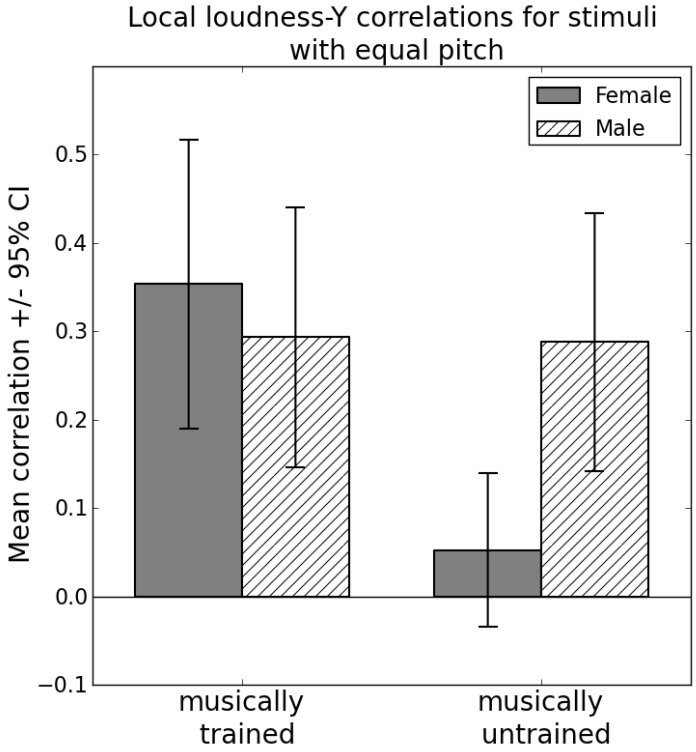


Figure 5-10 Influence of interaction between sex and musical training on local loudness–Y correlations for stimuli with equal pitch

5.3.4.3 Local correlations of loudness–Z for stimuli without change in pitch

Since the association between loudness and the z-axis for stimuli concurrently varied in pitch, loudness and tempo was too small to be interpreted meaningfully (mean absolute $\rho = .15$; see Absolute global correlation analysis), the focus is shifted to stimuli without change in pitch to investigate whether there was any association between loudness and distance/size when loudness was the only auditory feature being manipulated. Results from the ANOVA revealed two significant main effects: 'training' ($F(1, 60) = 7.38$, $p = .009$, partial $\eta^2 = .11$) and 'loudness' ($F(1, 60) = 24.83$, $p < .001$, partial $\eta^2 = .29$). Loudness–Z correlation coefficients were significantly higher for musically trained (*M* = .30, SEM = .06) compared to untrained

participants (*M* = .08, SEM = .06), and significantly higher when the loudness was increasing-decreasing (*M* = .39, SEM = .05) compared to decreasing-increasing (*M* = –.01, SEM = .06). This suggests that only musically trained participants associated loudness with the z-axis, and only if the loudness contour was increasing-decreasing.

5.3.5 Association between muscular energy (shaking events) and loudness

The question then arises whether participants did represent loudness at all when musical parameters were varied concurrently, since we have just seen that participants used neither height (representation of pitch takes precedence) nor distance/size (correlation coefficients too low). According to my hypotheses I expected an association between loudness and muscular energy (operationalized as shaking hand movements). Table 5-3 below shows an overview of the number of shaking events across the combined 1st + 3rd and 2nd + 4th quarters of all sound stimuli varied in loudness. It is evident that decreasing and increasing loudness contours show a very clear association with the number of shaking events, that is, when the loudness is decreasing across two quarters the number of shaking events decreases, and when the loudness is increasing the number of shaking events increases.

Table 5-3 Number of shaking events (muscular energy) per combined quarters

Loudness:	**decreasing**		**increasing**	
	1st+3rd quarter	2nd+4th quarter	1st+3rd quarter	2nd+4th quarter
all	2425	1867	1363	2517
non-visual	1001	567	500	971
visual	1424	1300	863	1546
untrained	1818	1484	1003	1916
trained	607	383	359	596

Note. Displayed are the added numbers of shaking events (fast hand movements with the Wii™ Remote Controller) for the first halves of each half (1st+3rd quarter) and for the second halves of each half (2nd+4th quarter) of sound stimuli varied in loudness. Numbers are displayed separately for the non-visual and visual condition, and musically untrained and trained participants, as well as overall such that n(visual) + n(non-visual) = n(all), and n(untrained) + n(trained) = n(all).

Table 5-4 below shows the results of testing the distribution of the number of shaking events across the first two and second two quarters with multiple binomial tests. Results are discussed in the Discussion section below.

Table 5-4 Association between muscular energy (shaking events) and loudness: multiple binomial tests for stimuli with equal, decreasing-increasing and increasing-decreasing loudness contours

Stimulus (ampl.)		No. of events in 1st half			No. of events in 2nd half		
		1st quarter	2nd quarter	*p* value	3rd quarter	4th quarter	*p* value
s1 (equal)	all*	285	375	0.000522	417	437	**0.515608**
	non-visual**	133	130	**0.901880**	153	148	**0.817707**
	visual*	152	245	0.000004	264	289	**0.307449**
	untrained*	237	328	0.000148	370	367	**0.941279**
	trained*	48	47	**0.918778**	47	70	0.041501
s2 (decr-incr)	all*	273	238	0.132483	231	401	**0.000000**
	non-visual*	121	77	0.002164	88	149	**0.000090**
	visual*	152	161	0.651204	143	252	**0.000000**
	untrained*	181	189	0.715977	148	262	**0.000000**
	trained**	92	49	**0.000368**	83	139	**0.000208**
s7 (decr-incr)	all	40	76	0.001065	36	38	0.907561
	non-visual	19	39	0.011928	8	18	0.075519
	visual	21	37	0.047940	28	20	0.312327
	untrained*[1]	24	76	**0.000000**	32	35	0.807195
	trained*	16	0	**0.000031**	4	3	0.726563
s8 (decr-incr)	all	102	116	0.378646	87	110	0.116787
	non-visual	59	53	0.636800	29	50	0.023820
	visual	43	63	0.064464	58	60	0.926704
	untrained*	84	109	0.083802	60	106	**0.000443**
	trained*[1]	18	7	0.043285	27	4	**0.000034**
s9 (decr-incr)	all	150	170	0.288160	188	135	0.003745
	non-visual	76	60	0.198181	49	59	0.386573
	visual*[1]	74	110	0.009684	139	76	**0.000021**
	untrained	87	128	0.006240	156	101	0.000725
	trained	63	42	0.050442	32	34	0.902159
s16 (decr-incr)	all*	58	55	0.850870	115	226	**0.000000**
	non-visual	43	34	0.362032	71	78	0.623194
	visual*	15	21	0.405032	44	148	**0.000000**
	untrained*	35	44	0.368188	84	159	**0.000002**
	trained*	23	11	0.057613	31	67	**0.000355**
s17 (decr-incr)	all*	44	49	0.678532	129	248	**0.000000**
	non-visual	20	17	0.742829	51	64	0.263054
	visual*	24	32	0.349682	78	184	**0.000000**
	untrained*[1]*	5	28	**0.000066**	82	194	**0.000000**
	trained	39	21	0.027340	47	54	0.550709
s18 (decr-incr)	all*	182	140	0.022175	99	254	**0.000000**
	non-visual**	77	30	**0.000006**	16	86	**0.000000**
	visual*	105	110	0.785084	83	168	**0.000000**
	untrained*	157	114	0.010598	63	177	**0.000000**
	trained*	25	26	1.000000	36	72	**0.000684**

s3 (incr-decr)	all**	155	298	**0.000000**	368	176	**0.000000**
	non-visual**	56	117	**0.000004**	131	47	**0.000000**
	visual**	99	181	**0.000001**	237	129	**0.000000**
	untrained**	120	207	**0.000002**	255	125	**0.000000**
	trained**	35	91	**0.000001**	113	51	**0.000001**
s10 (incr-decr)	all**	22	123	**0.000000**	108	21	**0.000000**
	non-visual**	8	65	**0.000000**	39	4	**0.000000**
	visual**	14	58	**0.000000**	69	17	**0.000000**
	untrained**	14	106	**0.000000**	102	16	**0.000000**
	trained	7	17	0.063915	6	5	0.774414
s11 (incr-decr)	all**	64	244	**0.000000**	201	115	**0.000002**
	non-visual**	33	103	**0.000000**	59	18	**0.000003**
	visual*	31	141	**0.000000**	142	97	0.004327
	untrained**	58	203	**0.000000**	163	67	**0.000000**
	trained*	6	41	**0.000000**	38	48	0.331834
s12 (incr-decr)	all	67	110	0.001522	290	219	0.001889
	non-visual*	34	46	0.218518	100	45	**0.000006**
	visual	33	64	0.002152	190	174	0.431783
	untrained*	34	72	**0.000285**	234	184	0.016442
	trained	33	38	0.635308	56	35	0.035450
s19 (incr-decr)	all	37	50	0.197979	188	131	0.001674
	non-visual*	18	26	0.291215	91	29	**0.000000**
	visual	19	24	0.542384	97	102	0.776838
	untrained	28	49	0.022033	165	124	0.018471
	trained	9	1	0.021484	23	7	0.005223
s20 (incr-decr)	all*	38	132	**0.000000**	250	207	0.049333
	non-visual*	26	84	**0.000000**	112	85	0.063691
	visual*	12	48	**0.000003**	138	122	0.352258
	untrained*	33	100	**0.000000**	205	156	0.011423
	trained*	5	32	**0.000007**	45	51	0.610068
s21 (incr-decr)	all	95	148	0.000815	171	154	0.374825
	non-visual	13	26	0.053252	54	29	0.008037
	visual	82	122	0.006184	117	125	0.652817
	untrained*	91	145	**0.000533**	121	124	0.898361
	trained	4	3	0.726563	50	30	0.032993

Note. Displayed are the numbers of shaking events for all four quarters of a stimulus (0-2 sec, 2-4 sec, 4-6 sec and 6-8 sec). Multiple binomial tests were run, separately for each half, to determine whether equal, decreasing and increasing loudness was associated with respectively equal, right-skewed ("decreasing") and left-skewed ("increasing") distribution of shaking events across two quarters. For the only stimulus with equal loudness, the null hypothesis was that their distribution is unequal, hence *p*-values exceeding the significance threshold of α = .20 were regarded as significant. For stimuli with decreasing and increasing loudness, the threshold was Bonferroni-corrected and testing was carried out one-sided: 0.10 / 140 = 0.000714. The number of asterisks indicates whether one half (*) or both halves (**) revealed a significant result (*p*-values in bold). [1] indicates that the distribution underlying the significant result is in opposite direction, i.e. left-skewed when right-skewed was hypothesized, or vice versa. Stimuli with both constant amplitude and changes in pitch were omitted because participants might have perceived changes in loudness due to the equal-loudness-level contour.

5.3.6 Association between muscular energy (shaking events) and tempo

Muscular energy was also hypothesized to be associated with tempo but as the overview in Table 5-5 below shows, there was no clear association between muscular energy and tempo.

While for stimuli decreasing in tempo there is—contrary to my hypothesis—a consistent increase in the number of shaking events across two quarters, there is no consistent pattern for stimuli increasing in tempo: overall, in the visual condition, and for musically untrained participants, the number is increasing, and in the non-visual condition and for trained participants, the number is decreasing. Thus, my hypothesis pertaining to muscular energy and tempo is rejected and the focus is shifted to other associations of tempo such as speed of hand movement.

Table 5-5 Number of shaking events (muscular energy) per combined quarters

Tempo:	**decelerando**		**accelerando**	
	1st+3rd quarter	2nd+4th quarter	1st+3rd quarter	2nd+4th quarter
all	1364	1786	2045	2188
non-visual	627	747	742	683
visual	737	1039	1303	1505
untrained	1052	1411	1583	1770
trained	312	375	458	413

Note. Displayed are the added numbers of shaking events (fast hand movements with the Wii™ Remote Controller) for the first halves of each half (1st+3rd quarter) and for the second halves of each half (2nd+4th quarter) of sound stimuli varied in tempo. Numbers are displayed separately for the non-visual and visual condition, and musically untrained and trained participants, as well as overall such that n(visual) + n(non-visual) = n(all), and n(untrained) + n(trained) = n(all).

5.3.7 How pitch, loudness, tempo and interactions thereof influence the speed of hand movement when representing sound gesturally

Investigating the speed of hand movement, only interaction effects of the ANOVA involving at least the factors ‘quarter’ and either ‘pitch’, ‘loudness’ or ‘tempo’ will be reported here since the aim is to analyse how changes of speed across either half of a sound stimulus are affected by changes in pitch, loudness and tempo. There were significant interaction effects between ‘quarter’ and ‘loudness’ ($F(1.29, 77.30) = 7.31$, $p = .001$, partial $\eta^2 = .11$), ‘quarter’ and ‘tempo’ ($F(1.68, 100.48) = 78.51$, $p < .001$, partial $\eta^2 = .57$), but not between ‘quarter’ and ‘pitch’ ($F(1, 60) = 1.16$, $p > .20$). When loudness decreased across two quarters, speed of hand movement was significantly reduced ($t(63) = –7.98$, $p < .001$, $r = .71$), and when loudness increased across two quarters, speed of hand movement was significantly reduced as well ($t(63) = –2.98$, $p = .004$, $r = .35$), as shown in Figure 5-11A below. On the other hand, when tempo decreased across two quarters, speed of hand movement was significantly reduced ($t(63) = –10.31$, $p <$

.001, r = .79), and when tempo increased across two quarters, speed of hand movement was significantly increased (t(63) = 2.59, p = .012, r = .31), as shown in Figure 5-11B below.

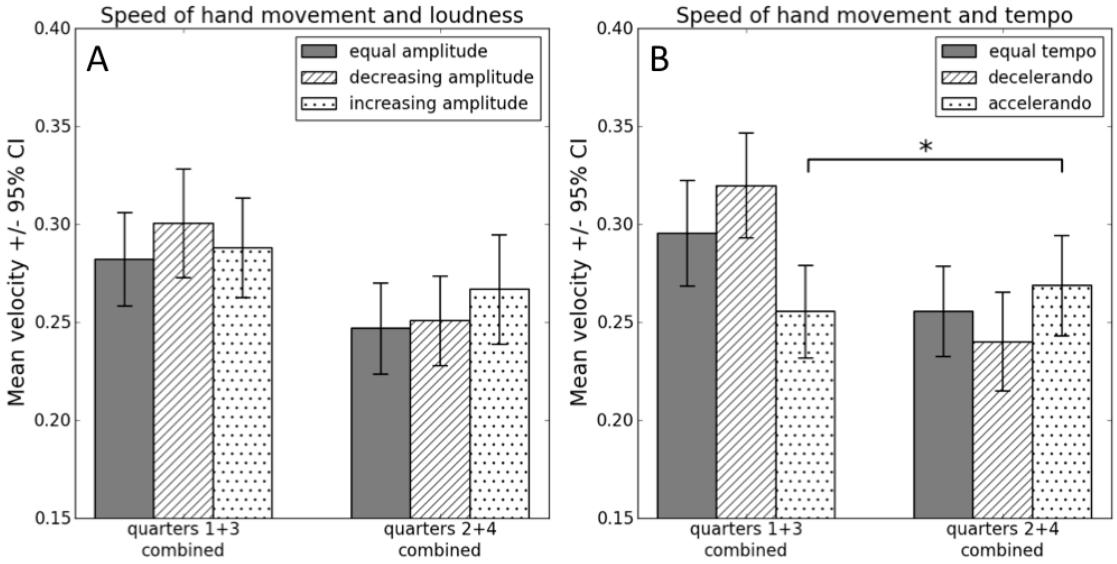


Figure 5-11 Influence of (A) loudness contour and (B) tempo profile on speed of hand movement. * indicates p < .05

Significant three- and four-way interaction effects further qualified the interaction effect between 'quarter' and 'tempo'. There was a significant interaction effect between 'quarter', 'tempo' and 'training' (F(1.68, 100.48) = 15.94, p < .001, partial η^2 = .21), revealing that the moderate effect size of the association between accelerando and increase in speed is due to musically untrained participants' lack of increase in speed. While musically trained participants' increase in speed across two quarters is highly significant when tempo is accelerating (t(31) = 3.98, p < .001, r = .58), there is no difference for untrained participants (t(31) = –.73, p > .40, r = .13), as shown in Figure 5-12 below.

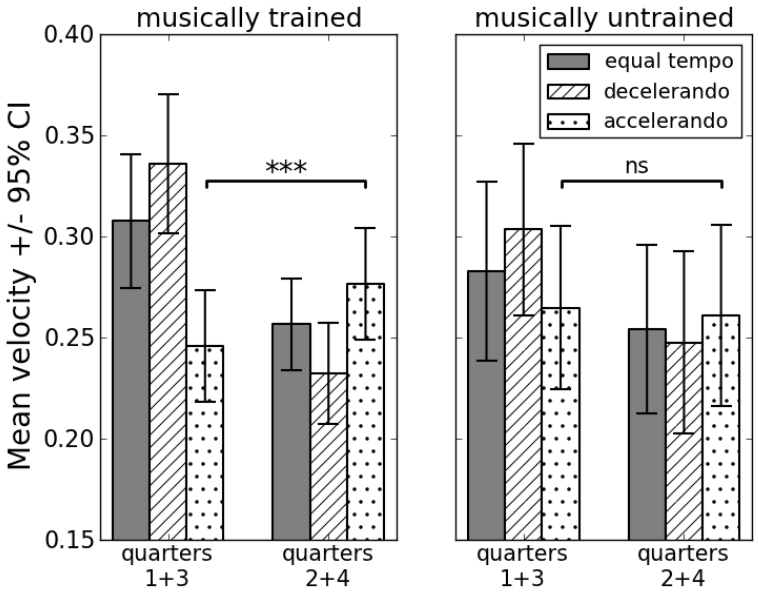


Figure 5-12 Influence of interaction between musical training and tempo profile on speed of hand movement. *** indicates $p < .001$, ns: not significant

There was also a significant interaction effect between 'quarter', 'tempo' and 'pitch' ($F(2, 120) = 3.95$, $p = .022$, partial $\eta^2 = .06$), revealing that the speed of hand movement associated with accelerando—but not decelerando—is affected by the pitch contour. When pitch was rising and tempo accelerating across two quarters, there was a significant increase in speed of the hand movement ($t(63) = 2.47$, $p = .016$, $r = .30$), but when pitch was falling and tempo accelerating, the increase in speed was not significant ($t(63) = 1.10$, $p > .20$, $r = .14$), as shown in Figure 5-13 below.

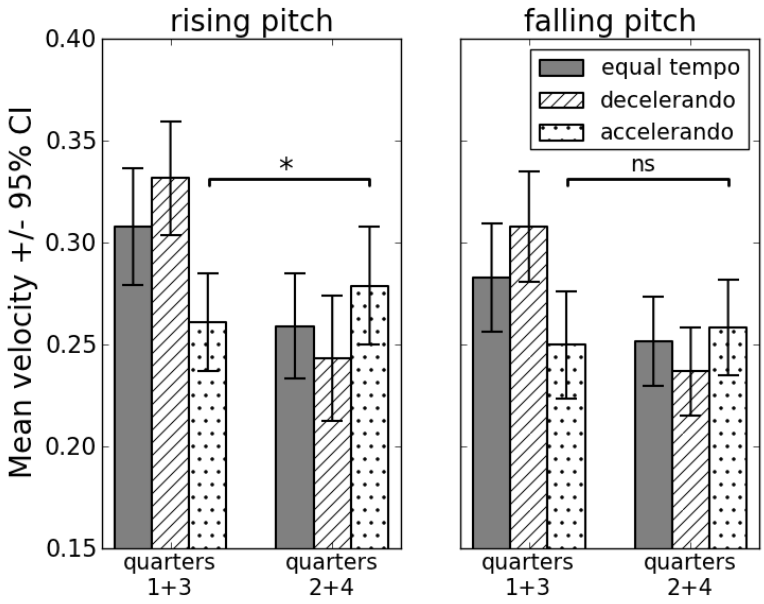


Figure 5-13 Influence of interaction between pitch contour and tempo profile on speed of hand movement. * indicates $p < .05$, ns: not significant

Moreover, there was a significant four-way interaction effect between 'half', 'quarter', 'loudness' and 'tempo' ($F(4, 240) = 4.69$, $p = .001$, partial $\eta^2 = .07$) which was broken down by running one ANOVA for each half.

For the first half, there was a significant interaction effect between 'quarter', 'loudness' and 'tempo' ($F(4, 240) = 3.90$, $p = .004$, partial $\eta^2 = .06$), revealing that accelerando across the first two quarters did not yield any significant changes in speed of hand movement, whether the amplitude was equal ($t(63) = –.71$, $p > .40$, $r = .09$) or decreasing ($t(63) = –.05$, $p > .90$, $r = .01$), but a marginally significant increase when the amplitude was increasing ($t(63) = 1.86$, $p = .068$, $r = .23$), as shown in Figure 5-14 below.

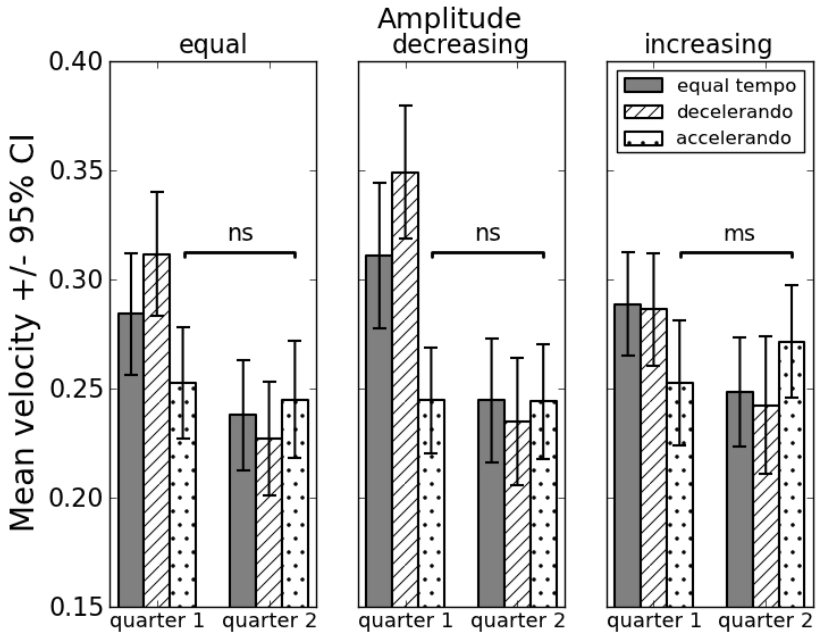


Figure 5-14 Influence of interaction between loudness contour and tempo profile on speed of hand movement in the first half of the auditory stimuli. ns: not significant, ms: marginally significant

For the second half, there was also a significant interaction effect between 'quarter', 'loudness' and 'tempo' ($F(3.45, 206.77) = 2.44$, $p = .048$, partial $\eta^2 = .04$), revealing that accelerando across the second two quarters did not yield any significant changes in speed of hand movement when the amplitude was decreasing ($t(63) = -.14$, $p > .80$, $r = .02$), but a significant increase in speed when the amplitude was equal ($t(63) = 3.09$, $p = .003$, $r = .36$) or increasing ($t(63) = 4.32$, $p < .001$, $r = .48$), as shown in Figure 5-15 below.

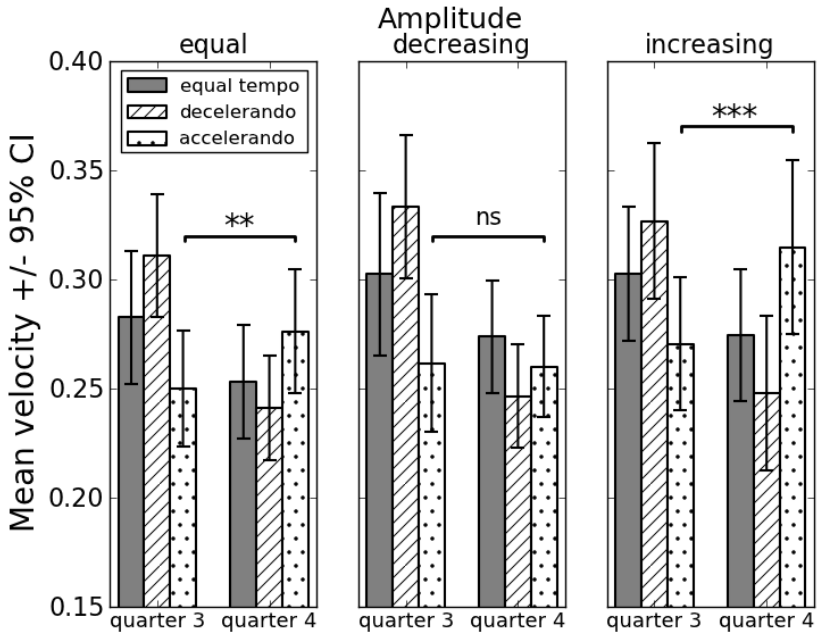


Figure 5-15 Influence of interaction between loudness contour and tempo profile on speed of hand movement in the second half of the auditory stimuli. ** indicates $p < .01$, *** indicates $p < .001$, ns: not significant

Finally, there was a significant interaction effect between 'quarter', 'pitch' and 'vision' ($F(1, 60) = 7.58$, $p = .008$, partial $\eta^2 = .11$), revealing that the speed of hand movement in the first halves of each half (quarters 1+3 combined) differed significantly between rising and falling pitch in the visual condition ($t(63) = 5.09$, $p < .001$, $r = .54$), but only marginally in the non-visual condition ($t(63) = 1.88$, $p = .065$, $r = .23$), whereas in the second halves of each half (quarters 2+4 combined) the effect was reversed such that the speed of hand movement differed significantly between rising and falling pitch in the non-visual condition ($t(63) = 2.13$, $p = .037$, $r = .26$), but not in the visual condition ($t(63) = 1.42$, $p > .10$, $r = .18$). Apart from this interaction, Figure 5-16 below also indicates significant main effects of 'pitch' and 'vision'. Regardless of any other influence, rising pitch compared to falling pitch ($F(1, 60) = 28.29$, $p < .001$, partial $\eta^2 = .32$), and the presence of the visualization compared to its absence ($F(1, 60) = 14.32$, $p < .001$, partial $\eta^2 = .19$), yielded significantly faster hand movements.

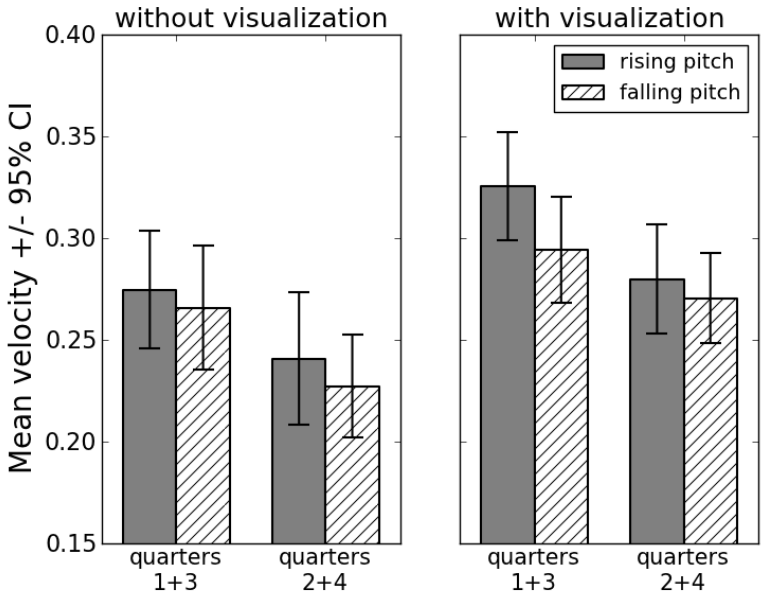


Figure 5-16 Influence of interaction between visualization and pitch contour on speed of hand movement

5.4 Discussion

5.4.1 Summary of main findings

Asking 64 participants to represent gesturally a set of pure tones, I analysed their representations of elapsed time, pitch, loudness and tempo, taking into account interactions between musical parameters within the sound stimuli. Regardless of direction, elapsed time was most strongly associated with the x-axis, pitch with the y-axis, and loudness with the y-axis as well, though the latter finding turned out to be a spurious effect caused by concurrent changes of pitch and loudness. All participants showed positive correlation coefficients between pitch and height, and this association was larger in the non-visual condition compared to the visual condition, and larger for musically trained compared to untrained participants. Moreover, rising-falling pitch contours led to higher correlation coefficients than falling-rising pitch contours, which is mainly due to musically untrained participants' lower values when presented with the latter contour. Equal tempo patterns compared to accelerandi patterns and constant amplitude compared to both decreasing-increasing and increasing-decreasing loudness contours gave rise to larger pitch–Y correlation coefficients (i.e. increase in pitch represented by upward movements). What is more, musical training and visual feedback influenced an interaction effect between pitch and tempo, which is discussed in detail below.

Notwithstanding the spurious loudness–height association for stimuli concurrently varied in pitch, loudness and tempo, those stimuli that only varied in loudness did reveal loudness–height associations: they were larger for increasing-decreasing compared to decreasing-increasing loudness contours, and musically trained participants showed higher values than untrained participants. There was also an interaction between training and sex revealing that female musically untrained participants showed lower values than their male counterparts, whereas no difference was observed between female and male musically trained participants. The hypothesized association between loudness and z-axis was only found in stimuli that only varied in loudness, and only for musically trained participants when the loudness contour was increasing-decreasing.

Muscular energy analysis revealed that mostly untrained participants used shaking hand movements to represent loudness, and further (interaction) effects of musical parameters—such as a clearer representation of increase compared to decrease in loudness—are discussed below. Finally, speed of hand movement was associated with tempo and influenced by musical training (untrained participants did not increase speed of hand movement when tempo increased) and interactions with pitch (falling pitch prevented increase in speed when tempo increased) and loudness (decreasing loudness prevented increase in speed when tempo increased). The visualization generally led to faster hand movements and differentially affected the speed of hand movements in an interaction with pitch (see detailed discussion below).

5.4.2 Elapsed time

Even though in language the metaphorical use of the z-axis (“ahead”, “behind”) seems to be more commonly used to refer to events in time, there is some evidence that gestural representations of time are mapped onto the x-axis (Casasanto & Jasmin, 2012). The strong association between elapsed time and the x-axis found in the present study—more specifically, the movement from left to right—is in line with human conceptualization of time as spatial (for a review see Núñez & Cooperrider, 2013). Note, however, that a positive correlation coefficient implies only an *overall* movement from left to right and does not exclude the possibility of (short) hand gestures including circular movements or instances of moving from right to left (e.g., if participants ran out of space). And comparing the present results with representations of elapsed time in drawings—although there might be different mechanisms at work here as discussed below—reveals that they appear to be less clear-cut. For instance, all of

Athanasopoulos and Moran's (2013) 25 British musically trained participants went from left to right to depict a comparable set of pure tones graphically, whereas in the present study, three (out of 32) trained participants showed no clear association between elapsed time and x-axis in either condition. Moreover, three trained participants, who predominantly went from right to left in the non-visual condition, showed movement from left to right only in the visual condition, which might have triggered the awareness of the numerous visual instances in our culture (e.g., written texts, calendars, timeline in graphs and, of course, music scores) where progression of time is represented as left-to-right movement. The fact that an even larger number of musically untrained participants went from right to left, and were seemingly not influenced by visual feedback, suggests that the amount of exposure to visual representations of time as progressing from left to right (as in music scores) might have an impact on the—conceptually or bodily—spatial mapping of time.

The analysis of local correlation coefficients revealed that the movement from left to right was more consistent when the stimuli contained accelerandi rather than decelerandi. Although speculative, this might reflect the fact that intensifying stimulus features—such as increasing tempo in this case—are more salient than attenuating ones because they are more significant in the environment: an object accelerating poses a greater potential threat than an object that decelerates (see Neuhoff, 2001, for a discussion of the adaptive value of changes in loudness). As I have shown, increasing-decreasing patterns give rise to more consistent mappings than decreasing-increasing ones, thus this effect might be part of a more general phenomenon of cross-modal mappings.

5.4.3 Pitch

The strong association between pitch and height corroborates findings from previous studies applying a range of different paradigms such as motion imagery (Eitan & Granot, 2006), drawings (Küssner & Leech-Wilkinson, 2014), gestures (Nymoen et al., 2013) and forced choices (R. Walker, 1987). Musically trained participants showing higher correlation coefficients than untrained participants is in line with previous studies, too (Küssner & Leech-Wilkinson, 2014; R. Walker, 1987), as is the finding that rising-falling pitch contours gave rise to higher correlation coefficients than falling-rising pitch contours (Kohn & Eitan, 2012). However, I was able to show that the latter effect is heavily influenced by training, revealing that only untrained participants, but not trained participants, show more consistent associations for rising-falling

pitch contours compared to falling-rising contours. Just as right-handed pianists need to train their left hand more extensively to achieve equally strong hands, musicians are trained to play rising-falling contours to the same standard as falling-rising contours, even though the latter contour might not feel as natural as the former. What is more, this interaction was further mediated by the tempo pattern: Both musically trained and untrained participants showed higher values when pitch and tempo patterns were concurrently increasing in the first half of the stimuli (i.e. rising pitch and increase in tempo) and moving contrarily in the second half of the stimuli (i.e. falling pitch and increase in tempo) compared to when pitch and tempo patterns were moving contrarily in the first half of the stimuli (i.e. falling pitch and increase in tempo) and concurrently increasing in the second half of the stimuli (i.e. rising pitch and increase in tempo). There are at least three different factors interacting here. First, the gestural pitch–height representation of decreasing pitch paired with an increase in tempo is facilitated by the laws of gravity: an object falling towards the ground accelerates. Secondly, faster processing of congruent semantic correspondences such as increasing pitch and increasing tempo, which both represent increasing intensity, facilitates accelerated upward movements. The third factor needs more explanation. The type of the pitch contour (rising-falling vs. falling-rising) is evidently crucial for the resulting association between pitch and height. While the roles of natural laws and conceptual metaphors have been discussed before in the context of cross-modal mappings (Johnson & Larson, 2003), the role of the pitch contour for embodied cross-modal mappings awaits further research. One mundane explanation could be the (lack of) effort to move the hand in a higher start position: it is simply more comfortable to wait for the beginning of a trial with the arm hanging loosely beside the body.

In addition to effects of pitch or interactions of pitch with other factors, there were also effects of loudness and tempo that influenced the size of the pitch–height correlations. Generally, constant amplitude and constant tempo led to higher correlation coefficients, which can be explained neatly by the attentional resources available: it is easier to represent pitch with the height of one's hand when pitch is the only sound feature that is varied. The role of attention might however be influenced by the amount of musical training, as the interaction effect between tempo and training illustrates. It was revealed that musically trained participants show higher correlation coefficients for accelerando-accelerando compared to decelerando-decelerando patterns, while untrained show the opposite behaviour: higher values for

decelerandi compared to accelerandi. The gradual increase in tempo—which is possibly more ecologically relevant than a decrease, as discussed earlier—might entrap untrained participants to shift their attention away from tracing the pitch with the height of the hand, focussing more on (representing) changes in tempo. One reason why this effect—though much smaller—is reversed for trained participants could be that an increase in tempo gives rise to heightened attention for the very same reason pertaining to the salience of intensifying sound features (see also Eitan & Granot, 2007), but with the result that pitch is now traced more alertly. Further empirical investigation is needed to find out whether this training-induced differential effect of attention on embodied cross-modal mappings is maintainable.

5.4.4 Loudness

The disclosure of the spurious loudness–height association in stimuli varied in several auditory features is perhaps not surprising for a musical culture largely based on pitch. When confronted with opposing pitch and loudness contours, participants chose to represent pitch, not loudness, on the y-axis. Importantly, this shows that pitch–height associations dominate loudness–height associations in a context of concurrently varied sound features, putting the results reported by Kohn and Eitan (2009)—that loudness–height associations of sound features varied in isolation are stronger than pitch–height associations—and the conclusion drawn by Eitan (2013a, p. 182)—that the "hierarchy of musical parameters delineating musical space and motion may conflict with the parametric hierarchy assumed by many music theorists" (i.e. pitch and duration first, loudness second)—into perspective. Of course, this does not mean people do not display loudness–height mappings (Eitan et al., 2008). As shown for stimuli only varied in loudness (Nos 2 & 3), there exists an association between loudness and the vertical axis, which is larger for increasing-decreasing than decreasing-increasing contours (see also Kohn & Eitan, 2012) and larger for musically trained compared to untrained participants. But compared to other mappings such as time–X and pitch–Y, this association turned out to be rather weak.

Similarly, the hypothesized association between loudness and the z-axis—particularly in the visual condition in which the z-axis was related to size (Lipscomb & Kim, 2004; R. Walker, 1987), but also in the non-visual condition in which it was related to distance (Eitan & Granot, 2006)—was almost non-existent for stimuli concurrently varied in pitch, loudness and tempo. One crucial difference between my experimental paradigm and that of Eitan and Granot—apart from the distinction between real and imagined movement—is possibly the fact that movement

in Eitan and Granot's study involved the relational movement of an imagined humanoid character to the stable position of the participant, whereas in the present study only one (real) person was involved. Even more importantly, moving forwards could be achieved either by moving only the arm or the whole body forwards. Thus, in both cases, though particularly in the latter, real sense of distance was unlikely to be involved.

Nevertheless, the analysis of stimuli without changes in pitch (Nos 2 & 3) revealed a very clear pattern: increasing-decreasing loudness contours—but not decreasing-increasing loudness contours—are represented by movements along the z-axis such that an increase (decrease) in loudness led participants to move forward (backward). And, as observed several times before, musically trained participants showed higher scores than untrained participants, whose mean correlation coefficient in fact suggests a complete absence of associations between loudness and the z-axis. Untrained participants did, however, represent changes in loudness as can be seen in the profiles of muscular energy (see Appendix 5.4) and will be discussed in the following section.

5.4.5 Muscular energy and loudness

The first observation to make is that increase in loudness is more strongly associated with increase in muscular energy (shaking events) than decrease in loudness with decrease in muscular energy, as can be seen from stimulus No. 2 (see *p* values in Table 4). However, if increase in loudness precedes decrease (No. 3), the latter *is* represented with decrease in muscular energy, suggesting that the order seems to play a crucial role for the existence of the association between decreasing loudness and decreasing energy. Interestingly, the strong association between increasing-decreasing loudness and increasing-decreasing muscular energy has been found in the drawing experiment as well (see Chapter 3): the association between pressure on a tablet—resulting in a thicker line when more pressure was applied—and loudness (more pressure for louder sounds) was more consistent when the loudness contour was increasing-decreasing rather than decreasing-increasing. What is more, these contour effects of loudness–energy are reminiscent of the contour effect observed for pitch–height associations: there seems to be a general preference for inverted U-shape contours, whether the mapping involves pitch–height or loudness–energy associations.

The second observation pertains to the interaction between pitch and loudness. When pitch is falling and loudness increasing, the association between increased loudness and increased muscular energy is annulled (Nos 7 & 19 and, to a lesser degree, No. 21) or even reversed (Nos 8 [trained participants] & 9 [visual condition]). Complementarily, when both pitch and loudness are increasing (No. 16), there is a very strong association between increased loudness and increased muscular energy, and this association is even stronger when the increasing pattern occurs in the first half of a stimulus (No. 10), with the exception of musically trained participants who, just as for the remainder of the stimuli, do not make much use of muscular energy, i.e. shaking events, when representing changes in loudness. Kohn and Eitan (2009) also report an association between muscular energy and pitch, which comes third after energy–loudness and energy–tempo associations. Taken together, this suggests that while the primary association of pitch seems to be spatial height, pitch does interfere with loudness to the extent that the usually observed distribution of muscular energy (increased loudness – increased energy) can be reversed when pitch and loudness contours are opposed.

Thirdly, the influence of tempo variations on pitch and loudness patterns seems to be unsystematic, if not puzzling. For instance, muscular energy patterns in response to sound stimulus No. 12, which is concurrently increasing and decreasing in pitch, loudness *and* tempo, does not reveal such clear results as muscular energy patterns in response to stimulus No. 10, which is "only" concurrently increasing and decreasing in pitch and loudness. Moreover, the decelerando-decelerando tempo profile of sound stimulus No. 11 seems to have (almost) no impact on the association between muscular energy and loudness/pitch observed for stimulus No. 10.

Finally, and perhaps most bafflingly, falling pitch and increasing loudness of the first half of stimulus No. 20 gave rise to strong associations between increased loudness and increased muscular energy, even though we have just seen that falling pitch overrules increasing loudness (e.g., No. 19). This difference can only be due to the concurrent decelerating tempo pattern of stimulus No. 20, which is absent in No. 19. Perhaps tempo does, after all, have an impact on the muscular energy profile such that the effect of falling pitch—presumably represented by a downward hand movement—and decelerating tempo—associated with, as we have seen before, a very stable decrease in speed regardless of pitch and loudness—cancel each other out (Eitan and Granot [2006], for instance, report that falling pitch is associated with increasing

speed) if, and only if, the increasing loudness profile occurs in the first (No. 20) rather than the second half (No. 8) of a stimulus. If that were true, the influence of the inverted U-shape would be more pervading and prevailing than previously thought.

5.4.6 Speed of hand movement

Although pitch had been associated with speed in adjective matching (P. Walker & Smith, 1986) and rating tasks before (Eitan & Timmers, 2010), no such association was found in the present study. Similarly, there was no clear association between loudness and speed – a result that might have been biased by the stimuli involved in this analysis. One third of them—i.e. the ones with equal tempo (Nos. 4, 7, 10, 13, 16, 19)—included one second of unchanged pitch at the end of each half of a stimulus (see also Methods).[58] It is possible that participants stopped gesturing briefly when reaching these points, creating a 'slowing down' bias at the end of each half.[59] This potential bias notwithstanding, the fact that the speed of hand movement decreased when the loudness decreased and that the speed decreased to a lesser extent when the loudness increased suggests that loudness did have an influence. At least partly, then, this finding suggests a gap between imagined and real bodily cross-modal mappings. While Eitan and Granot (2006) found no association between decreasing loudness and decreasing speed in a rating task, the present study, as well as that of Kohn and Eitan (2009), provides evidence for such a correspondence.

The association between tempo and speed of hand movement is more straightforward. With increasing tempo participants increase the speed of their hand movements, and with decreasing tempo they slow down. Musical training, however, significantly influences this effect, such that untrained participants do not show an increase in speed of hand movement when the tempo is accelerating but only a decrease in speed when the tempo is decelerating. While differences between musically trained and untrained participants pertaining to imagined speed have been reported before for stimuli varied in inter-onset intervals and articulation (Eitan & Granot, 2006), the present interaction effect between tempo and training presents a novel finding.

[58] Previous research has indicated that musically trained participants continue drawing a horizontal line when presented with pitch unchanged over time, while untrained participants stop drawing for a moment and only continue when pitch changes again (Küssner & Leech-Wilkinson, 2014). The absence of this effect in the interaction between 'quarter', 'tempo' and 'training' suggests, however, that gesturing sounds produces different results from drawing sounds.

[59] It is most likely for the same reason that the speed of hand movement decreases across two quarters of a stimulus (see Figure 5-11B, Figure 5-12, Figure 5-13, Figure 5-14 and Figure 5-15) when the tempo is equal.

Crucially, other concurrently varied auditory features such as pitch and loudness influence the association between tempo and speed too. First, while the direction of pitch has no influence on the association between decelerating tempo and decrease in speed, falling pitch inhibits increase in speed in response to accelerating tempo. Note that falling pitch—represented by a downward hand movement— paired with accelerating tempo manifests the prototypical *physical* prerequisites for accelerated movement: an object (here the hand) accelerating towards the ground. There is, however, no increase in speed, which could be explained by semantics taking precedence over gravity. If falling pitch is conceived of as LESS and accelerating tempo is conceived of as MORE, this might create a semantic conflict, preventing the speed of hand movement from increasing. Another explanation could be the sense of intensity that is felt when various musical parameters interact. When musical parameters are aligned (e.g., falling pitch and decreasing tempo), the resulting change in speed mirrors the feeling of intensity that is created by this alignment (e.g., decrease in speed). When musical parameters are opposed, however, the resulting change in speed (if any) is much harder to predict, as it depends on the salience of individual musical parameters that, in their sum, determine whether one feels the intensity increasing, decreasing or perhaps ambiguous.

Secondly, and similar to the pitch interference, decrease in speed of hand movement in response to decelerating tempo was unaffected by concurrently varied loudness. However, when the amplitude was equal or decreasing across the first two quarters, the association between accelerating tempo and increased speed disappeared. Only when both loudness and tempo increased did the speed of hand movement increase. This pattern is even clearer across the second two quarters, revealing that only decreasing loudness, but not equal or increasing loudness, results in a lack of increasing speed in response to accelerating tempo.

Taken together, these findings substantiate not only evidence of the association between tempo and speed in bodily cross-modal mappings (Kohn & Eitan, 2009), but also provide new insights into how interactions of auditory features affect the resulting speed of the hand movement.

5.4.7 The roles of musical training, sex and visual feedback

The findings from the present study provide further evidence that musical training is a factor influencing the consistency of cross-modal mappings. In line with previous research (Eitan & Granot, 2006; Rusconi et al., 2006), both pitch—particularly falling-rising pitch—and loudness

are mapped more consistently by musically trained participants. It needs to be tested to what extent sensorimotor skills play a role here (Küssner & Leech-Wilkinson, 2014) and how auditory, tactile and motor perception interact when mapping sound features cross-modally in real-time. What is more, musical notation might play a crucial role here, too, and it would be very valuable to compare cross-modal mappings of musicians who use notations with those who do not.

Importantly, the present findings provide evidence for qualitative differences as well. The almost complete absence of muscular energy to represent loudness in trained participants (see Appendix 5.5) raises the question of whether fast shaking hand movements with the Wii™ Remote Controller are perhaps too uncommon for musicians trained to manipulate loudness in very specific ways on their instruments. Studying percussionists' bodily cross-modal mappings therefore seems an obvious starting point for further investigations.

The two observed sex effects—that female participants showed higher time–X correlation coefficients than male participants when the amplitude was equal compared to decreasing-increasing and that female untrained participants showed lower loudness–Y correlations coefficients than their male counterparts (whereas there were no differences between male and female musically trained participants)—were unexpected and open new pathways for investigation. The latter effect hints at the possibility that musical training may be biased towards the perceptual tendencies of males. Although there is some research concerning the role of sex in choosing an instrument (Hallam, Rogers, & Creech, 2008) and neurophysiological sex differences in pitch processing (Gaab, Keenan, & Schlaug, 2003), the realm of individual differences in embodied cognition is still in its infancy (Keehner & Fischer, 2012), and I am not aware of any previously reported sex effects pertaining to cross-modal mappings of sound and music.

The situation is similar for the role of visual feedback in bodily cross-modal mappings of sound. Strikingly, the association between pitch and height was significantly stronger in the absence of the visualization, and this effect was further qualified by the interaction effect between pitch and tempo. While the pitch–height associations were generally higher when rising-falling pitch contours were paired with accelerando-accelerando patterns, there was no difference in the non-visual condition between equal tempo patterns paired with either rising-falling pitch

contours or falling-raising pitch contours, whereas there were increased values in the visual condition when equal tempo patterns were paired with rising-falling pitch contours rather than falling-rising ones. It is possible that the visualization encouraged participants to explore different ways of representing pitch, since "seeing the sound" on a screen is surely an invitation to move it. It is puzzling, however, that the visualization differentially decreased the pitch–height correlation coefficients in a manner that was dependent on interactions of various sound features. Finally, the visualization also influenced the speed of hand movement such that its presence led to faster movements overall, which might be accounted for by a feeling of increased mobility when participants "see" the sound moving as an object on a screen. And again, we observe that the visualization differentially varied the speed of hand movement at the beginning and ending of pitch rises and falls such that the end of a rising pitch contour compared to a falling contour was accompanied by faster hand movements when there was no visualization, whereas there was no difference when the visualization was present. For the beginning of a rising pitch contour the effect was reversed: it was accompanied by faster hand movements when the visualization was present, while there was no difference without visualization. More studies are necessary to shed light on these differential effects of visual feedback.

5.4.8 Preference for convex shapes

One recurring finding of the present study is the preference for convex shapes (increasing-decreasing contours). Although this effect was hypothesized for pitch mappings based on previous findings (Kohn & Eitan, 2009), its pervasiveness in other mappings (e.g., of loudness) and more complex interactions between musical parameters suggests a prominent role in gestural cross-modal mappings. Drawing on findings from dance and movement therapy, Kestenberg-Amighi and colleagues (1999; as discussed in Eitan, 2013a) propose a general preference for inverted U-shape contours based on the natural tendency of the body—and its various functions, e.g., respiration, heart rate—to grow first before shrinking. What is more, Kohn and Eitan (2009) remind us that 'rise before fall' is also a commonly observed pattern in music that has been widely discussed in musicology. For instance, analysing a large database of Western folk songs, Huron (1996) showed that convex melodic shapes are much more common than any other melodic contour, and Leech-Wilkinson (forthcoming) recently discussed the role of increasing and decreasing intensities ("feeling shapes"), drawing on Stern's

psychoanalytic theory of Forms of Vitality (Stern, 2010). Indeed, our life is saturated with rising-falling contours, starting with the structure of the day (from sunrise to sunset), over the speed of locomotion (increase in speed followed by decrease; either actively when walking, or passively when using public transport), to any grasping or more fine-tuned movements in order reach a pen or turn up the volume of the radio. And as discussed earlier, increasing stimulus properties in any sensory modality—higher, louder, brighter, warmer—imply the approach of a potentially harmful object, raising an organism's attention and alertness.

5.4.9 Limitations and future directions

There are a few limitations which need to be considered when interpreting the current dataset and designing future studies. To begin with, there was the lag of the real-time visualization (ca. 100-150 ms), which, in fact, was a lag of the data capturing procedure in general and was hence present in both conditions. The visual condition necessitated a compromise between smoothness of visualization and speed of writing to disk. In future studies not concerned with creating a real-time visualization, it may be advantageous to reduce the time-lag in order to achieve more accurate measurements.

A further limitation is that the order of the non-visual and visual conditions was not balanced since visualization first would have had a considerable impact on the way participants respond subsequently without visualization, having fresh memories of what their movements created visually on the screen in front of them only a few moments before. Inevitably, the spontaneity of participants' responses was reduced in the visual condition, which was regarded as the lesser of two evils.

Generally, one needs to be conscious of the nature of the cross-modal mappings measured experimentally—whether spontaneous or, as it were, mandatory—since apart from the paradigm itself, the instruction may crucially influence what is being measured (Rusconi et al., 2006). I chose the expression "represent sound gesturally" over instructions emphasizing a more communicative aspect of the gestures, e.g., "while listening to the music, move to it in an appropriate way, such that another child could recognize the music while watching your movements without sound" (Kohn & Eitan, 2009, p. 235) or, pertaining to sound drawings, asking participants to "represent the sound on paper in such a way that if another member of their community saw their marks they should be able to connect them with the sound"

(Athanasopoulos & Moran, 2013, p. 190). Although constituting seemingly negligible differences in instruction, the resulting drawings and gestures may give rise to different outcomes, particularly in a cross-cultural context as discussed by Eitan (2013b).

Another limitation is the decision to increase the size of the disk when moving the hand forward, i.e. away from the body. Moving forward was regarded as equivalent to 'more' (i.e. increase = bigger) in a three-dimensional space with its origin *in* the participant. Perhaps it would have been more intuitive to decrease the size when moving the hand forward, resulting in an alignment of distance and size: a small, soft object is far away, whereas a large, loud object is close. This could explain the absence of a stronger association between loudness and size in the visual condition but cannot account for the absence of loudness–distance associations in the non-visual condition. Thus, future experiments should include a systematic variation of the distance-size relationship.

What is more, muscular energy—conceptualized in the present study as fast (shaking) hand movements—does not account for instances in which muscles might be tense without any hand movements involved. Thus, in future studies, electromyography might be used to encompass further instances in which muscular energy is involved.

Next, the design of the stimuli needs attention. First, it should be acknowledged that tempo variations were not completely systematized to avoid an exponential increase in experimental stimuli: when the tempo was changed it consisted either of two decelerandi or two accelerandi, but never of a mixture of both tempi. Secondly, when several auditory features were varied concurrently, change of direction always happened at the same time after four seconds. Needless to say, in musical performances there can be all sorts of overlaps (e.g., a slow crescendo over several rising-falling pitch glides including a short decelerando at the end), creating a complex interplay of increasing and decreasing intensities that my set of pure tones is unable to match. And while my stimuli could have been much more complex to come closer to real musical stimuli, they could have included simpler variations as well—e.g., a single pitch ascent with concurrently varied loudness or tempo—to study the basic gestural mappings in more details. Thus, there is scope for future studies to investigate both ends of the spectrum. Thirdly, and perhaps most crucially, when varying pitch, loudness and tempo concurrently, the variations of individual sound features might be differentially salient. That is, it matters whether

the pitch range encompasses half an octave or four octaves, or whether the change in loudness occurs over 80% or 10% of the maximum amplitude. It is therefore not implausible that pitch—not loudness—was represented on the y-axis because it was perceptually more salient. Had the pitch range only included four semitones (or had it been in a different register) and the change in loudness been made more extreme, it might well have resulted in loudness–height associations. Researchers thus need to take great care when designing auditory stimuli that are varied in several sound features.

Finally, it should be pointed out that the findings presented here do not capture the unique ways in which participants might have represented sound gesturally, not only because the applied motion capture system is insensitive to fine-grained hand movements but also because participants might have used—consciously or subconsciously—other parts of their bodies to represent the sound. While the focus here was on averaged responses of hand movements to get insight into a largely under-researched field, the role of fine-grained movements of hands, fingers and other body parts provides a fruitful path to explore in future studies.

5.5 Conclusion and implications

In this chapter I investigated gestural representations of pitch, loudness and tempo, providing a solid empirical basis for future studies concerned with bodily cross-modal correspondences. I was able to show that musical training plays an important role in shaping bodily cross-modal mappings, e.g., giving rise to more consistent mappings and annulling the commonly observed bias for convex shapes. Loudness–size and loudness–distance associations appear to be less relevant if untrained individuals are provided with the opportunity to link loudness to energy level, which can be seen as the fundamental physical factor influencing amplitude (i.e. deflection of air molecules). Moreover, concurrently varied musical parameters have a significant effect on the ways in which people represent sound gesturally: interactions between pitch and loudness affect how participants adjust the speed of their hand movement. While it remains to be seen what the underlying mechanisms (e.g., perceptual, semantic) of these bodily cross-modal mappings are, the findings provided here may provide further support for the existence of recently developed concepts within embodied music cognition such as Godøy's (2006) 'gestural-sonorous objects', emphasizing the interconnection of motion and sound features in the mind of the listener. Facilitated by advances in multimedia technology (Tan, Cohen, Lipscomb, & Kendall, 2013) and the development of new musical instruments, the

increasingly complex role of movement in creating and manipulating sounds and music challenges findings of cross-modal correspondences that have been obtained with traditional paradigms. Future studies need to address whether findings from bodily cross-modal mappings can be integrated wholly into current theoretical frameworks or whether "embodied cross-modal mappings" might form a separate category worth studying in its own right. Besides theoretical implications, the outcome of the present study, as well as its low-cost motion capture devices, may be used in clinical settings where sounds and music are used to co-ordinate movement. For instance, music-based movement therapy has been found to be effective in treating Parkinson's disease (De Dreu, Van der Wilk, Poppe, Kwakkel, & Van Wegen, 2012; Rochester et al., 2010), and therapeutic approaches to stroke may benefit from musical activities, as shown in a study using the Wii™ Remote Controller to develop new forms of interventions (van Wijck et al., 2012). Having discussed gestural cross-modal mappings of pure tones, I will now move on to real musical excerpts.

Chapter 6: Gestural cross-modal mappings of musical excerpts in real-time

6.1 Introduction

In the previous chapter, I have shown how musically trained and untrained individuals represent gesturally a series of pure tones varied in pitch, loudness and tempo. As I have pointed out, there is scope for further studies including more complex auditory stimuli. In this chapter, I will investigate how the same 64 participants represented sixteen musical excerpts with their hand and arm gestures. There are many ways of studying the relationship between gestures and music (see also Chapter 1), some of which are covered in the volumes *Musical gestures: Sound, movement, and meaning* (Godøy & Leman, 2010), *Music and gesture* (Gritten & King, 2006) and *New perspectives on music and gesture* (Gritten & King, 2011). While these volumes are grounded in musicology, the topic of music and gesture has also attracted attention in other fields such as Human-Computer Interaction (HCI), where researchers try to develop new systems, tools and applications that map gestures onto sound in real-time (e.g., Françoise, 2013). What most of these approaches have in common, though, is their focus on human gestures. Hatten defines human gesture as "any energetic shaping through time that may be interpreted as significant" (Hatten, 2006, p. 1). While such a broad definition certainly encompasses the gestures encountered in this part of the experiment, in which the shaping of one's gestures over time is meant to carry some information about the sonic features of the music, it is possible to narrow down the definition further. Using the terminology provided by Jensenius, Wanderley, Godøy and Leman (2010) the representational gestures of this experiment fit into the category of 'communicative gestures', as they potentially signal the sonic features to someone else, or, perhaps more appropriately, 'sound-accompanying gestures', which cover gestures that are not part of the sound-producing action "but follow the music" (p. 24). However, shaping music gesturally in terms of a cross-modal representation is not directly addressed by these definitions nor is it covered in the volumes mentioned above. It is latently present, for instance, in studies of parent-infant interactions or empirical investigations of conducting, but it rarely emerges as the central research question at hand. This lack of research and the challenges involved in studying it are aptly summarized by Kozak, Nymoen and Godøy (2012, p. 69):

> "[…], human movement to music is quite idiosyncratic, showing substantial variability in how listeners interpret sounds with their bodies. One important question that remains far from settled, therefore, is whether sequences of musical sounds can have an effect on gestures that is quantifiable, and whether we can detect specific sound and movement features that can be correlated for the majority of listeners."

Note their subtle change of language as they are approaching the core issue: They begin by using the term 'music', followed by 'sound', before arriving at 'sequences of musical sounds'. Far from being a criticism of their study, this change of language from the holistic (music) to the particular (a musical sound) exemplifies a commonly observed behaviour when it comes to studying musical gestures quantitatively, in particular on a group level. To study gestural representations of music empirically (e.g., in experiments) usually means breaking down both gestures and music into smaller manageable chunks. The previous chapter is yet another example of such an approach, and while these methods can lead to exciting, thought-provoking new findings (see Introduction of Chapter 5), it may be very worthwhile digging at the other end of the spectrum too, and look at the whole rather than its parts.

There are only very few studies which investigated how adults represent music, that is, real musical excerpts as opposed to a set of musical features (pitch, loudness, timbre, etc.), with free hand and arm gestures. Haga (2008) asked three trained dancers and three untrained individuals to respond with spontaneous gestures to various musical excerpts including pieces by Vivaldi and Ligeti as well as one electronic piece of music composed for the purposes of the study. The results of this observational study showed that there was generally broad consensus among trained and untrained participants. The more detailed and complex the musical excerpt, the more variation was observed in the gestures. Interestingly, the dancers were often seen adding their own interpretative gestures to fill parts in the musical excerpts when a pulse was missing. Moreover, it was observed that dancers developed their gestures upon repeated presentation of a musical excerpt, remembering what they had done previously and exploring further gestural renderings of the music.

In another study, (Western) participants were asked to move a joystick in response to three pieces of traditional guqin Chinese music (Leman et al., 2009), thus their movements were relatively restricted in comparison with free gestural responses. Having participants repeatedly

listening and gesturing along with the music over four sessions in which each piece was presented twice consecutively, it was revealed that the relative number of consistent responses—i.e. similar velocity patterns—increased over the course of the experiment, especially for the two more melodic pieces of the experimental stimuli. These two melodic pieces also led to increasingly similar movement responses across participants, while the third piece—described by the authors as having "a more narrative character with less fluent melodic line" (p. 264)—gave rise to increasingly idiosyncratic movement responses. Besides recording participants' movements, Leman and colleagues also recorded the movements of the musician and correlated it with the listeners' movements. It was found that the correlation between the musician's shoulder movement and the participants' arm movement increased over the course of the experiment for the two melodic pieces, suggesting that the movement velocity patterns are not only shared between listeners but also to some extent between musician and listener.

In a recent study by the same group (Maes, Van Dyck, et al., 2014; Van Dyck, 2013), the relationship between music and movement was investigated by comparing listeners' free movement responses to music with their linguistic descriptions of the expressive qualities of the music. The piece of music used in this study was the beginning of the first movement of Brahms' First Piano Concerto. The participants were told to

> "translate [their] experience of the music into free full-body movement. Try to become absorbed in the music and express your feelings in a bodily fashion. There is no good or wrong way to execute this task (Van Dyck, 2013, p. 71)."

While the participants moved during the whole length of the excerpts, the authors identified three respective 'heroic' and 'lyric' passages, each 30 seconds long, for the purpose of their analyses. Based on Laban's Effort-Shape model, participants rated the expressive qualities of the excerpts on a bipolar scale consisting of twenty-four adjectives, sixteen pertaining to effort and eight to shape.[60] Using a motion capture system, seven movement features were extracted and matched to the effort and shape categories. Results revealed that all the movement features clearly differentiated between the two types of excerpts. For instance, if the average value for 'acceleration' was high for the heroic passages, it was low for the lyric passages. Interestingly, there was an effect of musical training, indicating that trained participants showed

[60] For more information about Laban's Effort-Shape model see Van Dyck (2013, pp. 67-68).

higher values for the movement features ‘size’ and ‘height’. This suggests that they moved more and filled more space with their gestures during the experiment, possibly because they felt more comfortable moving to the music.[61] Regarding the analysis of the linguistic expressions, there was large agreement among the participants as to how well a particular adjective described the expressive qualities of the music. Moreover, it was found that the extremes of the movement features correlated with the extremes of the adjective scales such that an excerpt which was rated, for instance, as conveying the expressive qualities ‘big’, ‘broad’, ‘thick’ and ‘exalting’, also gave rise to a high value for the movement feature ‘size’.[62] The authors interpret their findings as evidence for the sharing of expressive qualities of music in linguistic expressions and body movements.

These holistic approaches—where real musical excerpts, or perhaps even whole compositions, are used as experimental stimuli—might reveal many new insights into embodied cross-modal mappings and carve the way for future studies by addressing the questions raised by Kozak and collaborators. Since the experimental procedure has already been laid out in detail in Chapter 5, it will suffice to describe the musical stimuli at this point. These were mainly chosen to achieve a rich variety of genres, instrumentation, textures and complexities (e.g., comparison of monophonic and polyphonic pieces; representation of pure tones within a musical piece), and to include potentially interesting dynamic shapes. I chose short excerpts rather than complete compositions to enable a greater variety of musical stimuli in a short amount of time, trying to avoid any possible effect of participant fatigue. An overview of stimuli from both practice and experimental trials can be seen in Table 6-1 below.[63] For more detailed information about the excerpts, see Appendix 6.1.

[61] The authors assessed the participants’ perceived freedom of movement on a Likert scale from 1 (= totally disagree) to 5 (= totally agree) after the experiment. Results showed that the median of musically trained participants was 4, whereas the median of untrained participants was 2.
[62] Since the order of the two parts—movement and self-report—were counterbalanced, it is unlikely that the verbal descriptions influenced the movements (or vice versa).
[63] The sound stimuli can be downloaded at http://tinyurl.com/nqmn3ej.

Table 6-1 Overview of musical excerpts used in experimental and practice trials

Composer	Title	Length (in s)	Type
Bach, J. S.	Partita for Solo Violin No. 3 in E major BWV 1006 (Giga)	28.78	E
Bach, J. S.	Partita No. 1 in B-flat major BWV 825 (Courante)	24.00	E
Berg, A.	Wozzeck, Act 3	18.74	E
Berlin, I.	This Year's Kisses	18.63	P
Boulez, P.	Répons, Section 1	32.90	E
Bruckner, A.	Symphony No. 8, 4th movement	22.80	E
Carrothers, B.	I Can't Begin To Tell You (from "Keep Your Sunny Side Up")	24.06	E
Chopin, F.	Prelude Op. 28, No. 6 in B minor (M. Argerich)	7.32	E
Chopin, F.	Prelude Op. 28, No. 6 in B minor (A. Cortot)	8.09	E
Chopin, F.	Etude Op. 10, No. 5 in G-flat major	15.30	P
Debussy, C.	Prelude, Book 1, II. Voiles	31.06	P
Ferneyhough, B.	no time (at all)	13.01	E
Grupo Fantasma	El Consejo (from "El Existential")	16.81	E
Grupo Fantasma	Reconciliar (from "El Existential")	20.53	P
Messiaen, O.	Vingt Regards Sur l'Enfant Jésus, I. Regard du Père	21.85	E
Mozart, W. A.	Horn Concerto No. 4 in E-flat major K. 495 (Rondo)	23.80	E
Radiohead	The Butcher (from "Supercollider / The Butcher")	28.41	E
Satriani, J.	The Forgotten (Part Two) (from "Flying in a Blue Dream")	24.59	E
Schönberg, A.	Verklärte Nacht, Op. 4	30.15	E
Stravinsky, I.	Three Pieces for Solo Clarinet, 2. Quarter (crotchet) = 168	20.41	E

Note. E = Experimental Trial, P = Practice Trial

6.2 Exploring gestural representations of music

Before turning to the gestural responses themselves, it will be worthwhile considering a comparison of participants' perceptions—i.e. their subjective experience—of the two types of experimental stimuli, namely pure tones and musical excerpts. After the completion of the experimental blocks (see Figure 5-2), participants were asked to rate, separately for pure tones and musical excerpts, the perceived difficulty of the tasks (1 = not difficult at all, 5 = very difficult) and the perceived consistency of their responses (1 = very inconsistent, 5 = very consistent) on 5-point Likert scales. Data were analysed using mixed-measures ANOVAs with

the within-subjects factors ‘type’ (pure tones / musical excerpts) and ‘visual’, and the between-subjects factors ‘sex’ and ‘training’. Results of both ratings are presented and discussed below.

6.2.1 Difficulty ratings

There was a significant main effect of ‘type’ ($F(1, 60) = 11.77$, $p = .001$, partial $\eta^2 = .16$), indicating that participants’ perception of the difficulty of representing music gesturally was greater ($M = 2.67$, SEM = .12) than that of representing pure tones ($M = 2.16$, SEM = .09). There was also a main effect of training, showing that, overall, untrained participants ($M = 2.62$, SEM = .11) found the representation tasks more difficult than trained participants ($M = 2.21$, SEM = .11), $F(1, 60) = 7.28$, $p = .009$, partial $\eta^2 = .11$.

However, these main effects need to be considered in light of two interaction effects. First, there was a marginally significant interaction between ‘type’ and ‘training’ ($F(1, 60) = 3.82$, $p = .055$, partial $\eta^2 = .06$), suggesting that only musically trained participants perceived the difficulty of the representation task differently. This was confirmed by paired-samples *t*-tests, showing that musically trained participants perceived representing music ($M = 2.61$, SEM = .14) as more difficult than representing pure tones ($M = 1.81$, SEM = .13), $t(31) = –4.75$, $p < .001$, $r = .65$, whereas no difference of perceived difficulty was found among untrained participants (music: $M = 2.73$, SEM = .19; pure tones: $M = 2.51$, SEM = .14), $t(31) < 1$, $p > .30$. This outcome makes sense given the results of the drawing experiment (see Chapter 3), in which musically trained participants approached pure tones and musical excerpts with the same strategy, namely attempting to represent all discernible sound features in their drawings. Consequentially then, if presented with an excerpt from an orchestral piece consisting of complex textural layers, representing *all* its sonic features becomes incredibly difficult, if not impossible.

Secondly, there was a marginally significant interaction between ‘sex’ and ‘training’ ($F(1, 60) = 3.89$, $p = .053$, partial $\eta^2 = .06$), showing that male trained and untrained participants did not differ significantly (trained: $M = 2.40$, SEM = .15; untrained: $M = 2.51$, SEM = .18), $t(30) < 1$, $p > .60$, whereas female untrained participants ($M = 2.73$, SEM = .14) rated the tasks more difficult than female trained participants ($M = 2.02$, SEM = .13), $t(30) = –3.73$, $p = .001$, $r = .56$. Although the average values suggest sex differences also within the groups of untrained, and especially trained, participants these trends were not significant. Nevertheless, these findings lend further empirical support for a hypothesis briefly sketched in the previous chapter: that musical

education may be biased towards perceptual abilities more commonly observed in men than in women.

Moreover, there was a main effect of 'visual' ($F(1, 60) = 5.17$, $p = .027$, partial $\eta^2 = .08$), indicating that the condition with visualization on the screen was perceived as more difficult ($M = 2.54$, SEM = .10) than the condition without visualization ($M = 2.29$, SEM = .08). This effect was not further qualified by an interaction, thus it is reasonable to assume that the perceived difficulty applied equally to trained and untrained participants. The novelty that this experimental setting imposed on participants, as well as the slightly delayed response of the visualization, may well explain such an outcome.

6.2.2 Consistency ratings

There was a significant main effect of 'type' ($F(1, 60) = 5.57$, $p = .022$, partial $\eta^2 = .09$), indicating that participants rated the consistency of their gestural representations higher in the part with pure tones ($M = 3.56$, SEM = .08) compared to the part with music ($M = 3.30$, SEM = .11). This is understandable insofar as the pure tones were much simpler, thereby making it easier for the participants to apply their strategies consistently.

There was also a significant interaction effect between 'training' and 'visual' ($F(1, 60) = 12.26$, $p = .001$, partial $\eta^2 = .17$), and follow-up t-tests revealed that trained participants rated their gestural representations more consistent in the condition without visualization ($M = 3.72$, SEM = .12) compared to the condition with visualization ($M = 3.23$, SEM = .16), $t(31) = 3.10$, $p = .004$, $r = .49$, whereas there was no significant difference in consistency ratings—but a tendency in the opposite direction (without visualization: $M = 3.27$, SEM = .12; with visualization: $M = 3.50$, SEM = .15)—among untrained participants, $t(31) = –1.73$, $p = .093$, $r = .30$. This pattern is also reflected by a significant difference between trained and untrained participants' consistency ratings in the condition without visualization ($t(62) = 2.76$, $p = .008$, $r = .33$), while a comparison in the condition with visualization revealed no significant difference ($t(62) = –1.23$, $p > .20$, $r = .15$). In fact, several musically trained participants reported in the feedback interview that they found the visualization distracting, while some even chose to close their eyes. One mentioned that it "made the music less enjoyable" and that "the visualizations in his mind were far more sophisticated", complaining that it was "impossible to visualize tone colour, accent, harmony and complex rhythms." Indeed, large inter-individual differences and preferences for

visualizations of music make it impossible to develop what might be called a 'generic' visualization of music. But to think that there are certain commonalities in the ways people perceive the dynamic changes of music is perhaps not too far-fetched. How these are mapped visually depends largely on a (highly individualised) lifetime of listening experiences. Next, I will consider what people actually did when asked to represent short musical excerpts gesturally.

6.2.3 Ways of representing music gesturally

What it means to represent music gesturally is a delicate question. As reviewed in the previous chapter, short action-related sound snippets can give rise to highly idiosyncratic responses (Caramiaux et al., 2014), which is why it will be worthwhile investigating the responses to various short musical excerpts. While it is conceivable that individuals' representational responses to music and sound differ between cultures (Athanasopoulos & Moran, 2013; Eitan, 2013b)—although drawing tasks might be of a different kind compared to gesturing tasks—I have shown in Chapter 3 that training influences representational strategies also within cultures. It is thus expected that similar differences will be found when comparing gestural cross-modal mappings of music. Participants were asked in this part of the experiment to "depict/represent the music with [their] arm and hand movements as the music is played. [They] should start gesturing when the music starts and stop gesturing when the music stops." It was emphasized "that there are NO 'right' or 'wrong' ways to represent music with arm and hand gestures! We would like you to depict the music with your gestures in a way that feels natural to you." When participants had follow-up questions and asked whether certain kinds of approaches were acceptable, the experimenter encouraged them to pursue whatever strategy they thought might be appropriate to represent music with gestures. During the practice and experimental trials the experimenter was seated behind a screen and was unable to see what the participants were doing. This served as a way of increasing participants' "free" responses since moving one's body to music in front of a stranger might have caused some individuals to feel intimidated.

Exploring the videos of participants' movements after completion of the experiment revealed that the interpretation of "representation" differed sometimes substantially between participants. To get a better idea of what the participants did during the experiment, two music students were trained to annotate the videos. Each of them only annotated one half of the stimuli due to the large total number of videos (N = 2560, including practice and experimental trials, and conditions with and without visualization). They were trained to indicate categorically whether a

video clip contained any of the following four action categories: ‘abstract’, ‘dancing’, ‘conducting’, or ‘air playing’. ‘Abstract’ here refers to any arm and hand gestures intended to capture some sonic feature of the music (pitch, loudness, tempo etc.). To provide a reference for ‘dancing’ both student assistants were shown video clips featuring one participant who consistently represented the music through dance. ‘Conducting’ and ‘air playing’ such as air guitar or air violin were not further exemplified since both student assistants stated that they were confident with these categories. Importantly, the categories were not mutually exclusive, and student assistants were instructed to annotate as many actions per video clip as necessary: two different actions might be blended in one instance (e.g., dancing *and* conducting at the same time), as well as occurring consecutively (e.g., playing air guitar followed by abstract representations of the loudness). If none of these categories fitted, the student assistants were asked to choose the category ‘other’, which required providing a brief description of the observed movement. One example could be a participant performing continuous circles that are neither related to any sonic features—although they might be, such as repetition—nor recognizable as conducting gestures. What is more, the student assistants were instructed to estimate the time in percentage participants spent looking at the screen in front of them in the condition with visual feedback.

Although the choice of these categories might be debatable (e.g., ‘conducting’ could be seen as a subcategory of ‘abstract’) and there is bound to be some variation in the application of the annotation criteria across the two student assistants, the point here is to provide an overview of the diversity of participants’ approaches to representing music gesturally, rather than breaking down participants’ movements into smaller chunks and assigning precise labels. A distribution of participants’ ‘abstract’, ‘dancing’, ‘conducting’ and ‘air playing’ gestures can be seen in Table 6-2 below.

Table 6-2 Gestural representational strategies (in %)

	abstract		dancing		conducting		air playing		looking time
Musical excerpt by composer	nv	v	nv	v	nv	v	nv	v	v
Bach, J. S. (violin)	71.9	70.3	1.6	0.0	14.1	7.8	9.4	3.1	74.5
Bach, J. S. (keyb.)	71.9	73.4	0.0	0.0	18.8	9.4	4.7	3.1	72.1
Berg, A.	60.9	64.1	0.0	1.6	12.5	7.8	7.8	3.1	78.9
Boulez, P.	84.4	82.8	0.0	0.0	3.1	1.6	0.0	0.0	79.1
Bruckner, A.	42.2	54.7	3.1	1.6	20.3	12.5	4.7	1.6	75.5
Carrothers, B.	45.3	53.1	10.9	6.3	26.6	10.9	4.7	3.1	76.6
Chopin, F. (Argerich)	65.6	56.3	0.0	0.0	9.4	3.1	3.1	0.0	77.3
Chopin, F. (Cortot)	65.6	67.2	0.0	0.0	6.3	1.6	3.1	0.0	74.3
Ferneyhough, B.	89.1	92.2	0.0	1.6	1.6	1.6	3.1	1.6	73.5
Grupo Fantasma	53.1	57.8	14.1	14.1	15.6	7.8	7.8	4.7	75.9
Messiaen, O.	50.0	59.4	0.0	0.0	9.4	6.3	4.7	1.6	81.2
Mozart, W. A.	43.8	62.5	4.7	4.7	32.8	17.2	4.7	1.6	74.1
Radiohead	56.3	62.5	9.4	9.4	10.9	6.3	10.9	3.1	75.9
Satriani, J.	46.9	54.7	4.7	3.1	7.8	7.8	9.4	7.8	77.3
Schönberg, A.	67.2	75.0	0.0	0.0	10.9	7.8	4.7	1.6	78.2
Stravinsky, I.	87.5	84.4	0.0	0.0	9.4	3.1	1.6	0.0	77.7

Note. nv = non-visual, v = visual. 'Looking time' refers to the estimated time (in %) participants were looking at the screen in the visual condition.

First, it should be noted that while the majority of participants used abstract gestures to represent the musical excerpts, for a considerable number of participants ‘representation’ also meant conducting, playing air instruments or even dancing. However, it should be stressed that for *all* musical excerpts, the main type of gesture was abstract. Of course, the various types of representation are also stimulus-dependent. It is conceivable that the approach to the Ferneyhough piece, for instance, is largely abstract and does not involve many dance movements (only one out of 64 participants showed dance movements), whereas the Latin piece by Grupo Fantasma elicits more dance movements, conducting gestures and air playing, and comparably fewer abstract movements. One might say that the latter kind of music affords dance-like movements (for a discussion of gestural affordances of music see Godøy, 2010a) and that it is in fact unnatural to move only one’s right arm and hand to ‘represent’ this kind of

music. In fact, previous research has shown that music with a clear pulse engages people's whole bodies when moving to music (Burger, Thompson, Luck, Saarikallio, & Toiviainen, 2013), and individuals show more engaged dance activity and entrainment as the intensity of the bass drum increases (Van Dyck et al., 2013). To investigate the association between pulse clarity and types of gestures, MIRtoolbox 1.5 (Lartillot & Toiviainen, 2007) was used to extract pulse clarity (Lartillot, Eerola, Toiviainen, & Fornari, 2008). The obtained values were then correlated with the percentage scores from Table 1.

Results revealed a very clear pattern: There was a negative correlation between pulse clarity and abstract gestures, $r = -.70$, $p = .003$ (with visualization: $r = -.58$, $p = .019$), and a positive correlation between pulse clarity and dancing, $r = .81$, $p < .001$ ($r = .80$, $p < .001$), conducting, $r = .60$, $p = .014$ ($r = .66$, $p = .006$) and air playing, $r = .55$, $p = .027$ ($r = .63$, $p = .009$). In other words, the clearer the pulse of the music, the more dancing, conducting and air playing, and the fewer abstract gestures there were.

Next, I will examine the distribution of the types of gestures in more detail. In the previous chapter, we have seen remarkable differences in representing sound gesturally between musically trained and untrained participants, as well as some sex effects. Therefore, in Table 6-3 and Table 6-4, the results of chi-squared tests are listed, comparing the percentages of male and female, and trained and untrained participants, respectively. No differences were found for the 'dancing' category, probably due to the relatively low occurrence overall.

Table 6-3 Chi-squared tests comparing the distribution of representational strategies among male and female participants

	abstract	conducting	air playing
Musical excerpt by composer	nv	nv	v
Bach, J. S. (keyb.)		M: 14.1% F: 4.7% $\chi^2(1) = 3.70$ $p = .055$, $\phi = -.24$	
Berg, A.	M: 23.4% F: 37.5% $\chi^2(1) = 5.32$ $p = .021$, $\phi = .29$		
Bruckner, A.	M: 14.1% F: 28.1% $\chi^2(1) = 5.19$ $p = .023$, $\phi = .29$		
Mozart, W. A.	M: 15.6% F: 28.1% $\chi^2(1) = 4.06$ $p = .044$, $\phi = .25$	**M: 25.0% F: 7.8%** **$\chi^2(1) = 8.58$** **$p = .003$, $\phi = -.37$**	
Satriani, J.			M: 0% F: 7.8% $\chi^2(1) = 5.42$ $p = .053$, $\phi = .29$

Note. nv = non-visual condition, v = visual condition, M = male, F = female. The chi-squared test in bold indicates significance at the Bonferroni-corrected significance threshold of $\alpha = .003125$.

Looking at sex effects, it can be seen that for three excerpts—namely, Berg, Bruckner and Mozart—female participants show more abstract gestures than male participants in the condition without visualization. At least for the Mozart excerpt, this pattern can be explained fairly straightforwardly, as significantly more men than women chose to conduct while listening to this piece. Conducting was also displayed more often by men than women when listening and gesturing along with Bach's Partita for Keyboard (marginally significant at $p = .055$). And more women than men used air playing to represent the Satriani piece in the condition with visualization. One might argue, however, that the significance threshold needs to be adjusted since sixteen tests were run. When treating each musical excerpt as a separate analysis the Bonferroni-corrected significance threshold is thus $\alpha = .003125$. Applying this stricter criterion, the only 'surviving' sex effect refers to conducting gestures in response to the Mozart excerpt. Similar to the sex effects in the previous chapter, such effects are still poorly understood and

should be interpreted with great caution only. That more men than women used conducting gestures to represent music could be seen as an effect largely shaped by cultural factors. The vast majority of conductors in the classical music business is male and all “first times” of women in this sector—such as conducting a famous orchestra or at a special occasion—usually draw large media attention. Such a (socially constructed) gender distinction is not supported by empirical findings. Wöllner and Deconinck (2013) found that when presented with point-light displays of male and female conductors ‘in action’, individuals are unable to guess the sex correctly – an effect that increases with increasing skill of the conductors involved. What is more, studying children’s kinaesthetic responses to music, Kerchner (2000) provides evidence that girls, rather than boys, choose to conduct in response to a musical excerpt. She describes the case of two female fifth-grade participants who

> “chose to depict their music listening experience by pretending to be the conductor of the orchestra that played the musical excerpt. Through their conducting gestures, I observed their reaction to melodic rhythm, dynamics, articulation, beat, subdivided beat, phrase, subphrase, embellishment, and the families of instruments in the orchestra. They chose a musical activity—conducting—to describe that which they perceived and responded to musically (p. 43).”

Kerchner (2000, p. 42) also observed that “fifth-grade males tended to use reserved foot and arm movements as they remained in one place in the movement area”, whereas “fifth-grade females used their entire body in the entire movement area (p. 43).” These findings seem to suggest, then, that cultural factors are the driving force behind differences found in adults, and as long as conducting remains a profession dominated by males, it should perhaps not come as a surprise if only a very small number of women chose to conduct to the Mozart piece.

Table 6-4 Chi-squared tests comparing the distribution of representational strategies among musically trained and untrained participants

	abstract		conducting		air playing
Musical excerpt by composer	nv	v	nv	v	nv
Bach, J. S. (violin)			T: 12.5% UT: 1.6% $\chi^2(1) = 6.34$ $p = .026$ $\phi = -.32$	T: 7.8% UT: 0% $\chi^2(1) = 5.42$ $p = .053$ $\phi = -.29$	
Bach, J. S. (keyb.)			T: 15.6% UT: 3.1% $\chi^2(1) = 6.56$ $p = .01$ $\phi = -.32$	T: 9.4% UT: 0% $\chi^2(1) = 6.62$ $p = .024$ $\phi = -.32$	
Berg, A.				T: 7.8% UT: 0% $\chi^2(1) = 5.42$ $p = .053$ $\phi = -.29$	
Bruckner, A.			**T: 18.8%** **UT: 1.6%** **$\chi^2(1) = 11.68$** **$p = .001$** **$\phi = -.43$**	T: 12.5% UT: 0% $\chi^2(1) = 9.14$ $p = .005$ $\phi = -.38$	
Carrothers, B.	T: 29.7% UT: 15.6% $\chi^2(1) = 5.11$ $p = .024$ $\phi = -.28$	T: 34.4% UT: 18.8% $\chi^2(1) = 6.28$ $p = .012$ $\phi = -.31$	T: 18.8% UT: 7.8% $\chi^2(1) = 3.93$ $p = .048$ $\phi = -.25$		
Grupo Fantasma					T: 7.8% UT: 0% $\chi^2(1) = 5.42$ $p = .053$ $\phi = -.29$
Messiaen, O.		T: 35.9% UT: 23.4% $\chi^2(1) = 4.15$ $p = .042$ $\phi = -.26$			
Mozart, W. A.				**T: 15.6%** **UT: 1.6%** **$\chi^2(1) = 8.89$** **$p = .003$** **$\phi = -.37$**	
Radiohead	T: 34.4% UT: 21.9% $\chi^2(1) = 4.06$ $p = .044$ $\phi = -.25$				
Satriani, J.	T: 29.7% UT: 17.2% $\chi^2(1) = 4.02$ $p = .045$ $\phi = -.25$			T: 7.8% UT: 0% $\chi^2(1) = 5.42$ $p = .053$ $\phi = -.29$	
Schönberg, A.			T: 10.9% UT: 0% $\chi^2(1) = 7.87$ $p = .011$ $\phi = -.35$	T: 7.8% UT: 0% $\chi^2(1) = 5.42$ $p = .053$ $\phi = -.29$	

Note. nv = non-visual condition, v = visual condition, T = trained, UT= untrained. The chi-squared tests in bold indicate significance at the Bonferroni-corrected significance threshold of $\alpha = .003125$.

Next, I will consider the effects of training. Comparing the numbers of abstract gestures between musically trained and untrained participants, there was a significant difference for four

musical excerpts: Carrothers, Messiaen, Radiohead and Satriani. In all these instances, trained participants showed more abstract representations of the music than untrained participants. As can be seen from Table 6-2, the excerpts by Carrothers, Radiohead and Satriani led quite a few participants to dance, conduct and play air instruments. Since previous research has shown that musicians are very keen on representing sonic features (Küssner, 2013)—regardless of whether the auditory stimuli consist of pure tones or musical excerpts—it fits into the picture that trained participants display this behaviour for gestural responses to a broad variety of musical genres too. Untrained participants seem to be more inclined to play air guitar or dance – if the music affords such behaviour. Note, however, that this does not mean that trained participants never dance or play air instruments to represent the music they are presented with (otherwise there would be a difference between trained and untrained participants in these categories). But they are more likely to accompany these kinds of representational movements with abstract gestures that refer to the underlying sonic features. The case for the Messiaen excerpt is quite different, as it does not invite one to dance, nor to play air instruments. But also the overall percentages in the 'abstract' category are fairly low (visual: 50%; non-visual 59.4%), indicating that a substantial number of participants used different approaches for this particular piece (see below for a more detailed description of the movement shapes). Interestingly, musically trained participants showed more air playing than untrained participants for one excerpt only. However, since this effect was only marginally significant (p = .053) and only 5 out of 32 trained participants engaged in playing an air instrument, it is perhaps safer to assume that by and large, the difference between trained and untrained participants with regards to air playing is negligible.

Quite a different picture emerged, however, from the comparison of conducting gestures. There were twelve significant effects, all of which revealed the same pattern: musically trained participants conducted more than musically untrained participants. In many cases this is true for both the visual and non-visual condition, suggesting that the influence of visual feedback did not interfere much with the choice of representation. One exception is the Mozart excerpt. With its very clear beat, it lends itself splendidly to conducting along, and even several untrained participants (eleven, in comparison to seventeen trained) opted for this approach in the condition without visual feedback. With the visualization present, however, this changed and the difference is now statistically significant. One might argue that it is hardly surprising that more

trained than untrained participants conduct in response to music – after all, this is part of a musician's training and especially orchestral players are regularly exposed to conductors' gestures. But it should be noted that for some musically trained individuals, this was the predominant, and—disregarding the Ferneyhough excerpt—in two cases the exclusive mode of representing the music gesturally.

6.3 Detailed piece-by-piece analysis: Movement data and verbal descriptions of alternative ways of representing music gesturally

I will now consider each musical excerpt in more detail by looking at some of the verbal descriptions of participants' movement in combination with data from the Kinect™ and Wii™ Remote Controller. As in the previous chapter, speed profiles and distribution of shaking events were calculated overall, as well as separately for musically trained and untrained participants and for the visual and non-visual conditions (see Appendix 6.2–6.6). Unlike the analysis of gestural responses to pure tones though, the present analysis does not include correlations between sonic features of the music and hand movements because it is not possible to extract pitch information from an orchestral piece the same way as can be done from a pure tone changing in frequency. Although other sonic features, such as perceived loudness or intensity may be extracted, it makes little sense to enter them in a correlation analysis because the stimuli, and also the nature of the task, required a different, more holistic, approach. To that end, a more qualitative approach was applied, comparing the speed profiles visually and discussing also alternative kinds of gestures – those that did not fit in any of the four action categories outlined above.

6.3.1 Bach: Partita No. 3 for Solo Violin

Comparing trained and untrained participants' speed profiles, there is at least one striking difference: the regularity of velocity peaks in the group of trained participants. These are presumably artefacts of clearly marking the beat of this piece. Note that this would be in line with the finding that more trained than untrained participants conducted along as a representation of this excerpt, as shown above. As the summary of alternative representation strategies suggests (see Table 6-5 below), there was a greater diversity in the visual compared to the non-visual condition, and the speed profile shows larger variation—with higher peaks—in

the visual condition. However, there does not seem to be a strict differentiation between trained and untrained participants in terms of varied representation strategies. In both groups, there are individuals "experimenting" with lines, circles and waves, for instance. Interestingly, one musically trained participant (see also Satriani excerpt) stood motionless during the whole excerpt.

Table 6-5 Gestures in response to Bach (Partita No. 3 for Solo Violin)

Untrained	Non-visual	Up and down motion; shaking controller; drawing small Us; figures of eight
Untrained	Visual	Small zigzags; horizontal figures of eight; drawing Us and diagonal lines; shaking controller; drawing circles and moving hand forwards and backward; drawing horizontal waves; pushing away from himself
Trained	Non-visual	Drawing waves; drawing horizontal lines
Trained	Visual	Moving right arm up and down; standing motionless with arm up and left; moving arm from side to side while shaking wrist; drawing lines / circles / triangles

6.3.2 Bach: Partita No. 1 in B-flat major (keyboard)

Even though the data from the Partita for Solo Violin suggest that there are, apart from the conducting gestures, no major differences between musically trained and untrained participants, a somewhat different picture emerges from this second Baroque piece. As can be seen in Figure 6-1 below, the trained participants' speed profile can be roughly divided into three parts: part one lasting from 0 to 13 seconds, part two from 13 to 21 seconds and part three from 21 to 24 seconds. In each part there is a slow increase in speed followed by a drop in speed before this process repeats itself. This shaping is not trivial but related to structural events that are brought to life in Leonhardt's performance. (I encourage the reader to listen to the recording. As outlined in Chapter 1, my focus is on music as sound rather than text. Thus, the score below in Figure 6-3, as well as all other scores that follow, should be seen as additional information rather than the main point of reference.) The first drop in speed marks the beginning of a new phrase with a rhythm of dotted quavers in the right hand. The second drop coincides with a single dotted quaver (E4), which stands out as the longest note (in the middle) of five bars, preparing the listener for the end of the phrase. These subtle aspects of a musical performance are picked up by the trained ear and implemented into the gestural representation, whether deliberately or not. Although it is possible that musically untrained participants are able to perceive these nuances—especially if their attention is directed towards these sonic events—

there is no hint of their taking into account of these subtleties on a group level (see Figure 6-2). Moreover, comparing visual and non-visual conditions, it can again be observed that the visualization led to more variation in the speed profile.

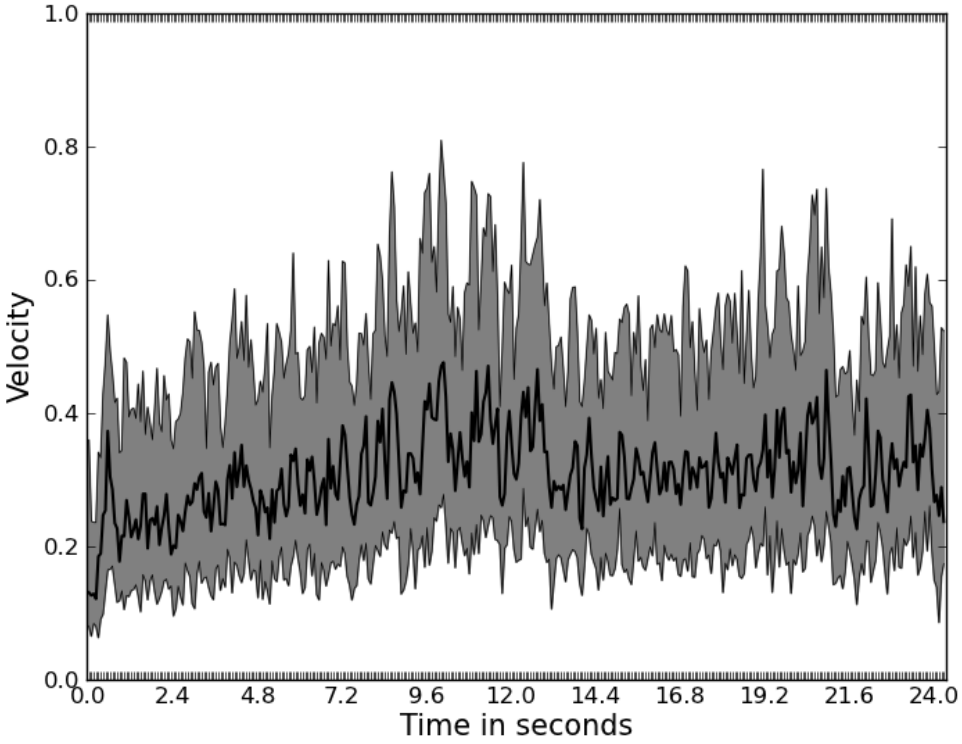


Figure 6-1 Speed profile of Bach's Partita No. 1 (keyboard) averaged across all musically trained participants

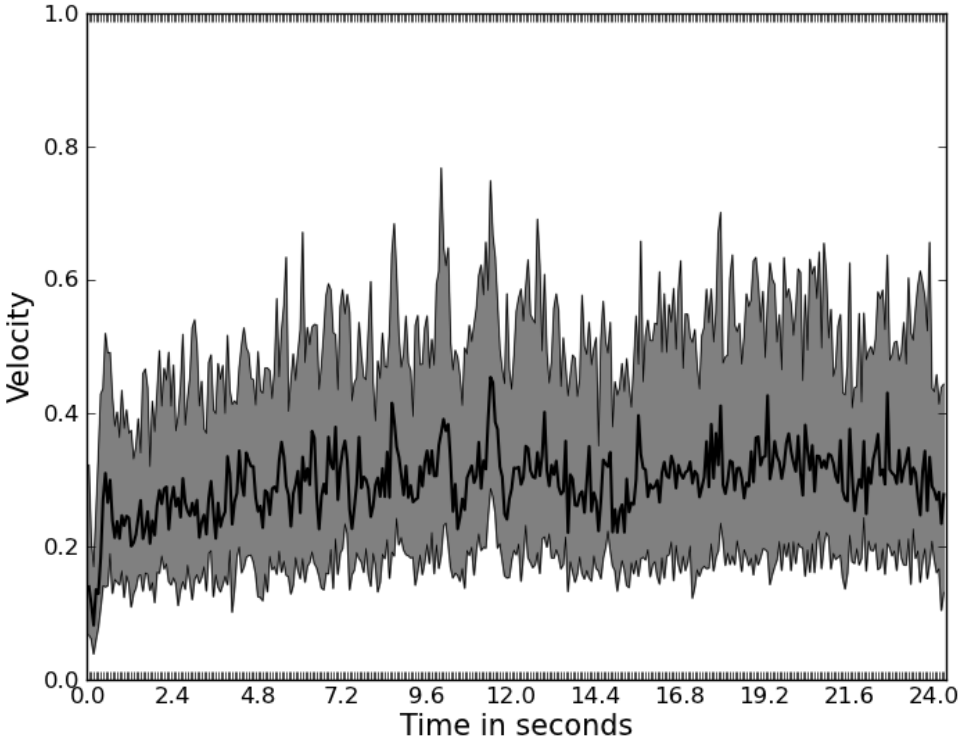


Figure 6-2 Speed profile of Bach's Partita No. 1 (keyboard) averaged across all musically untrained participants

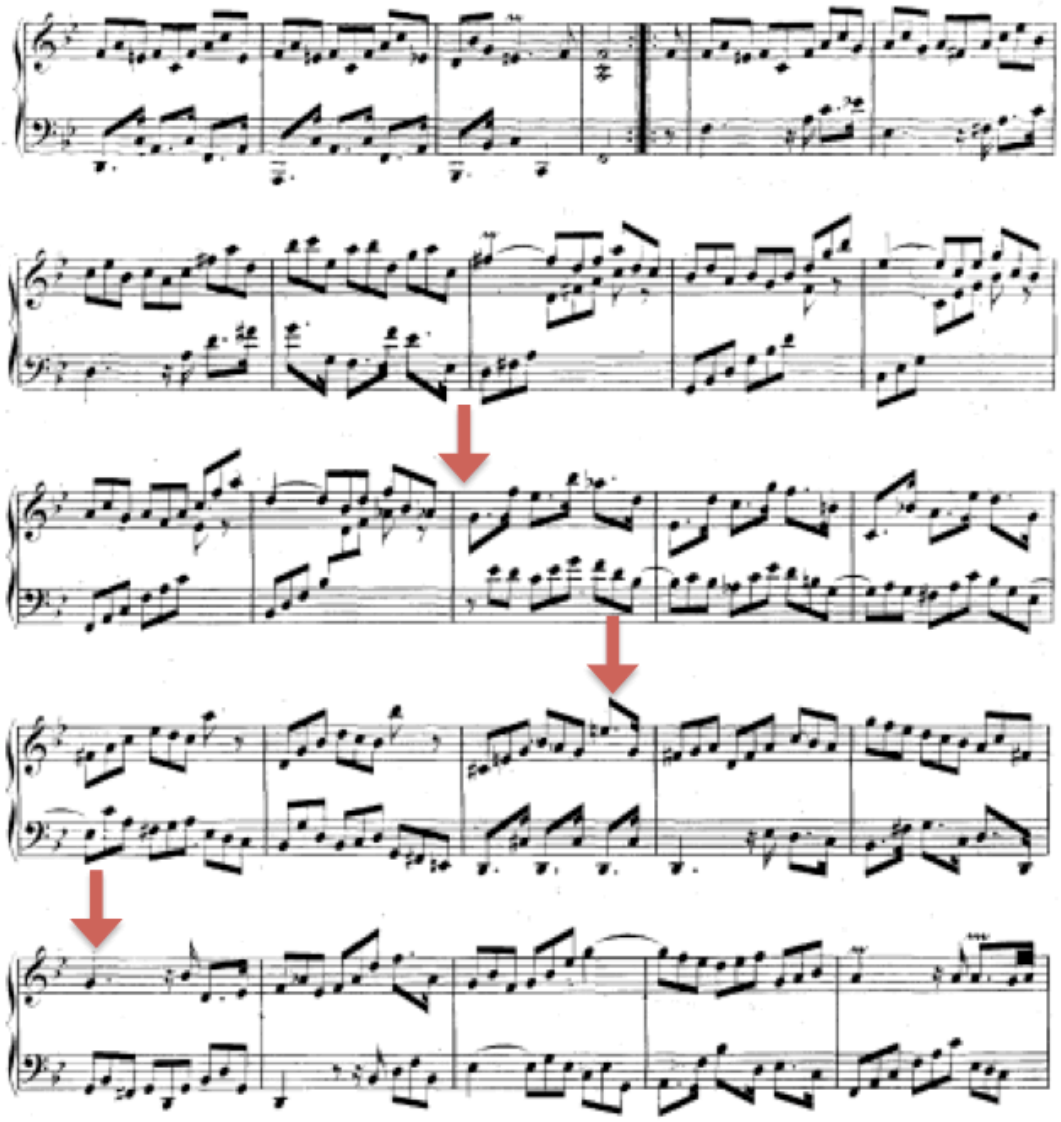


Figure 6-3 Score of Bach's Partita No. 1 (keyboard). The first two arrows indicate the drops in musically trained participants' speed of hand movements (see Figure 6-1). The last arrow indicates the end of the excerpt.

Table 6-6 Gestures in response to Bach (Partita No. 1 in B-flat major [keyboard])

Untrained	Non-visual	Outward pushes; up and down wrist motion; drawing small circles; one slow diagonal whilst shaking controller; disjointed figure of eight; drawing Us; shaking handset rapidly; drawing small waves
Untrained	Visual	Shaking controller; shaking wrist up and down whilst moving arm from side to side; drawing horizontal lines; drawing small spirals; drawing circles and moving hand forwards and backward; shaking handset rapidly
Trained	Non-visual	Drawing small circles / waves; moving arm from side to side whilst shaking wrist; drawing horizontal lines; shaking handset
Trained	Visual	Moving arm from side to side whilst shaking wrist; drawing diagonal lines up; drawing circles

6.3.3 Berg: Wozzeck

The first observation to make is that musically trained participants' speed profile shows more extreme peaks than those of untrained participants. The excerpt starts with enormous force and energy before gradually losing its momentum and becoming calmer and quieter. This pattern is mimicked in the speed of trained participants' hand movements, which, overall, gradually decreases as the excerpt unfolds. On the other hand, there is first a slight increase in speed of hand movement among the untrained participants before the speed decreases, making the profile look arc-shaped with a thinner end on the right hand side. Although speculative, it seems as though musically trained participants are more "present in the moment" (cf. Stern, 2004, 2010). They seem to gesture along with the music, from moment to moment, rather than following—or lagging behind— the music. Put differently, it could be argued that musically trained participants *act*, while untrained participants *react*. Recall that all participants first listened to each piece once before they were asked to represent it gesturally. And although many participants already used the first presentation to try out different gestural approaches, given this particular excerpt, only the trained participants seemed to have benefitted from a "practice round". The point being that while the sudden forceful beginning of this excerpt should have come equally expected for both groups, musical training seems to provide the skill (or advantage) of grasping the dynamic contour of a (novel) musical excerpt quickly and translating its dynamic shape into motor responses. Suffice it to say that the comparison of visual and non-visual conditions revealed once more a more varied speed profile in the condition with visualization.

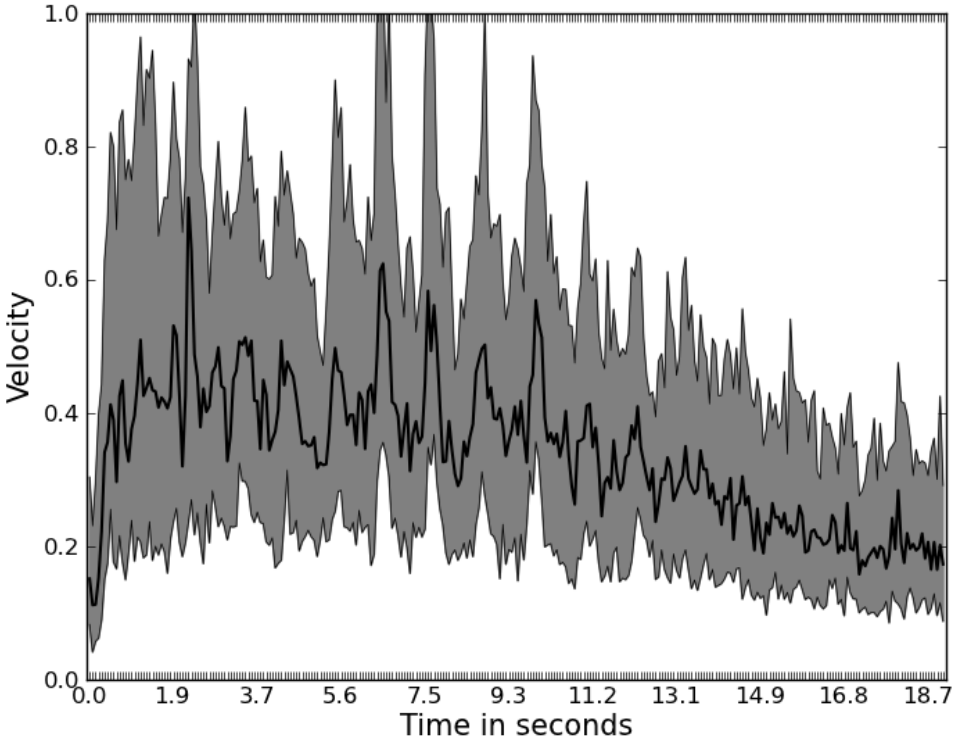


Figure 6-4 Speed profile of Berg excerpt averaged across all musically trained participants

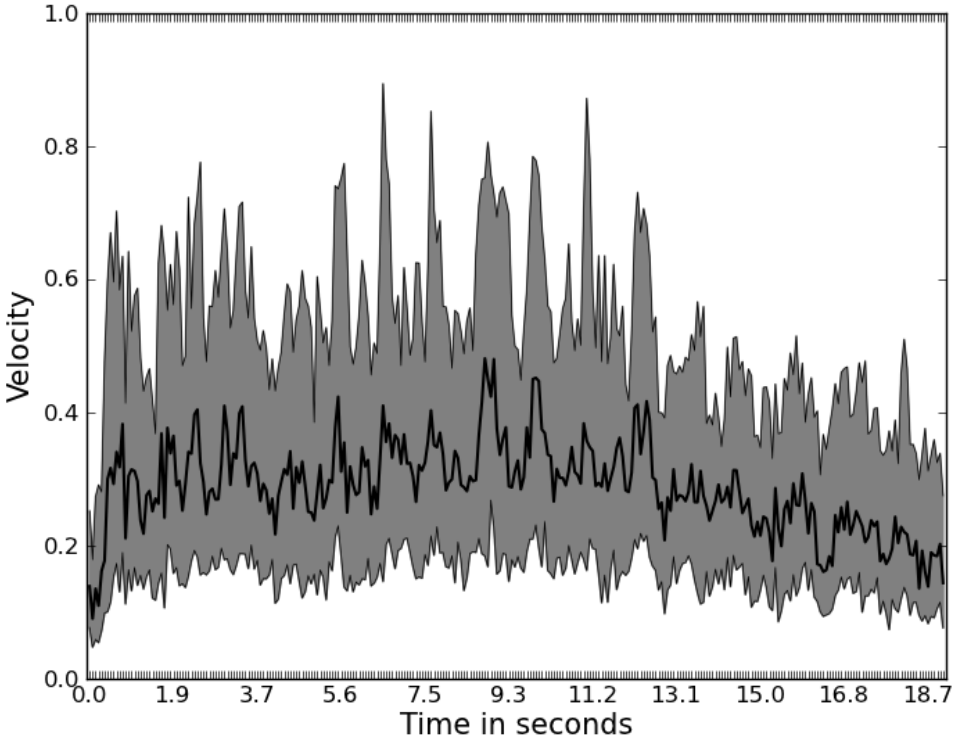


Figure 6-5 Speed profile of Berg excerpt averaged across all musically untrained participants

Table 6-7 Gestures in response to Berg (Wozzeck)

Untrained	Non-visual	Slow horizontal figure of eight; outward pushes with drum beat; drawing circles with drum beat; drawing Us; moving controller up and down; moving wrist up and down whilst moving arm side to side; shaking controller
Untrained	Visual	Slow backwards spirals; drawing Us; large horizontal arcs; drawing inverted Us with drums; large circles; leaning whole body forwards and backwards; shaking controller; drawing figures of eight; pushing arm in and out
Trained	Non-visual	Moving both arms up and down with beat; moving up and down diagonally with drum beat; drawing horizontal lines; shaking the handset
Trained	Visual	Moving both arms in and out to beat; drawing large figures of eight; drawing small circles; drawing horizontal lines

6.3.4 Boulez: Répons

Musically trained and untrained participants' gestural representations of this excerpt appear to be quite similar, except that the speed profile of trained participants is once again more distinct in the extremes. The excerpt is characterised by the interaction of various solo instruments (piano, harp, vibraphone, glockenspiel, cimbalom) which introduce short but distinct musical motives that seem to be echoed in a bubbling musical surface to the point where musical fore- and background become indistinguishable (and a new motive is introduced). The onset of these short motives—clearly visible in the waveform (see Figure 6-6 below)—are easily detectable in both trained and untrained participants' speed profiles, and the types of gestural representations (waves, circles, lines) are very similar across both groups too.

Table 6-8 Gestures in response to Boulez (Répons)

Untrained	Non-visual	Waves at various speeds; shaking controller; waving hands above head; holding arm out front whilst twisting body side to side; drawing small Us; drawing small circles; descending continual 'S' shape
Untrained	Visual	Shaking controller up and down; drawing circles above head; drawing figures of eight; waving arm above head
Trained	Non-visual	Drawing waves; drawing giant circles; drawing horizontal lines; shaking the handset
Trained	Visual	Moving right arm up and down; drawing circles; drawing waves, drawing horizontal lines

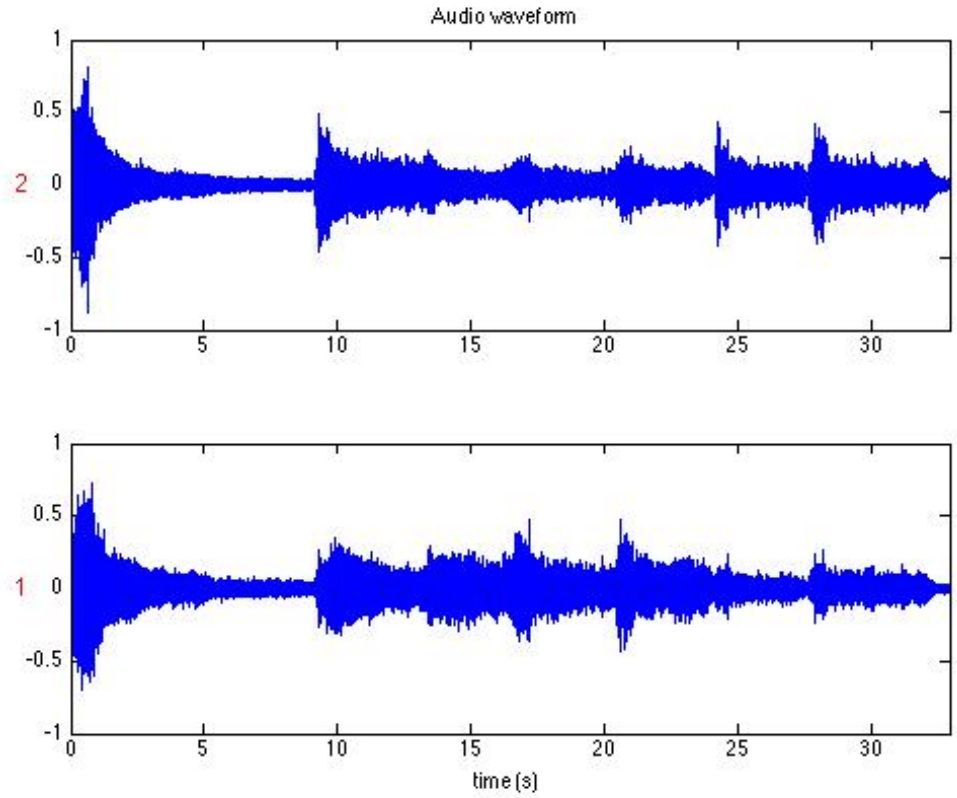


Figure 6-6 Audio waveform of Boulez excerpt (Répons)

6.3.5 Bruckner: Symphony No. 8

Examination of the speed profiles revealed one previously observed pattern—trained participants show more extreme speed variation than untrained participants—and one novelty: the visualization seems to have led to a somewhat more structured response compared to the non-visual condition (see Figure 6-7 and Figure 6-8 below). There are clear peaks and troughs in the speed profile that are related to the melody of the brass section. Thus, not only did the visualization engage faster and more varied speed of hand movements but it can also help carve out features of the musical surface. What is more, the regular staccato playing of the strings was apparently a very prominent feature, and trained as well as untrained participants used pushing or punching movements to mark the beat (see Table 6-9 below).

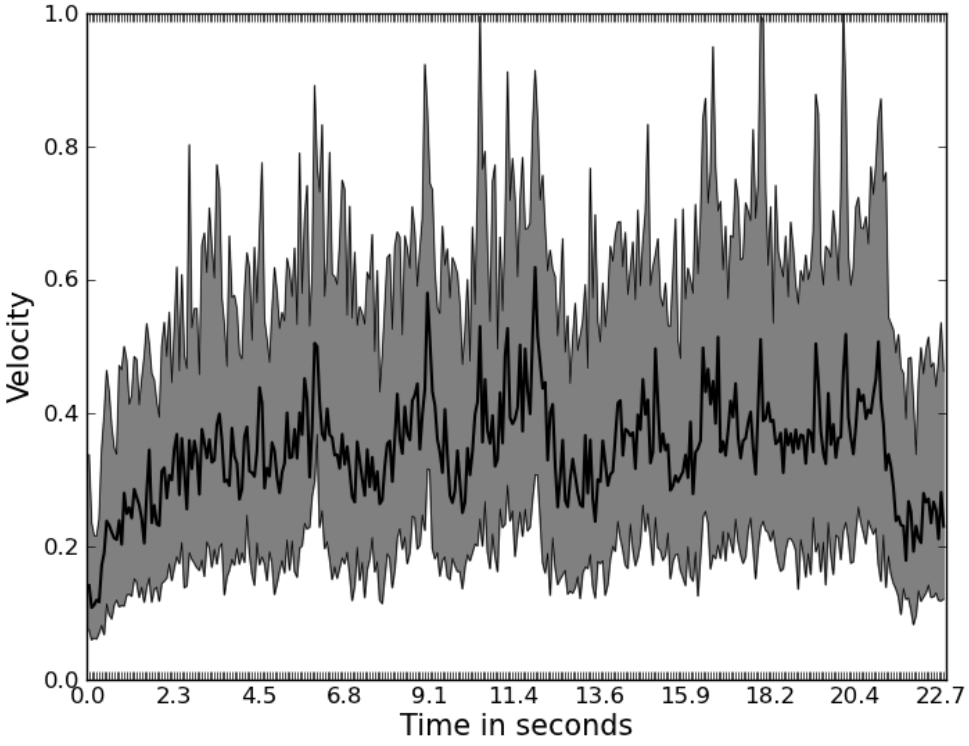


Figure 6-7 Speed profile of Bruckner excerpt averaged across visual condition

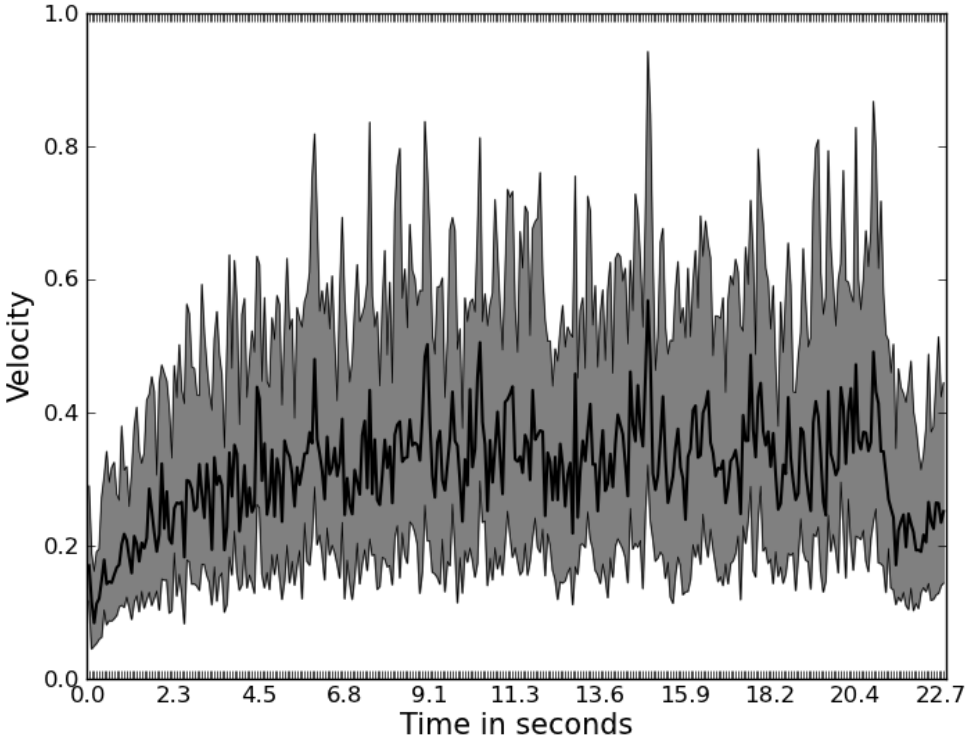


Figure 6-8 Speed profile of Bruckner excerpt averaged across non-visual condition

Table 6-9 Gestures in response to Bruckner (Symphony No. 8)

Untrained	Non-visual	Constant side to side motion of arm; shaking controller forwards and backwards; outward pushes (growing with intensity); up and down wrist motion; side to side and spirals; swinging arm side to side, then up and down; drawing Us
Untrained	Visual	Side to side swipes; drawing Us at various speeds; outward pushes; shaking controller; up and down arm motion that starts low and ends high; holding arms above head and pushing hands up; moving wrist up and down with beat; walking forwards whilst moving arm up and down; waving arm above head; drawing circle
Trained	Non-visual	Drawing horizontal waves; drawing wavy inverted Us; drawing horizontal lines
Trained	Visual	Horizontal chopping action; moving arm up and down; pushing in and out with beat; punching in front of his body and drawing large circles; drawing diagonal lines up; drawing circles and squiggly lines

6.3.6 Carrothers: I Can't Begin To Tell You

We have seen above that untrained participants do not readily take the subtleties of performance into consideration. If the musical shapes are clearly defined though—such as in the solo introduction of the excerpt by Bill Carrothers, which contains very noticeably separated musical figures—there is no apparent difference in trained and untrained participants' speed profiles nor in the distribution of the shaking events. Familiarity with the genre might generally be another factor involved, though as we will see later on, it seems to be secondary to the distinctiveness of a sonic event. It is also noticeable that this jazz piece triggered a broad variety of gestural responses in the untrained participants, whereas the trained participants showed the usual narrow range of responses.

Table 6-10 Gestures in response to Carrothers (I Can't Begin To Tell You)

Untrained	Non-visual	Swaying; drawing Us; small waves from side to side; outward pushes with beats 1 and 3; figure of eight; moving controller up and down; moving wrist up and down while moving arm from side to side; shaking controller forwards and backwards; drawing horizontal waves; shaking handset rapidly; freezes in the rests in the solo intro; drawing horizontal lines
Untrained	Visual	Bouncing figure of eight; large swipes up, down, side to side, spirals; side to side motion; horizontal line and wave; outward pushes with the beat; moving controller up and down; shaking arm up and down whilst twisting side to side; shaking controller forwards and backwards
Trained	Non-visual	Drawing Ws; moving arm up and down with beat; moving arm side to side and shaking wrist; drawing horizontal lines
Trained	Visual	Moving right wrist side to side; moving arm up and down (with beat); drawing waves; drawing horizontal lines

6.3.7 Chopin: Prelude Op. 28, No. 6 (performances by Argerich and Cortot)

The two shortest excerpts of the current set of stimuli were perhaps too short to reveal any particularities of their gestural renderings. It may be suggested that the performance by Argerich gave rise to a slightly faster speed of hand movement at the beginning of the excerpt, and that the performance by Cortot elicited slightly more shaking events at the beginning of the excerpt, in the group of musically trained compared to untrained participants, respectively. Other than that, the speed profiles appear to be interchangeable. And also the representation strategies did not differ much in both excerpts: neither between trained and untrained participants, nor between visual and non-visual conditions (see Table 6-11 [Argerich] and Table 6-12 [Cortot] below). The horizontal arcs that are chosen by many participants seem to follow the melodic line and are in accordance not only with findings from gestural representations of pure tones—where pitch is represented on the y-axis—but also with drawings of these pieces, as we have seen in previous chapters.

Table 6-11 Gestures in response to Chopin (Prelude Op. 28, No. 6 performed by Argerich)

Untrained	Non-visual	Slow U shapes; slow waves; large horizontal arcs; twisting both wrists; up and down wrist motion; figures of eight; shaking controller forwards and backwards; moving arm side to side; drawing small circles; descending diagonal line
Untrained	Visual	Drawing large horizontal lines; side to side waving; large horizontal arcs; drawing Us (in both hands); shaking controller; drawing circles; figures of eight; drawing spirals in both hands; raising and lowering arm; drawing large diagonal; pushing arm in and out
Trained	Non-visual	Drawing horizontal waves; drawing wavy inverted Us; drawing horizontal line
Trained	Visual	One big arch; drawing spirals; drawing large/wavy inverted Us (in both hands); drawing horizontal lines; drawing upwards diagonal line; drawing circles; drawing waves

Table 6-12 Gestures in response to Chopin (Prelude Op. 28, No. 6 performed by Cortot)

Untrained	Non-visual	Slow waves; outward pushes at various speeds; shaking controller; spirals; drawing big vertical zigzag; moving wrist up and down whilst moving arm side to side; swinging arm forwards and backward; holding arm up near shoulder and bending torso; drawing horizontal waves; drawing figures of eight; drawing a slowly descending diagonal line
Untrained	Visual	Drawing Us; large horizontal arcs while shaking wrist; shaking controller; figures of eight; leaning whole body forwards and backwards; drawing slow inverted Us; drawing wavy horizontal lines; drawing a large W; drawing descending zigzags; moving arm up and down
Trained	Non-visual	Drawing waves; drawing wavy inverted Us; drawing horizontal lines; drawing an M
Trained	Visual	Drawing circles of various sizes at various heights; drawing large / wavy inverted Us; drawing horizontal lines; drawing diagonal lines; drawing figure of eight

6.3.8 Ferneyhough: no time (at all)

The first thing to notice is that both musically trained and untrained participants showed a relatively limited range of gestural representations (see Table 6-13 below). Played by two guitarists, the piece has got a highly complex structure (many sonic events in a short space of time), making it difficult for the listeners to trace this excerpt with their hand. Nevertheless, it becomes obvious from the speed profiles that both trained and untrained participants were able to pick up a sudden burst that appears approximately 7 seconds into the excerpt. That and another, if less immediate, sonic feature are clearly visible in all speed profiles. And also a rest roughly in the middle of the excerpt—immediately prior to the sudden burst and only a split second long—is represented in the drop of shaking events (see Figure 6-9 below). In line with

previous observations, the visual condition and musical training seem to have led to a more distinct speed of hand movements.

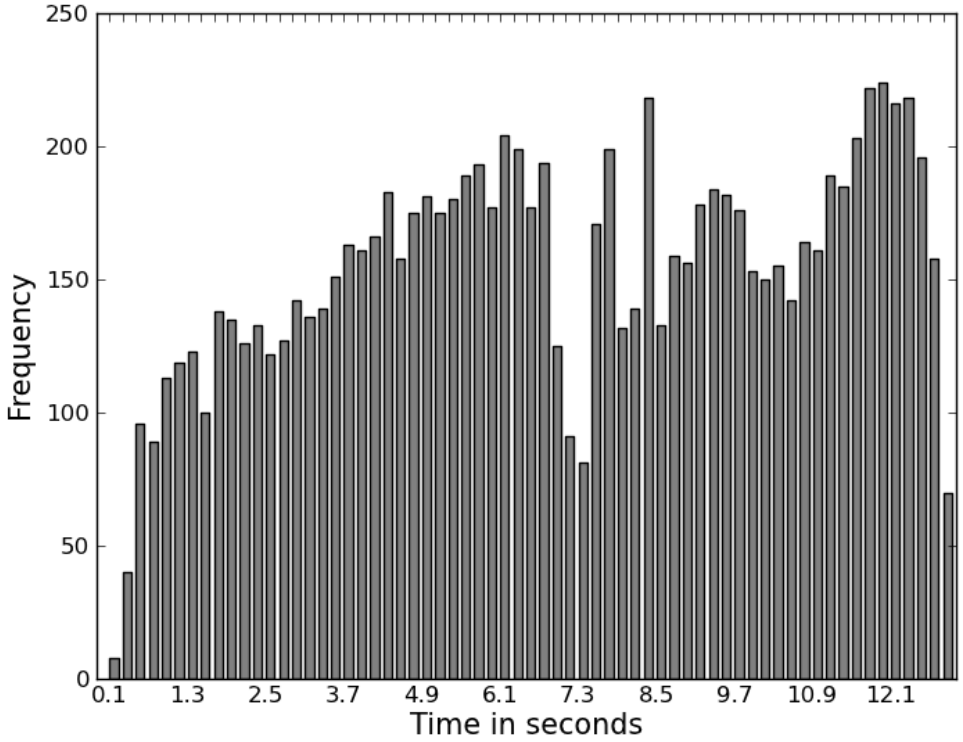


Figure 6-9 Total number of shaking events during the Ferneyhough excerpt

Table 6-13 Gestures in response to Ferneyhough (no time [at all])

Untrained	Non-visual	Shaking controller; holding arm up near shoulder and shaking; whisking motion, drawing circles
Untrained	Visual	Shaking controller (side to side / forwards and backwards); moving controller up and down
Trained	Non-visual	Drawing horizontal line; descending diagonal line; shaking the handset
Trained	Visual	Punching in front of his body; drawing horizontal lines; shaking the handset; drawing circles and squiggly lines

6.3.9 Grupo Fantasma: El Consejo

If Brian Ferneyhough's 'no time (at all)' were to be placed on one side of a spectrum of predictability, this piece by the group Fantasma would probably be found somewhere at the other end. In fact, it contains the same motive twice, repeated from ca. 8 seconds into the

excerpt. What is noticeable is the faster speed of hand movements in the visual compared to the non-visual condition, including more extreme velocities. Besides that, the speed profiles look rather flat and uniform – as does the waveform of this excerpt. However, listening to this piece, it does not sound "flat" at all: its energetic rhythms in combination with the melody by the brass section drives the music forward and make it seem very rich, even though one might argue that, musicologically speaking, there is not much happening. This perceived richness is reflected also in an objective measure extracted again with the MIRtoolbox. 'El Consejo' comes out on top of a table measuring the "event density" by estimating the number of note onsets per second. With 5.9 note onsets per second on average, it is followed by Radiohead (3.4) and Bach's Partita No. 1 (3.3) on the places two and three, respectively. From a perspective of embodied music cognition, it could be suggested that too high event density gives rise to participants' focussing on a larger-scale level, because their understanding (or perhaps better: grasping) of the sonic events is constraint by their sensorimotor abilities. That is to say, if it is not possible to make sense of the sonic events on a micro-level, for instance, because there is too much going on, participants switch to a meso-level, which has been identified as a more appropriate level before (Godøy, 2006). In terms of representational approaches there is a large variety, as can be seen in Table 6-14 below. And being by far the most danceable of all stimuli in the experiment, this is indeed what many participants did to represent the music: they danced.

Table 6-14 Gestures in response to Grupo Fantasma (El Consejo)

Untrained	Non-visual	Small flaps with both arms; drawing waves; outward pushes; shaking controller; drawing Us and spirals; moving wrist up and down whilst moving arm side to side; holding arm up near shoulder and shaking; "clawing" motion
Untrained	Visual	Drawing large, wide Us; side to side waving; shaking controller; drawing waves and walking side to side; shaking arm up and down whilst twisting side to side; drawing horizontal lines then twists wrist with drum fills; pushing arm in and out
Trained	Non-visual	Moving arm side to side whilst shaking wrist; drawing horizontal line; shaking the handset
Trained	Visual	Moving right wrist side to side, arm up and down; drawing horizontal lines; drawing giant circles; drawing M-shaped lines; vertical strokes

6.3.10 Messiaen: Vingt Regards Sur l'Enfant Jésus

This excerpt is characterized by groups of regularly spaced chords (three per group) whose pitch moves upwards within the groups and overall. The peaks in the speed profiles coincide

nicely with what would have probably also been the peaks in the speed profile of the pianist's hand movements, had someone measured Aimard's movements during this recording. They mark the preparatory movements for the first chord of each group, which the pianist had to carry out to transfer his hands quickly and safely from the right end of the keyboard to the left end. Although not many participants chose to play air piano during this excerpt, it might still be the closest they came to mimicking the speed of the performer's hand and arm movements in their representational gestures. And although there are no differences apparent in the speed profiles of musically trained and untrained participants, the overview in Table 6-15 below shows that untrained participants used a greater variety of representational strategies than untrained participants.

Table 6-15 Gestures in response to Messiaen (Vingt Regards Sur l'Enfant Jésus)

Untrained	Non-visual	Drawing (slow) Us; flapping/raising and lowering both arms; up and down wrist motion (whilst moving arm side to side); moving arm up and down in front of body; drawing small Ws in both hands; holding arms above head and waving; shaking controller forwards and backwards; drawing horizontal waves
Untrained	Visual	First half clear forwards and backwards motion, second half side to side; continuous spirals; drawing upwards diagonals; drawing Us (of various sizes); shaking controller forwards and backwards; moving arm up and down; drawing horizontal wave stepping from left to right
Trained	Non-visual	Drawing Us; moving both arms in and out, changing direction each bar; drawing diagonal line from bottom left to top right; moving arm up and down
Trained	Visual	Drawing vertical lines; drawing tall waves; drawing diagonal line from bottom left to top right; moving arm up and down; drawing circles and punching in the air

6.3.11 Mozart: Horn Concerto No. 4 in E-flat major K. 495

The speed profiles of this excerpt are characterized by the extremely regular (conducting) gestures that mark the clear beat of this piece. Comparing musically trained and untrained participants, it could be argued that, for the first time, untrained participants showed in fact clearer beating gestures – with more extreme speed of their hand movement. Inspecting the speed profiles closely, we can also see a rough division into three parts: 0–8 seconds, 8–16 seconds and 16–24 seconds (see Figure 6-10 below). The middle part is slightly elevated indicating faster hand movements. This three-part structure fits nicely with the solo / tutti structure—and hence the loudness contour—of the excerpt: the first and last eight seconds

consist of solo parts for the horn, while in the middle eight seconds the orchestra is playing tutti. The speed profiles of the visual and non-visual conditions revealed once more the previously observed pattern of more variation in the condition with visual feedback.

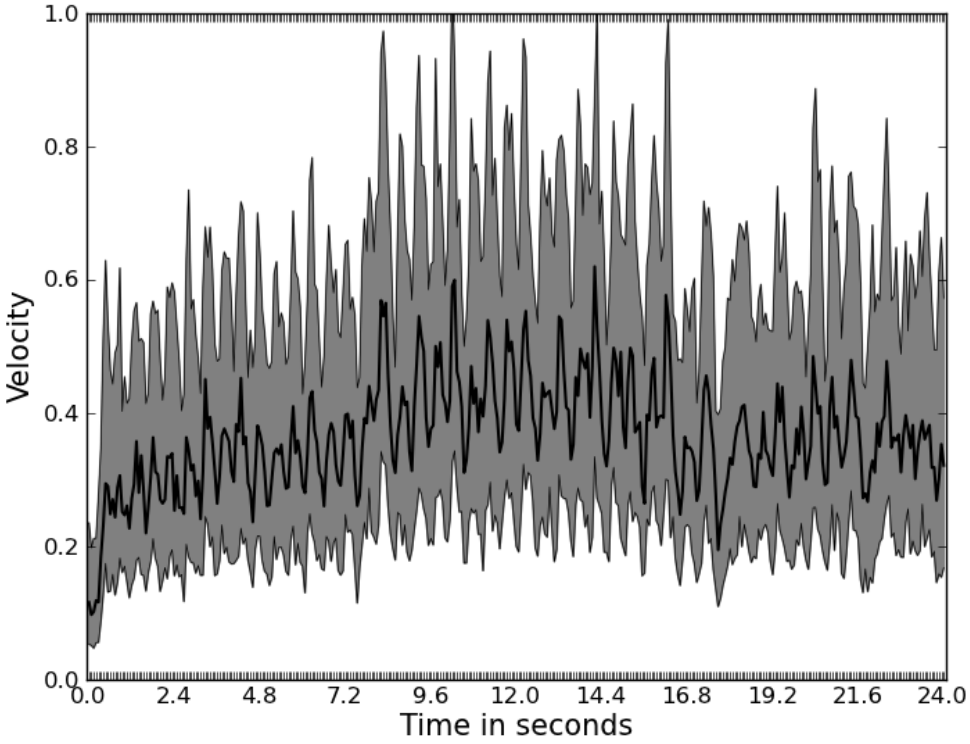


Figure 6-10 Speed profile of Mozart excerpt averaged across all participants and conditions

Table 6-16 Gestures in response to Mozart (Horn Concerto No. 4 in E-flat major K. 495)

Untrained	Non-visual	Waving side to side; drawing Us; outward pushes; up and down wrist motion; drawing figures of eight; shaking controller / drawing circles; drawing waves; swinging arm side to side, forwards and backwards; drawing spirals; drawing Ws
Untrained	Visual	Waving, then figures of eight; outward pushes; alternating up and down motion (two hands on controller) and circles; shaking controller (forwards and backwards); fast and large spirals; shaking wrist up and down whilst moving arm side to side; holding arms above head and bouncing knees
Trained	Non-visual	Drawing Us; moving arm side to side; drawing wavy inverted Us; drawing horizontal lines
Trained	Visual	Drawing juddering vertical lines; moving arm up and down; moving arm side to side; drawing diagonal lines up; drawing circles and figures of eight

6.3.12 Radiohead: The Butcher

The first half of this excerpt is characterised by its clear rhythmic structure in the absence of any pitch. While the synthesized drum beats—including some more elaborate rhythmic figures as the piece unfolds—continue throughout the whole excerpt, the melody starts after 13 seconds. It consists of five pure-tone-like notes, four of which descend slowly, before the last one goes up again. In terms of timbre, these five notes come thus very close to the pure tones of the first part of this experiment but now in the context of a real musical piece. Comparing the musically trained and untrained participants' speed profiles and shaking events (see Figure 6-11 and Figure 6-12 below), a telling difference emerges. While the untrained participants' speed profile looks very uniform—and the distribution of their shaking events suggests that they followed the clear pulse of this excerpt—the trained participants' profile contains very distinct velocity peaks during the first 13 seconds (very similar, in fact, to that of the untrained participants) before their profile becomes more flat. And also the number of shaking events during the first 13 seconds is greater than in the remainder of the excerpt. I mentioned in the previous chapter that Western musical culture is largely based on pitch, and this becomes apparent in the musically trained participants' approaches to represent gesturally the second part of this excerpt. As soon as clear pitch information was available, trained participants stopped representing the rhythmic structure and focused on an accurate representation of the pitches, as we have seen in the part with the pure tones. This tendency was confirmed by watching the individual video clips. Regarding the influence of visual feedback, there was again more variability in the condition with the visualization.

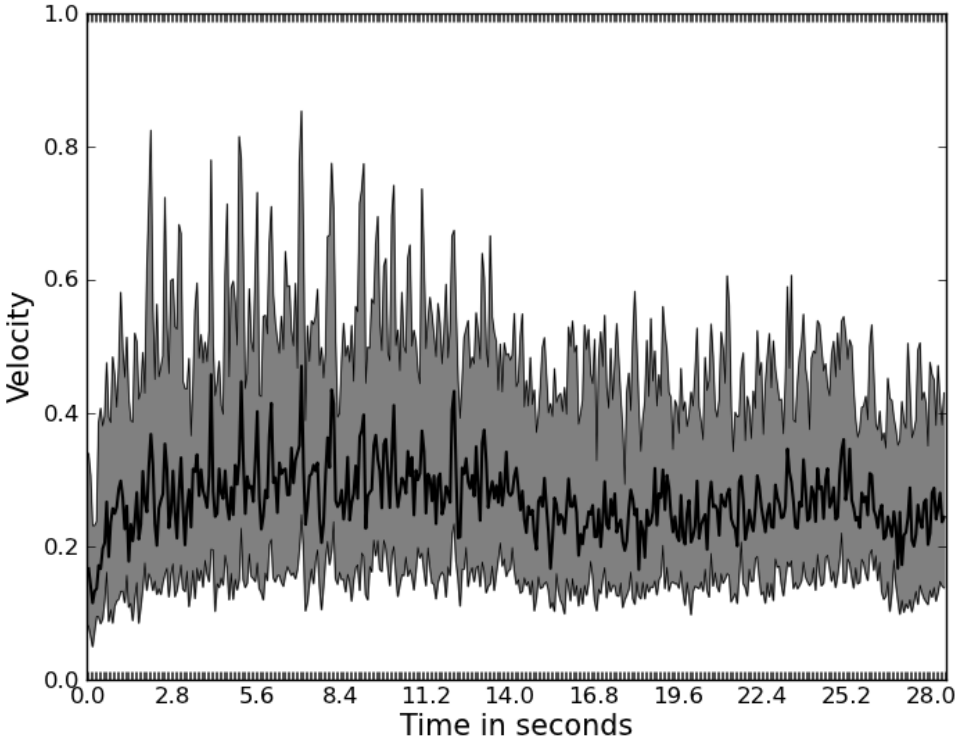


Figure 6-11 Speed profile of Radiohead excerpt averaged across all musically trained participants

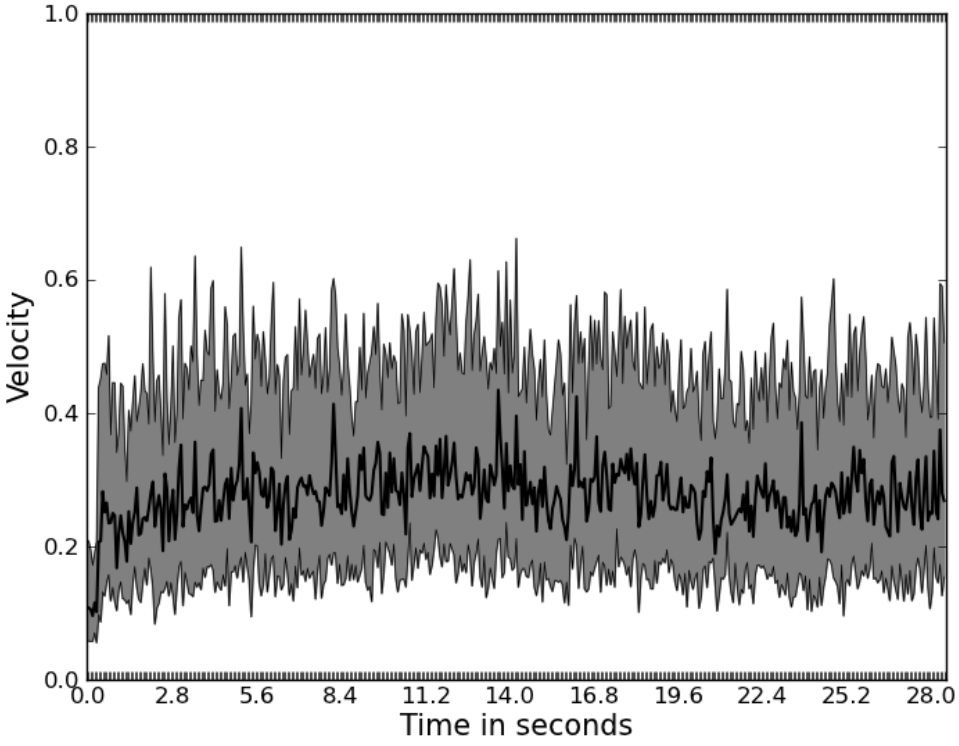


Figure 6-12 Speed profile of Radiohead excerpt averaged across all musically untrained participants

Table 6-17 Gestures in response to Radiohead (The Butcher)

Untrained	Non-visual	Outward pushes; up and down wrist motion/shakes; shaking controller (whilst moving arm up and down / forwards and backwards); drawing large Z; moving arm forwards and backwards; moving arm side to side (whilst shaking wrist); mixture of jabbing the air in front of him and drawing small waves
Untrained	Visual	Large sweeps, horizontal lines, spirals; outward pushes; shaking controller (forwards and backwards); up and down wrist motion; moving controller up and down; drawing Us; drawing inverted Us for one half, then up and down; jabbing the air in front and above him
Trained	Non-visual	Moving both arms up and down / from side to side; drawing horizontal lines; punching in front of his body; shaking the handset
Trained	Visual	Swinging controller forwards and backwards; moving right wrist side to side; drawing horizontal lines; punching in front of his body

6.3.13 Satriani: The Forgotten (Part Two)

Inspecting musically trained and untrained participants' speed profiles, the first observation to make is that the velocity peaks in the group of trained individuals were much more regular and distinct than in the group of untrained participants. Note also that this difference only occurred in the speed profiles, not the distribution of the shaking events, which look very similar in both groups. These peaks coincide with the regular drumbeats in the excerpt. Given the finding that more trained than untrained participants chose to conduct while listening to this excerpt (at least in the visual condition), this pattern of velocity peaks fits into the picture. At the same time, however, both groups show a great diversity of representational strategies, as can be seen in Table 6-18 below. One musically trained participant, whose main activity is composing and his main instrument guitar, chose to stand still with his hand up in the air. In the feedback interview, he reported that he would stand still for the solo violin piece (Bach's Partita for Solo Violin) and the guitar piece (Satriani) because they were solos. Followed up on this remark, he further stated that "for things like that I like to just listen rather than respond to it." Although this is a single opinion, it demonstrates that gestural representation of music might entail more than simply accounting for the sonic features. However, this sort of behaviour might also be linked again to conducting gestures. After all, a conductor often stands motionless during a solo part of an orchestral player and, of course, during a solo part of a violin or piano concerto. Interestingly,

this musically trained participant did not stand still during the solo keyboard or clarinet pieces – perhaps an indication that the definition of 'solo' is linked to the personal main musical activity.

Table 6-18 Gestures in response to Satriani (The Forgotten [Part Two])

Untrained	Non-visual	Drawing figures of eight; moving up and down a diagonal; slow outward pushes; up and down wrist motion (whilst moving arm side to side); side to side motion; drawing Us and swaying from side to side; pushing / shaking controller forwards and backwards; holding arm up near shoulder and shaking
Untrained	Visual	Drawing Us at various speeds (in both hands); shaking controller forwards and backwards; drawing tall, spiralling waves; drawing large circles; pushing arm in and out
Trained	Non-visual	Moving both arms up and down; wiggling controller up and down to the right and left; pushing arm out, then drawing waves; drawing horizontal lines; drawing circles; drawing figures of eight
Trained	Visual	Moving right arm up and down; standing motionless with arm up and left; drawing waves; drawing horizontal lines; drawing circles

6.3.14 Schönberg: Verklärte Nacht

One thing that becomes immediately obvious when inspecting the tempo and shaking profiles is the increase in speed and number of shaking events at the end of this excerpt, which can be found in the non-visual and visual condition, as well as for musically trained and untrained participants. And there is not much doubt about the underlying sonic features that triggered such behaviour. There is an intensification in several musical parameters towards the end of this excerpt, namely in tempo and dynamics, and most prominently, the increase in tremoli: first the second violin, second viola and second cello only, joined in by the first viola two bars later, and finally all members of the sextet (apart from the first viola which has a rest) play tremoli. The perceptual effect is one of high intensity and fast movement, and it seems as though, regardless of musical training, listeners are able to pick up these characteristics easily and turn them into gestures by moving their hands quickly. There is only one nuanced difference between these two groups such that musically trained participants show an even more extreme increase in speed of their hand movements than untrained participants. Similarly, the presence of the visualization led participants to move their hand more quickly (compared to the non-visual condition), thus the visualization can be seen as a contextual factor that further intensifies the gestural representations of music.

Table 6-19 Gestures in response to Schönberg (Verklärte Nacht)

Untrained	Non-visual	Large U-shaped gestures; large slow sweeps; side to side motion and drawing Ws; shaking handset rapidly; upwards diagonal lines
Untrained	Visual	side to side waving; large horizontal arcs; drawing U-shapes gestures; large figures of eight; shaking handset rapidly
Trained	Non-visual	drawing Us; drawing waves; two horizontal lines; drawing figures of eight and circles; shaking handset rapidly
Trained	Visual	U-shapes gestures; moves both arms in and out and up and down; drawing horizontal lines; drawing figures of eight and circles; shaking handset rapidly

6.3.15 Stravinsky: Three Pieces for Solo Clarinet

The pattern observed in musically trained and untrained participants' speed profiles is reminiscent of those of Bach's Partita No. 1 in B-flat major such that the structure of the performance of this piece becomes more obvious in trained than in untrained participants' speed profiles (see Figure 6-13 and Figure 6-14 below). In the trained individuals, there is a three-part structure, with the first part from 0 to 6 seconds, the second part from 6 to 15 seconds, and the third part from 15 seconds till the end. The second part may be further subdivided into 6 to 10, 10 to 11, and 11 to 15 seconds, while the third part may be subdivided into 15 to 17 seconds and 17 seconds till the end. All these time points, which reflect changes in speed of hand movements, coincide with structural aspects of the performance (again, the reader should listen to the recording before referring to the annotated score below in Figure 6-17), showing how carefully musically trained participants shaped their gestural responses in terms of speed. On the other hand, the untrained participants' speed profile looks relatively uniform, only showing a small velocity peak after 15 seconds. But consider the distribution of the shaking events. Here we find this three-part structure in both musically trained and untrained participants (see Figure 6-15 and Figure 6-16 below), suggesting that untrained participants did notice at least some rough structural boundaries and represented them—if not in fine-grained differentiations of speed of hand movement—by shaking (or rather not shaking) the controller. Also the alternative gestural representations (see Table 6.20 below) appear to be quite similar between trained and untrained participants, with the visualization having no major impact on the gestures – apart from the usually observed enhancement of speed and variability.

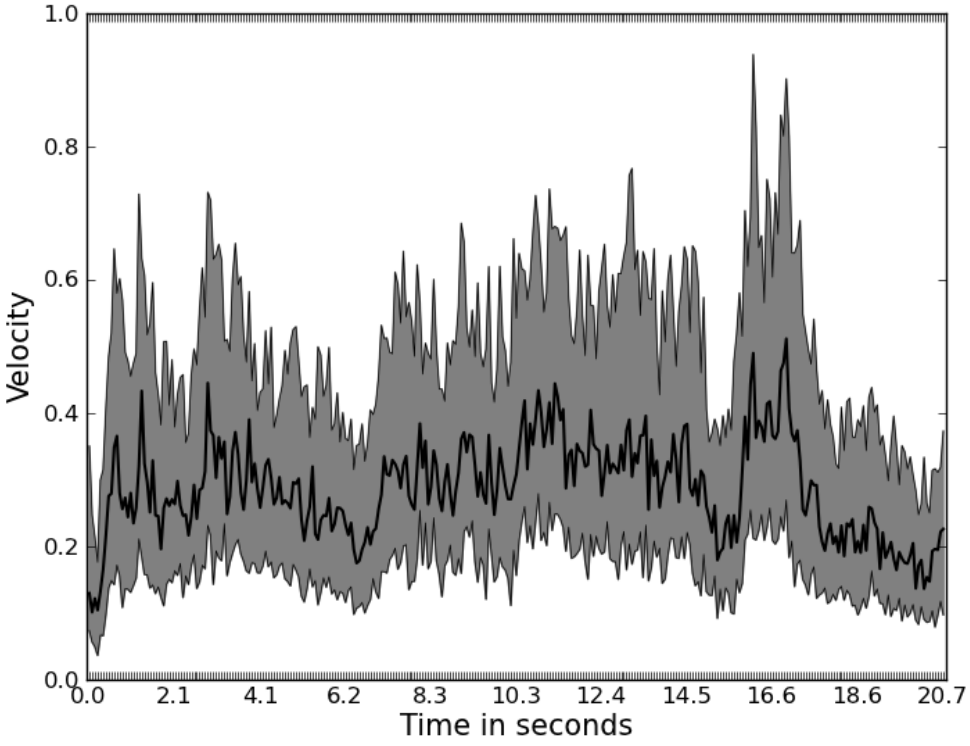


Figure 6-13 Speed profile of Stravinsky excerpt averaged across all musically trained participants

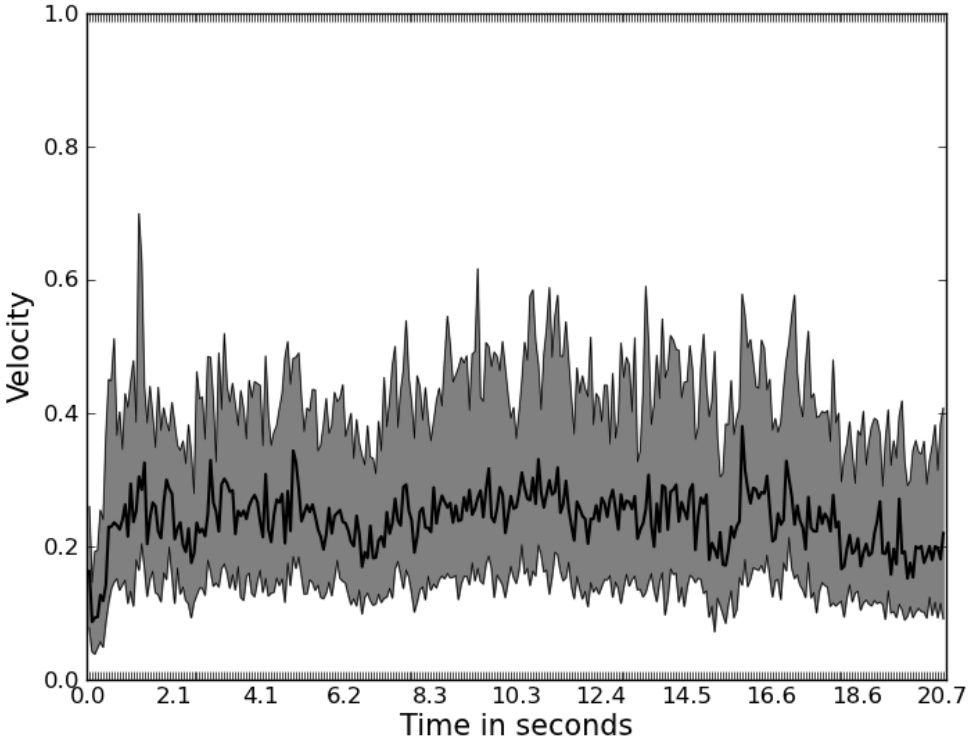


Figure 6-14 Speed profile of Stravinsky excerpt averaged across all musically untrained participants

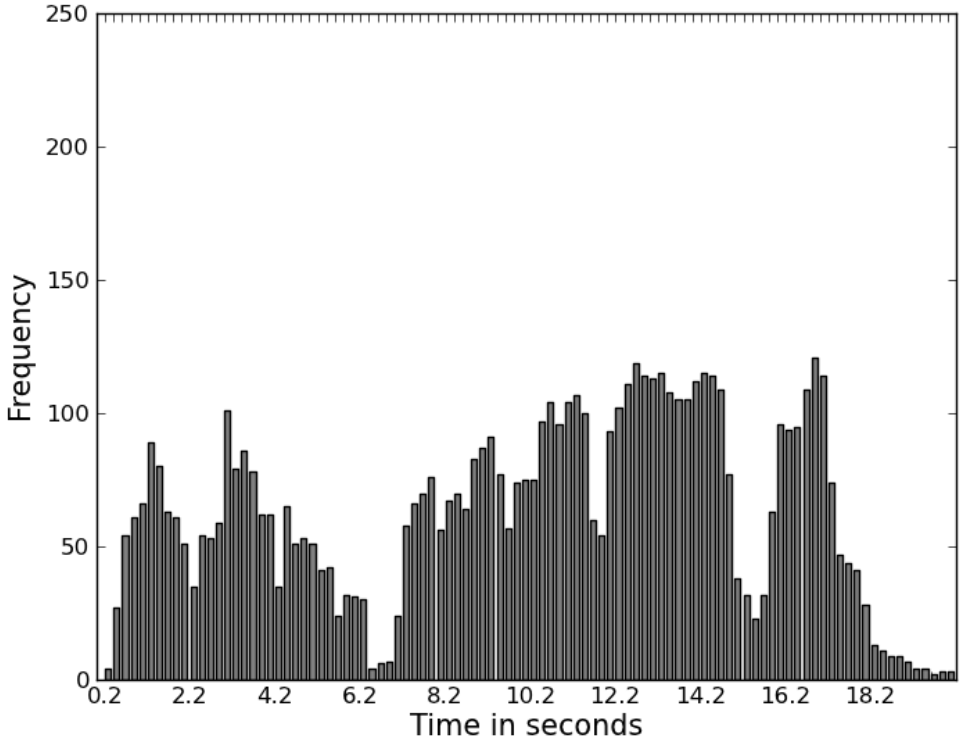


Figure 6-15 Number of shaking events of musically trained participants during the Stravinsky excerpt

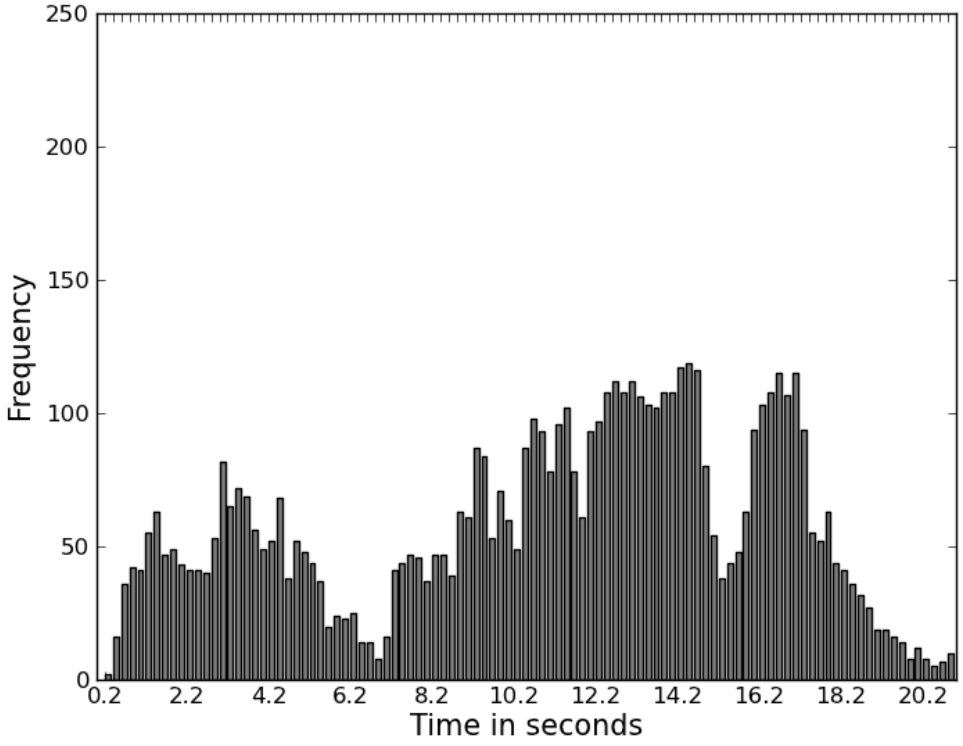


Figure 6-16 Number of shaking events of musically untrained participants during the Stravinsky excerpt

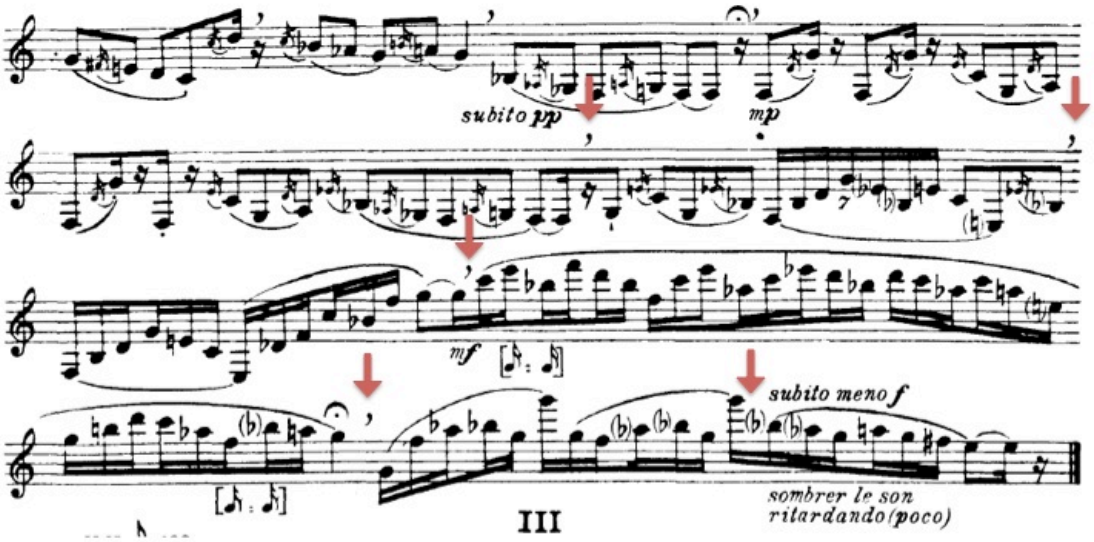


Figure 6-17 Score of Stravinsky excerpt (Three Pieces for Solo Clarinet). Arrows indicate changes in speed of hand movements in musically trained participants (see Figure 6-13).

Table 6-20 Gestures in response to Stravinsky (Three Pieces for Solo Clarinet)

Untrained	Non-visual	Drawing small circles in both hands; shaking controller; drawing Us
Untrained	Visual	Large backwards spirals of the arm, small spirals with wrist; waves at various speeds; drawing small circles in both hands; shaking controller; drawing Us
Trained	Non-visual	Drawing horizontal lines; shaking the handset; drawing circles
Trained	Visual	Drawing circles and moving arm on an upward diagonal; drawing M-shaped lines; rapidly shaking handset

Table 6-21 below provides an overview of the number of different gestures occurring across all musical excerpts, reported separately for trained participants, untrained participants, non-visual condition and non-visual condition. The types of gestures included were ‘up and down’, ‘forwards and backwards’, ‘from side to side’, ‘shaking controller/wrist’, ‘U-shaped gestures’, ‘figures of eight’, ‘zigzags’, ‘lines’, ‘circles’, ‘waves’, ‘spirals’, ‘arcs/inverted U shapes’, ‘outward pushes’, ‘drawing Ws’, ‘drawing Ms’, ‘sweeping’ and ‘no movement’. To be included in the table, a gesture had to occur in at least two different musical excerpts. A detailed overview including the types of gestures can be seen in Appendix 6.7.

The clearest pattern arising is that musically untrained participants used a greater variety of gestures than trained participants in all musical excerpts. The comparison between the visual and non-visual condition is less straightforward. Given that three pieces—the ones by Bach

(violin), Ferneyhough and Stravinsky—led to twice as many different gestures in the visual compared to the non-visual condition and given that such an extreme difference was never observed vice versa, it might be suggested that there is a tendency towards greater gestural variety in the visual condition.

Table 6-21 Overview of different gestures by excerpt

	Number of different gestures observed			
Musical excerpt by composer	UT	T	NV	V
Bach (violin)	10	7	6	12
Bach (keyboard)	11	5	8	7
Berg	10	5	8	9
Boulez	7	5	6	6
Bruckner	9	7	10	9
Carrothers	11	6	11	9
Chopin (Argerich)	12	5	10	11
Chopin (Cortot)	13	6	12	10
Ferneyhough	5	4	3	7
Grupo Fantasma	8	6	7	9
Messiaen	9	7	8	10
Mozart	11	7	13	10
Radiohead	11	6	7	10
Satriani	10	8	10	10
Schönberg	9	8	9	9
Stravinsky	5	4	4	7

Note. Only gestures occurring in at least two excerpts are reported here. UT: untrained participants, T: trained participants, NV: non-visual condition, V: visual condition.

6.4 Consistency of speed of hand movements within participants (across visual and non-visual conditions)

The function of the visualization might be seen as emphasizing or supporting the role of the body as a mediator between the physical world and musical intentions (Leman, 2007). The question then arises whether the availability of visual feedback influenced participants' gestural representation of music. To investigate this question, the speed of participants' hand movements in the non-visual condition was correlated with the speed in the visual condition,

separately for each of the sixteen musical excerpts, resulting in 1024 correlation values. Since the speed values represent time series, Spearman's rank correlation coefficient ρ was used (see previous chapter). Next, a mixed-design ANOVA was run with the within-subjects factor 'music' (varied on sixteen levels) and the between-subjects factors 'training' and 'sex'. This analysis aimed to investigate whether there were any differences between musically trained and untrained, as well as male and female, participants.

Results revealed a main effect of 'music' ($F(15, 900) = 8.78$, $p < .001$, partial $\eta^2 = .13$) and a main effect of 'training' ($F(1, 60) = 6.22$, $p = .015$, partial $\eta^2 = .09$), but no effect of 'sex' ($F(1, 60) < 1$, $p > .50$). It will suffice to present an overview of the mean correlation values in Table 6-22 below, rather than computing all possible pairwise comparisons. The excerpts at the top of the table may be interpreted as examples of musical stimuli that led to more similar gestural responses in terms of speed of hand movement across the non-visual and visual conditions compared to the excerpts at the bottom of the table. Generally, the correlation coefficients are quite small but what is of interest here is not their absolute size but how values of different musical excerpts and/or groups of participants compare with one another. The fact that, on average, the visualization seemed to have had the largest impact—that is, the most dissimilar gestural response in terms of speed compared to the non-visual condition—on the Chopin excerpt performed by Cortot is striking because not only is this excerpt relatively short but it is also texturally and rhythmically less complex than many of the other excerpts.[64] Perhaps this has to do with Cortot's rubato which is very different from modern norms and suggests something animate in motion (Leech-Wilkinson, 2011). Overall, there does not seem to be a singular factor accounting for the size of the correlation coefficients. For instance, there are longer and danceable excerpts at both the top (Schönberg, Carrothers) and at the bottom (Bach [violin], Grupo Fantasma) of the table. To explore this hypothesis, the mean correlation values were correlated with the length of the excerpts, as well as various features extracted from the audio files with the MIRtoolbox.[65] Since none of these correlations reached the Bonferroni-corrected significance threshold of $\alpha = .00227$, it is suggested that what accounts for the

[64] Although the mean correlation coefficient for the same excerpt performed by Argerich is more than twice the size, this difference is not statistically significant.

[65] Features pertaining to dynamics were root mean square (RMS) and low energy; features pertaining to rhythm were attack time, tempo, pulse clarity and event density; features pertaining to timbre were spectral centroid, spectral entropy, roughness, mel frequency cepstral coefficients (MFCC) 2–7 and spectral flux; features pertaining to tonality were key clarity, mode and harmonic change detection function (HCDF); and the feature pertaining to register was salient pitch. For a more detailed description of these features, see for instance Laukka, Eerola, Thingujam, Yamasaki and Beller (2013).

reliability of the representational gestures in terms of speed across the two experimental conditions is largely driven by the interplay of specific sonic characteristics of each individual excerpt and their proneness of invoking visualizations.

Table 6-22 Mean correlations of speed of hand movement between visual and non-visual condition

Musical excerpt by size of mean correlation	Mean	SEM
Schönberg	0.156	0.013
Mozart	0.143	0.013
Carrothers	0.142	0.013
Berg	0.138	0.015
Messiaen	0.131	0.014
Bruckner	0.130	0.014
Boulez	0.124	0.013
Stravinsky	0.094	0.014
Chopin (Argerich)	0.090	0.016
Satriani	0.089	0.013
Ferneyhough	0.072	0.012
Bach (keyboard)	0.070	0.012
Radiohead	0.061	0.013
Grupo Fantasma	0.059	0.012
Bach (violin)	0.057	0.010
Chopin (Cortot)	0.041	0.017
Note. SEM = Standard Error of the Mean		

Next, I will zoom into the main effect of 'training'. To explore this effect further, multiple independent-samples *t*-tests were computed, comparing musically trained and untrained participants. Using a Bonferroni-corrected significance threshold of α = .003125, there were no differences between these groups. However, some comparisons revealed $p < .05$. Musically trained participants showed higher mean values than untrained participants for the excerpt by Berg, $t(62) = 2.03$, $p = .047$, $r = .25$, (trained: $M = .17$, SD = .14; untrained: $M = .11$, SD = .10); for the excerpt by Bruckner, $t(62) = 2.11$, $p = .039$, $r = .26$, (trained: $M = .16$, SD = .13; untrained: $M = .10$, SD = .10); and for the excerpt by Chopin performed by M. Argerich, $t(62) = 2.41$, $p = .019$, $r = .29$, (trained: $M = .13$, SD = .14; untrained: $M = .05$, SD = .12). All these

excerpts belong to the classical canon of composers and it is highly likely that musically trained participants were much more familiar with these pieces than untrained participants. Since many musically trained participants reported that they found the visualization distracting, it is perhaps not surprising that the influence of the visualization on the speed of hand movements seems to have affected untrained more than trained participants: untrained participants were less distracted and more keen on exploring different gestures than trained participants. For 13 out 16 excerpts, trained participants' mean correlation values were larger than those of untrained participants. Only for the excerpts by Grupo Fantasma, Ferneyhough and Messiaen this tendency was reversed, though not statistically significantly. In other words, what we have observed numerous times before in the previous chapters—that musically trained participants' responses are more consistent than those of untrained participants across various auditory stimuli—holds true also when comparing the consistency *within* participants: musically trained individuals are, on average, less prone to change the speed of their hand movements in the presence of a visualization on a screen in front of them than musically untrained participants.

6.5 Shape perception

Thus far I have examined in detail gestural representations of short musical excerpts. In this final section, I will investigate how these movement shapes are *perceived* by observers who are unfamiliar with the experiment and its context. Two male musically trained individuals (undergraduate students in the Music Department at King's College London) were asked to "indicate how well the shape of the music was represented by the arm movement on a scale from 1 (= not at all) to 5 (= very well)." Since it was revealed that the only consistent difference between the visual and non-visual condition was the (variability of the) speed of hand movements, it was decided to present the two observers with video clips from the non-visual condition only. Each observer thus rated the same 1024 video clips (64 participants x 16 musical excerpts) in 4–5 separate sessions over the course of ten days, and received £80 as compensation for their time. Inter-rater reliability was assessed using a two-way mixed, absolute, average-measures intra-class correlation (Hallgren, 2012), revealing that the intra-class correlation was in the good range, ICC = .60. To investigate whether sex and/or musical training of the original participants had an influence on the ratings, data were analysed with independent-samples *t*-tests separately for the two observers and the 16 musical excerpts, as well as across both observers and all musical excerpts.

Results revealed that, overall, there was a significant difference between the perceived movements shapes of musically trained and untrained participants ($t(62) = 3.16$, $p = .002$, $r = .37$), but not between those of male and female participants ($t(62) < 1$, $p > .80$), showing that the ratings for the representation of the shape of the music were higher when participants were musically trained ($M = 3.47$, SD = 0.65) compared to untrained ($M = 3.01$, SD = 0.49). Investigating this effect for each observer separately, it was revealed that the ratings of the first observer showed a marginally significant effect ($t(62) = 1.97$, $p = .054$, $r = .24$), while those of the second observer were highly significant ($t(62) = 3.70$, $p < .001$, $r = .43$).

Looking at each musical excerpt individually, the significance threshold was Bonferroni-corrected to $\alpha = .003125$. Using this stricter criterion, only the Messiaen excerpt reached significance when comparing the first observer's ratings of musically trained and untrained participants (trained: $M = 3.84$, SD = 0.68, untrained: $M = 3.34$, SD = 0.48; $t(62) = 3.40$, $p = .001$, $r = .40$). Two further pieces that came close (i.e. below the threshold of .05) were Berg's Wozzeck (trained: $M = 3.59$, SD = 0.95, untrained: $M = 3.19$, SD = 0.64; $t(62) = 2.01$, $p = .049$, $r = .25$) and Chopin's Prelude played by Alfred Cortot (trained: $M = 3.53$, SD = 0.92, untrained: $M = 2.94$, SD = 0.88; $t(62) = 2.65$, $p = .01$, $r = .32$). Note that for fifteen out of sixteen musical excerpts the ratings of video clips featuring musically trained participants showed higher scores than those featuring untrained participants. Only the movement shapes in response to the Satriani excerpt were rated higher on average—though not statistically significant—by the first observer when the participants were untrained ($M = 3.22$, SD = 0.71) compared to trained ($M = 3.19$, SD = 0.74). Applying the stricter criterion, there were no differences between male and female participants according to the first observer but in two cases the p value was below the threshold of .05. Female participants ($M = 3.31$, SD = 0.78) tended to represent the shape of Bach's Partita No. 3 better than male participants ($M = 2.84$, SD = 0.81), $t(62) = –2.36$, $p = .021$, $r = .29$, while male participants ($M = 3.59$, SD = 0.71) tended to represent Berg's Wozzeck better than female participants ($M = 3.19$, SD = 0.90), $t(62) = 2.01$, $p = .049$, $r = .25$.

Looking at the second observer's data in more detail, it can be seen that for all musical excerpts, he rated the representations by musically trained participants higher on average than those of untrained participants. Four excerpts reached the stricter significance criterion, namely Berg's Wozzeck (trained: $M = 3.72$, SD = 1.17, untrained: $M = 2.69$, SD = 0.90; $t(62) = 3.96$, $p < .001$, $r = .45$), Chopin's Prelude played by Martha Argerich (trained: $M = 3.28$, SD = 0.99,

untrained: M = 2.31, SD = 1.00; $t(62) = 3.90$, $p < .001$, $r = .44$) and by Alfred Cortot (trained: M = 3.31, SD = 1.23, untrained: M = 2.13, SD = 1.01; $t(62) = 4.23$, $p < .001$, $r = .47$), and the excerpt by Stravinsky (trained: M = 3.69, SD = 1.18, untrained: M = 2.75, SD = 0.95; $t(62) = 3.51$, p = .001, r = .41). Furthermore, eight excerpts were below the threshold of .05, and only four excerpts were above. These were the pieces by Bruckner, Grupo Fantasma, Messiaen and Mozart. Finally, there were no significant differences in the ratings of male and female participants.

Note that these results are somewhat limited by the fact that both observers were musically trained, leaving the question open whether untrained observers would have given similar ratings, or whether they would perhaps have preferred the untrained participants' gestures. Nevertheless, these findings suggest the existence of some kind of unspoken consensus among musically trained individuals about what makes a good gestural representation of music. Although both observers were blind to the level of training of the individuals in the video clips, they showed a clear preference for gestures by musically trained participants. On the item level, however, there were also some differences. For instance, the first observer rated the largest difference between trained and untrained participants' gestures for the Messiaen excerpt, which, in turn, was part of the group of excerpts which showed no significant differences according to the second observer (although the mean values showed the same tendency). On the other hand, the ratings of the second observer of the gestural representations of the Satriani excerpt came close to statistical significance (p = .016), whereas this excerpt was the only piece that showed a different tendency of the means in the ratings of the first observer. Note also that all excerpts showing significant differences in the ratings (using the Bonferroni-corrected results) between trained and untrained participants could be grouped together under the umbrella term 'classical music' (first observer: Messiaen; second observer: Berg, Chopin, Stravinsky). This suggests that familiarity with the genre may play an important role here as well. Another striking finding is that the gestural representations of the two short Chopin excerpts—and particularly the recording by Alfred Cortot—whose speed profiles and distribution of shaking events did hardly allow for any distinction between musically trained and untrained participants gave rise to the largest differences of means between trained and untrained participants in both sets of ratings. There is thus something about the quality of trained participants' gestural responses to the Chopin excerpts—possibly the accuracy of the

representation of pitch on the vertical axis—that make them seem adequate or inadequate to the trained observer. Finally, it should be noted that the sex of the participants did not seem to play an important role for the shape ratings. The second observer's ratings came nowhere close to significance, and the ratings of the two excerpts—Bach's Partita for Solo Violin and Berg's Wozzeck—suggesting a sex difference according to the first observer showed opposite effects.

Although the existence of a training effect and the absence of a sex effect are both quite robust, one should bear in mind that the ratings only reflect the opinions of two observers. Future studies focussing on the perception of movement shapes in response to music should therefore test a larger sample of raters.

6.6 Summary and conclusion

To sum up the main findings of this chapter, I have reported differences in participants' perceptions of the experimental tasks. Whereas untrained participants found the representational tasks in general more difficult than trained participants, the latter group reported that they found it more difficult to represent music than pure tones gesturally. In addition, female untrained participants found the tasks more difficult than female trained participants, and the condition with visualization was generally rated as more difficult. In terms of perceived consistency, both groups rated their behaviour more consistent for the pure tones, and trained participants reported that they were more consistent in the condition without visualization.

After this brief comparison between the two types of stimuli, I examined the ways in which participants represented the musical excerpts gesturally, classifying the responses into the four broad categories 'abstract', 'conducting', 'dancing' and 'air playing'. It was revealed that the majority used abstract movement shapes to represent musical features gesturally, and this type of gestural response was the main activity for all musical excerpts. There was a correlation between pulse clarity and type of gestural response, indicating that the clearer the pulse of the excerpt was, the fewer participants showed abstract gestures and the more participants showed dancing, conducting and air playing. Investigating sex effects revealed that women show more abstract gestures and men conduct more. Furthermore, I showed that musically trained participants use more abstract gestures—at least to a few excerpts—and that trained participants conduct more.

A detailed piece-by-piece analysis revealed that trained participants' speed profiles look more distinct in the sense that the variability of the speed of hand movement is more extreme. The same extreme pattern was found more often in the visual compared to the non-visual condition. Even though the gestural variety of trained participants was smaller than those of untrained participants, it became obvious that trained participants pick up on subtleties of the performances and incorporate them in their gestural renderings of the musical excerpts. When the musical figures were less nuanced, untrained participants structured their movements accordingly. Correlating the speed of hand movements across the visual and non-visual conditions, it was revealed that untrained participants explore more in the non-visual condition, while trained participants show higher correlation values, indicating more consistent approaches across the two conditions. Finally, the perception of these movement shapes was tested, and it was found that two musically trained observers prefer the gestural shapes by musically trained participants, even though there was some variation on the item level.

In this chapter, I explored various ways of analysing gestural responses to real musical excerpts, including both quantitative and qualitative approaches. The aim was to sketch a rich picture of individuals' gestural responses, taking into account the variation in the musical excerpts. The findings provided here are in line with what Godøy (2010a, p. 111) had suggested previously:

> "[…] there may be very many alternative choreographies to the one and same musical excerpt, yet there are in most cases particularly salient events and features in the musical sound that the choreographies will tend to mirror and synchronize with."

Just what salience means differs as a function of musical training. The excerpts by Bach (keyboard) and Stravinsky are two examples of how musical training shapes musical cognition. That is not say that musically untrained participants are unable to perceive the structural nuances of the performance but they appear less salient to untrained participants, or—at the very least—are not mirrored in the speed profiles.

The result that trained participants are on the one hand more consistent but on the other hand less variable in their responses fits with findings obtained in my drawing experiment. And it is possible that the same motivation—representing the underlying sonic features—is at the core of other types of gestural representations (conducting, air playing) as well. For musically trained

participants, conducting and air playing in response to a musical excerpt might represent something more generic than directing an orchestra or playing an instrument, respectively. These types of gestures might be used to represent specific sonic features such as the pulse or a melodic line.

It is also intriguing to observe that in some cases where one would expect clear(er) distinctions between trained and untrained participants due to the amelodic, ametric and atonal nature of a piece—think of the Boulez excerpt—the differences turn out to be negligible – both in terms of the observed gestures and quantitative measures of speed of hand movement. What this highlights is that any differences found on a group level need to be interpreted in light of the musical stimulus used, suggesting that bottom-up information (i.e. the sonic features of a musical excerpt) play an important role in gestural responses to music. In fact, this might be seen as evidence for the ontological gap between simple sound stimuli and real musical excerpts. In the following chapter I will bring all main empirical findings of this thesis together in order to discuss and evaluate the outcome of testing cross-modal mappings of sound and music within a paradigm of embodied music cognition.

Chapter 7: General discussion, limitations and conclusion

7.1 Revisiting initial goals and introduction to the final chapter

This section is a reminder of what I aimed to do in this thesis and of how these aims were addressed. The main goal of my thesis work was to investigate the perceived shapes of sound and music within the broad research programme of embodied cognition, exploring a wide range of analytical techniques. In Chapter 1, I provided the general theoretical framework for my thesis, discussing how scholars from different academic backgrounds have challenged the classical, disembodied approach to cognition. I introduced the theory of conceptual metaphor (Lakoff & Johnson, 1980), the motor theory of speech perception (Liberman & Mattingly, 1985) and the theory of mirror neurons (Rizzolatti & Craighero, 2004) in light of embodied cognition, before turning to two accounts within musicology: embodied music cognition (Leman, 2007) and music cognition by shapes (Godøy, 1997). I reviewed the literature on drawings and gestures in response to sound and music, leading to the formulation of the motivation for the empirical studies carried out within my thesis.

In Chapter 2, I introduced the methods, experimental tools and analytical techniques. I highlighted a distinction between products and processes of cognition and showed their relevance for this thesis. Since I applied novel paradigms in both the drawing and gesture experiments, it was important to discuss the broader context of cross-modal mappings and some of the traditional experimental paradigms before introducing the specific tools, software and stimuli used in my experiments. I also highlighted some wider issues of research practices and analysis in (music) psychology, and introduced the main analytical technique of my thesis: non-parametric correlations.

Having carried out the empirical work (see Chapters 3–6), in this final chapter, I discuss critically the findings of my experiments in the context of embodied music cognition and other suitable accounts, drawing also on current theories in the cognitive sciences. I examine the limitations of this thesis work and how these might be addressed and overcome in the future. I provide a discussion of issues to be considered when designing cross-modal studies with sound and music, and suggest potential future paths for embodied music cognition.

7.2 Critical discussion

This section concerns the achievements of this thesis and the ways in which it might contribute to the wider field. One crucial question is the extent to which the findings from my experiments fit within the embodied music cognition paradigm; another concerns whether thinking of music and sound as shapes may provide us with new insights into music perception and cognition. Within the embodied paradigm, the interaction between the body and its environment plays a crucial role for cognition. Asking people to represent sound and music with their bodies—either in drawings or with hand gestures—should then be perceived as a natural way of making sense of auditory phenomena. Since I have provided detailed discussions of the results of my experiments in Chapters 3–6, I will focus here on some individual findings in an attempt to discuss them in the wider context of embodied music cognition as well as other appropriate theories. I will discuss separately the findings from the drawing experiment, the gesture experiment and the exploration of advanced mathematical tools.

7.2.1 Drawings

7.2.1.1 Summary of main findings

I have shown in Chapter 3 that the majority of the participants in the drawing experiment used the height on the tablet to represent pitch (higher up for higher pitches) and the thickness of the stroke to represent loudness (thicker lines for louder sounds). Results revealed two main differences between musically trained and untrained participants that are suggestive of two generic differences in the ways these two groups make sense of sound and music. First, I showed that musically untrained participants' approaches to representing a series of pure tones and two musical excerpts are more varied than those of trained participants. Secondly, comparing musically trained and untrained participants who explicitly reported using height to represent pitch and thickness to represent loudness, I found that trained participants are more accurate than untrained participants. Moreover, there was a difference in the way the two groups approached representing the duration of pitch. Whereas most musically trained participants accounted for the length of a pitch by drawing a horizontal line (longer line for longer-sounding pitch), untrained participants often stopped drawing momentarily when a pitch remained unchanged over time. On the other hand, untrained participants seemed to take greater care than trained participants with the relative size of their visual representations along the x-axis, drawing shorter shapes for shorter sound stimuli.

7.2.1.2 Achievements and theoretical context

To the best of my knowledge, the experiment described in Chapter 3 was the first study investigating individuals' drawings of sound and music in real-time. The finding that pitch is mapped onto the vertical axis is in line with traditional experimental paradigms investigating cross-modal correspondences (Spence, 2011). This is important to note because it shows that I was able to replicate a well-studied effect within this novel paradigm, which might be seen as evidence for the validity of my approach. The main value of this experimental approach, however, lies in its possibility to examine more detailed and varied aspects of cross-modal mappings. For instance, the perception of note duration is one aspect that can be studied *directly yet implicitly* by using a real-time drawing paradigm. It is direct because one can measure the length of a drawn line in response to a note without pitch change; and it is implicit because participants need not be aware of the exact purpose of the experiment, enabling researchers to study spontaneous responses to sound and music. In that sense, an embodied paradigm in which the participants are asked to make sense of sound by engaging in overt body movements has a clear advantage over traditional paradigms. Another advantage is the often-mentioned ecological validity. Having participants respond freely by drawing along with sounds and music gives them significantly more freedom than pressing buttons or choosing from provided response categories, and drawing along with music is something that may well be observed in a natural setting.

The results from the drawing experiments also show that the shapes people associate with very simple sounds can vary considerably, especially among musically untrained participants. Thus, while traditional reaction-time paradigms are restricted to quantitative differences, real-time drawings allow for the assessment of qualitative (as well as quantitative) differences. This has important implications for the interpretation of data. Shorter reaction times of musically trained compared to untrained participants are usually interpreted as faster processing, *assuming that the underlying processing mechanism is the same for both groups*. However, as the results of the drawing experiments indicate, this might not necessarily be the case. Studying some of the "drawing outliers" in more depth might reveal new insights into music cognition (cf. Küssner, 2013). For instance, there have been various cases of participants drawing circular shapes in response to changes in pitch. In a similar vein, Antović, Bennett and Turner (2013) report that a blind child in their study described changing pitch with circular movements. These findings

might suggest that helical models of pitch representation deserve closer attention in future studies.[66]

As outlined at the end of Chapter 1, I have envisaged the drawing responses as 'graphical attunings' or 'sound tracings'. Whereas the latter, more generic term appears unproblematic, the question arises whether participants in the drawing experiment attuned to the musical excerpts. According to Leman (2007, p. 115)

> "[…] attuning aims at addressing higher-level features such as melody, harmony, rhythm, and timbre, or patterns related to expressiveness, affects, and feelings. Attuning […] draws upon the idea that the world is perceived in terms of cues relevant to the subject's action-oriented ontology."

The drawings in response to the two musical excerpts provide evidence that participants have indeed (graphically) attuned to the excerpts (cf. Küssner, 2013). There is a clear representation of the melody in many drawings, but also—and this is a crucial observation—instances of representing extra-musical ideas or feelings (e.g., drawing waves). Interestingly, more musically untrained than trained participants chose to represent elements of expressiveness. This is a notable finding that—together with musically untrained participants' more varied drawing responses in general—raises the question to which extent musical training alters the quality of embodied attuning to music. Apart from more consistent cross-modal mappings, recent evidence suggests that professional musicians, compared to amateur musicians, also show more consistent emotional responses to music (Mikutta, Maissen, Altorfer, Strik, & Koenig, 2014). Is it possible that years of highly specialized body-environment interactions give rise to a bias towards higher-level features at the expense of (more varied) affective responses? Although I have no answer to this question the important point to note is that applying an embodied approach such as real-time drawing has the potential to carry out research with a more finely woven net, enabling empirical musicologists and music psychologists to draw on a rich pool of data.

[66] Antović and colleagues (2013) refer to Fugiel (2011) and Shepard (1982) for more information on helical models of pitch representation.

7.2.2 Gestures

7.2.2.1 Summary of main findings

In Chapter 5, I investigated how individuals map pure tones varied in pitch, loudness and tempo onto the kinaesthetic domain. In line with findings from the drawing experiment, pitch was mapped onto the vertical axis: increase (decrease) in frequency led participants to raise (lower) their arm. Elapsed time was most strongly associated with the x-axis, revealing that most participants went from left to right as the sound stimuli were played. The mapping of loudness was less straightforward. Although loudness was most strongly associated with the y-axis when comparing all three spatial axes, this mapping seems to be a spurious effect driven by sound stimuli concurrently varied in pitch and loudness. When varied in isolation, however, loudness was mapped onto the y-axis—as well as onto the z-axis—increasing loudness giving rise to upward and forward movements with the arm. More consistent cross-modal mappings of increasing-decreasing (compared to decreasing-increasing) pitch and loudness contours were observed. I showed that this bias was mainly driven by untrained participants' smaller correlation coefficients in conditions with decreasing-increasing contours, thus suggesting that musical training annuls such asymmetric mapping effects. It was also revealed that concurrently varied sound features affect the size of pitch–height mappings. The latter were most prominent when tempo and loudness remained unchanged. The visualization seemed to have a detrimental effect on pitch–height mappings, revealing that individuals showed stronger associations between pitch and height in the non-visual condition. While most musically untrained participants used movements with the Wii™ Remote Controller to indicate loudness, shaking the controller faster with increasing loudness, trained participants rarely showed such behaviour. Musical training also influenced the association between tempo and speed of hand movement such that trained participants, but not untrained ones, increased the speed of their hand movements when the tempo of the sound stimuli increased. Direction of pitch and loudness interfered with these tempo-speed mappings, indicating that decrease in either pitch or loudness did not lead to increase in speed of hand movement when the tempo was increasing. Following this detailed analysis of gestural representations of pure tones, I investigated gestural responses to various real musical excerpts in Chapter 6.

It should be noted that musically untrained participants found gesturing along to sounds and music generally more difficult than trained participants, and that the latter group found it easier

to represent pure tones gesturally. The visualization seemed to have distracted the musically trained participants, whereas both groups reported that they were more consistent for the pure tones. The ways in which people chose to represent the musical excerpts included various approaches ranging from abstract shapes aimed to capture sonic features, over conducting and dancing gestures, to air instrument playing. The most commonly observed type of gestural representation was abstract shaping: it was the main activity for all musical excerpts. I showed that, at least to some extent, the representational strategy is related to auditory features. It was revealed that pulse clarity was correlated positively with conducting, dancing and air instrument playing, and negatively with abstract shapes. Another influential factor was sex, indicating that men generally conducted more often than women, who, in turn, showed more abstract gestures. Analysing the musical extracts individually revealed that musically untrained participants showed a wider range of gestures than trained participants – a finding that is in accordance with the drawing experiment. The visualization led to more extreme speed profiles, and musically trained participants showed more distinct speed profiles than untrained participants. Importantly, trained participants' speed profiles revealed that musical training leads to clearer representations of structural aspects of the performances. In line with the observation that musically trained participants found the visualization distracting, untrained participants explored more gestural representations in the visual condition, indicated by lower correlation coefficients across the two conditions in comparison with trained participants. Investigating the perception of movement shapes, I showed that two musically trained observers preferred the gestural representations of the trained participants.

7.2.2.2 Achievements and theoretical context

Although more and more researchers have started to investigate the relationships between music and (free) bodily movements, I am not aware of another study investigating free gestural cross-modal mappings of systematically varied sound features and a range of real musical excerpts. As mentioned in Chapter 1, there are, on the one hand, studies investigating free movement responses to music in which the focus is more on bodily expressions or feelings (Burger, 2013; Van Dyck, 2013) and those investigating free gestural representations of the music, as is the case for my gesture experiment. While it remains to be seen to what extent these approaches measure the same underlying cognitive mechanisms, there is certainly some overlap in the behavioural responses of participants (e.g., dancing).

As with my drawing experiment, the finding that pitch is mapped onto the vertical axis may be seen as support for the validity of my approach of studying cross-modal mappings with gestures. The replication of this effect should merely be seen as the starting point for a more extensive investigation of gestural cross-modal shapes. For instance, the possibility of studying cross-modal mappings of dynamic auditory stimuli with free hand gestures has revealed insights into more consistent and more accurate representations of increasing-decreasing pitch (as well as loudness) contours compared to decreasing-increasing contours, extending findings from previous studies carried out with children (Kohn & Eitan, 2009) or by applying rating and forced-choice matching tasks rather than overt bodily movements (Kohn & Eitan, 2012). What is more, I have attempted to start disentangling some of the interaction effects of cross-modal mappings of *concurrently* varied sound features such as pitch, loudness and tempo. As highlighted in Chapter 5, the ways in which different sound features interact to form a common percept are still poorly understood. Using an embodied gesture paradigm I hope to have provided a starting point for other researchers to study the ways in which actions might influence our perception of concurrently varied sound features. Recent theoretical developments of action-perception couplings in music perception provide an adequate framework in which such interaction effects may be investigated further (Maes, Leman, Palmer, & Wanderley, 2014).

The finding that musically untrained participants (see Chapter 5) and children (Kohn & Eitan, 2009) use muscular energy to represent loudness is yet another example of how embodied approaches may contribute to knowledge in a way that traditional paradigms cannot. To get a clearer idea of how muscular energy is involved in the formation of musical meaning, researchers may use a continuous measurement of muscular activity such as electromyography (EMG). Although EMG has been used to measure facial muscle activation as a means of investigating physiological correlates of music emotions (Lundqvist, Carlsson, Hilmersson, & Juslin, 2009), it may be applied to other parts of the body (chest, arms, hands, legs etc.) to study how musical sounds affect our body and to demonstrate how this bodily involvement might be crucial for making sense of music.

The fact that children and untrained participants are more likely to use muscular energy to represent loudness raises also once more the question of how musical training affects cross-modal mappings of sound. Surely, there is a straightforward relationship between dynamics and muscular energy when playing an instrument such that playing louder requires more energy.

But this relationship does not seem to be musicians' primary association when it comes to representing loudness with bodily gestures. One reason might be that musical training equips individuals with the skill of achieving the widest dynamical contrasts by using only a minimal amount of physical effort. Unlike musically untrained participants then, highly skilled musicians may feel the changing dynamics of a musical piece less strongly in terms of muscular activation. In other words, musicians' body-environment interactions might have been adapted to cope with the extreme physical demands of becoming an expert on a musical instrument (Williamon, 2004).

Another remarkable finding is the effect of the visualization on musically trained and untrained participants' gestures, especially when presented with musical excerpts. Whereas most untrained participants enjoyed exploring the relationship between music, gesture and visualization, musically trained participants appeared to be distracted or even put off. How might this finding be considered in light of Godøy's (1997) account of musical shapes (see also Chapter 1)? In his work, we find a tight coupling of sound, action and image – what he calls the triangular model of cross-modality. According to Godøy (2003, pp. 317-318) this model

> "[…] depicts inextricable relationships between *action, vision* and *sound* in music perception and cognition. The basic assumption of this triangular model is this: *Any sound can be understood as included in an action-trajectory*. Furthermore, this image of sound-production will have visual and motor components in addition to that of the "pure" sound, and I believe images of sound-producing actions can play the role of mediating between the visual and the sonorous, i.e. that actions can translate from the sonic to the visual and, conversely, from the visual to the sonic."

What the finding from my gesture experiment might suggest is that musically trained participants have a well-defined triangular model in their mind. They have a very clear idea of how their actions translate into sound and what the accompanying sensory feedback in the visual domain is. Moreover, the fact that there was a slight delay in the visualization on the screen is perhaps the "worst" that can happen to a highly skilled musician: that their actions do not have the immediate and expected effects specified by the triangular model of cross-modality. Timing is such a crucial aspect of musical performance that even a small delay in any of the senses involved may cause a disruption to the musical experience. Godøy (2003)

discusses the mapping from musical sounds to visual images in terms of internal mimicking of the sound-producing actions, which fits into the broader context of the motor theory of perception (Galantucci et al., 2006). What Godøy calls “motor-mimesis” might then be seen as the central mechanism by which people perceive sound and music as shape. He writes:

> “Motor-mimesis translates from musical sound to visual images by a simulation of sound-producing actions, both of singular sounds and of more complex musical phrases and textures, forming motor programs that re-code and help store musical sound in our minds” (Godøy, 2003, p. 318, original in italics).

What the results of the gesture experiment have shown as well is that sound-producing gestures are not only simulated internally: they sometimes become overt during air instrument playing.

The fact that both musically trained and untrained participants chose to represent music through playing air instrument in my gesture experiment deserves closer attention. In a study specifically investigating air instrument playing, Godøy, Haga and Jensenius (2006b) proposed that the type of gestures involved in air instrument playing may be located on a continuum from novices to experts. The authors report that all participants were able to imitate the coarse gestures of air piano playing such as moving the hands from left to right when the pitch of the musical excerpts was increasing. However, only trained participants showed sensitivity to finer gestural details such as note onsets, dynamics or articulation. Godøy and colleagues regard air instrument playing as a type of ‘motormimetic sketching’. In this view, it is an imitation of the sound-producing gestures underlying the musical excerpts, but on a less detailed (i.e. sketchy) level. Making sense of piano sounds does not require expert skills in piano playing, even though the knowledge of what a piano looks like and how it is played will give a considerable advantage to a listener over those who do not have this knowledge. Engaging in an overt sound-producing action—in the absence of any instrument—is a (prototypical) situation of embodied cognition: we understand sounds through a bodily, goal-directed interaction with the physical environment. Musical training—i.e. years of repeated body-environment interactions on real as well as imagined instruments—mediates the degree of detailed movements. For expert pianists, this even leads to activation of motor areas in the absence of any overt movements (Haueisen & Knösche, 2001) – a finding that is in line with the motor theory of perception (Galantucci et al.,

2006) and auditory-motor coupling (Drost, Rieger, Brass, Gunter, & Prinz, 2005). In other words, when representing music via air instrument playing musically trained and untrained participants engage in the same activity on different levels. Recalling the distinction between products and processes (see Chapter 2), it is probable that the underlying process of air instrument playing is the same for both groups – even if the resulting products (i.e. the visible movement shapes) differ due to the amount of musical training.

7.2.3 Advanced mathematical techniques

7.2.3.1 Summary of main findings

The dataset of the drawing experiment (Chapter 3) was used to explore more advanced analytical techniques in Chapter 4. One of the main objectives was to take into consideration non-linear relationships between the sound features and the drawing characteristics. Another was to take into account the role of time more directly. Applying Gaussian processes for a regression analysis, I demonstrated that using an additional non-linear kernel considerably improves the modelling of the drawing outputs—the values of X, Y and pressure applied to the pen—over using a linear kernel only. The resulting hyperparameters were used in clustering and classification analyses. Of the clustering analyses, spectral clustering and Gaussian mixture models appeared to be more promising than Principal Component Analysis, but conclusions need to be drawn cautiously because it is possible that the lack of clear clusters *per se* played an important role for the resulting pattern.[67] The classification analysis was more revealing, showing that a classifier consisting of both linear and non-linear kernels successfully distinguishes between musically trained and untrained participants, with the SE hyperparameters from the regression analysis as input.

7.2.3.2 Achievements and theoretical context

The most important achievement of the exploration of advanced mathematical tools is arguably the modelling of the data, using a GP regression. It provided the basis—a set of 33 hyperparameters—for the subsequent analyses, including the building of a successful classifier. The results from the regression and classification analyses are sufficient to justify the exploration of more advanced analytical tools, which might be applied to different datasets in different research contexts in the future. At least two aspects are worth mentioning here. First, the regression model took into account both linear as well as non-linear relationships in the

[67] The use of more than two dimensions for the spectral clustering analysis and the PCA might have been another, perhaps more fruitful, option and should be followed up in the future.

data. This is not the norm in music psychology, where the majority of analyses are based on the General Linear Model (GLM). However, there is no reason to believe that all human responses to music should be modelled linearly. Indeed, relevant information might be lost when using linear models, and only more complex models are able to capture the full variety of individuals' responses adequately. Secondly, when studying the process of visualizations of sound and music it is important to consider temporal aspects. While non-parametric correlations achieve this indirectly, including time as a further input into the GP regression provided the opportunity to study effects of time directly. Given that sound and music are unfolding over time, there are still comparatively few established analytical techniques in music psychology that take time into consideration. Such analytical approaches include time series analysis (Schubert, 2004), differential calculus, phase-plane plots and Functional Data Analysis (Levitin et al., 2007; Vines, Nuzzo, & Levitin, 2005a). In the realm of embodied music cognition, Nymoen, Godøy, Jensenius and Torresen (2013) have made a first attempt at comparing systematically various approaches to studying movement responses to sound and music. Using four different approaches—pattern classification, frequentist hypothesis testing (of extracted motion features), non-parametric correlation analysis and Canonical Correlation Analysis—the authors conclude that the choice of approach is dependent on the specific research question, especially with regard to the timescale. It is suggested that researchers should apply various analytical tools if they are interested in capturing gestures on both micro- and meso-levels in terms of time (cf. Godøy, 2010b). Gaussian processes might be added to the arsenal of analytical techniques, though further experiments are crucial to improve existing, and develop novel, tools for analysing time-dependent (non-linear) data in the future.

7.3 Limitations

Having discussed some contributions of my findings to the field of empirical musicology and music psychology, I will now focus on some more general limitations of my approach. As with any empirical study, the methods applied come with several strengths and weaknesses. I have highlighted some of the strengths in the previous section and will now discuss its weaknesses. While some of the limitations are inevitable due to choices that had to be made prior to my studies (e.g., the experimental design), others emerged only during the testing or the analysis phase. Unpacking the weaknesses of my studies will help other researchers to consider their

experimental designs and methods for future studies more carefully. As such, reporting the limitations is a vital process in ensuring the highest research standards are maintained.

7.3.1 Experimental stimuli

The comparison of the findings from the drawing and the gesturing experiments is limited because different sets of pure tones were used. In the drawing experiment, I used discrete tones with different intervals, and the overall length of the auditory stimuli varied. While the rationale was to explore different kinds of variations in the auditory stimuli and their effect on people's visualizations, I decided to be more rigorous in the gesture experiment and chose stimuli with the same length, all of which sounded continuously. Whereas discrete auditory stimuli may be used in a drawing experiment in which participants are able to lift the pen and "jump" across the tablet, such a scenario is impossible in a gesture experiment (without visualization) since any movement with one's hand will result in data being collected. And even though auditory stimuli with different lengths have provided some insights into how musical training shapes the use (drawing) space, it was deemed more important to reduce as many confounding variables as possible when studying the effects of gestural cross-modal mappings of concurrently varied sound features. As I have mentioned in the limitations section of Chapter 5, there is room for further simplification of my pure tone stimuli. The parts with unchanged pitch over time (at the end of an increase and decrease, respectively) could be removed to achieve strictly increasing and decreasing functions of pitch (see also limitations of analysis below), and hence one confounding variable fewer. Parts with unchanged pitch were kept in the gesture experiment because the way in which musically trained and untrained participants represented them in the drawing experiment pointed to an important difference in the representation of unchanged pitch over time.

7.3.2 Experimental setup

There is at least one general issue with the applied soft- and hardware that should be addressed in future studies: its temporal resolution. As reported in Chapter 5, the time lag in the gesture experiment was partly due to the compromise of having a smooth visualization on the screen in front of the participants. Thus in future studies without real-time visualizations, the time lag may be further minimized to ensure the sampling of the gesture data and the sound features are synchronized as much as possible. A similar issue was involved in the drawing experiment. When participants moved the pen very quickly across the graphics tablet, the

resulting trace consisted of a dashed, instead of a continuous, line. This was probably a shortcoming of the software, which collected data at 48 frames per second, rather than a shortcoming of the tablet, which has refresh rate of 200 points per second. However, since it is unclear how often the tablet reached its maximum refresh rate, further testing is needed to arrive at a solid conclusion.

7.3.3 Participants

The musically trained participants recruited for the drawing experiment included keyboard players, wind/brass players, string players, composers and singers, while only the first four categories were recruited for the gesture experiment. Since I aimed to be more rigorous by having a group of musicians balanced by sex and main musical activity in the gesture experiment, I excluded the group of singers based entirely on pragmatic reasons.[68] The basic idea behind controlling the main musical activity in the gesture experiment more strictly was to obtain a heterogeneous sample of musicians, trying to avoid the problem of an underrepresented group of musicians (such as singers in the sample of the drawing experiment) causing outliers in the dataset. Of course, the fact that I did not include any singers (or percussionists, for that matter) in the gesture experiment limits the generalizability of the findings. In future studies, researchers should therefore aim to include an even broader range of musically trained participants, including perhaps some musicians who are unable to read music to study the effects of the ability of score-reading on cross-modal mappings (which might be an important mediating variable).

The criteria for musically trained and untrained participants were chosen arbitrarily due to the lack of agreement in the community of music psychologists as to what counts as "musically trained" and "musically untrained". The recent arrival of the Goldsmiths Musical Sophistication Index (Gold-MSI) may present a better way of assessing musicality in future studies, enabling researchers to enter 'musicality' as a continuous variable into their analysis (Müllensiefen, Gingras, Musil, & Stewart, 2014).

The findings are also limited since I did not assess participants' general drawing or gesturing skills prior to the experiment in a non-musical context. Such skills (or the lack thereof) might have influenced the results and should be controlled for in future experiments.

[68] Since there had been only very few responses from singers—all of whom were female—to the call for participants for the drawing experiment, I wanted to avoid difficulties in the recruiting process for the gesture experiment.

7.3.4 Analysis

One problem with applying non-parametric correlations to my datasets is that they do not take into account time directly.[69] They are therefore insensitive to micro-scale variations in the drawing and gestural responses to sound and music. The use of cross-correlations might have provided further insights into the accuracy with which (groups of) participants synchronized with the sound stimuli (for a recent example of the use of cross-correlations in music research see Himberg & Thompson, 2011). In such an analysis, one time series is gradually shifted against another to identify the position with the largest correlation coefficient in a given time window (e.g., one second).

Another problem of using non-parametric correlations for the present datasets may be the existence of rank ties, resulting in an underestimation of the size of the correlation coefficients. Since the sound stimuli contained passages with unchanged pitch over time, it is likely that there were tied ranks present. To minimize the effect of rank ties in future studies, a non-parametric correlation coefficients adjusting for ties (e.g., Kendall's τ) may be used. Alternatively, researchers could use auditory stimuli whose sound features of interest are constantly changing. However, in order to investigate some specific research questions, it might be necessary to keep passages with sound features unchanged over time and look for an alternative analysis tool instead. These and other decisions researchers have to make can have a great impact on the experiment to be carried out. In the following section, I will discuss some of the choices researchers might face when studying cross-modal shapes of sound and music.

7.4 Future considerations

When studying cross-modal mappings of auditory stimuli, the outcome will depend to a large extent on the specifics of the experiment such as the choice of stimuli, the experimental setting and the instruction given to participants. By discussing some of the issues involved I hope to provide a helpful, if by no means exhaustive, overview for researchers who wish to carry out experiments on cross-modal mappings of sound and music.

[69] Drawing/gesture responses occurring too early or too late result in a shift that may affect the size of the correlation coefficient. In that sense, the correlation analysis is indirectly sensitive to participants drawing or gesturing "out of synchrony".

7.4.1 Music vs. sound

This dichotomy is not specific to the study of cross-modal mappings but can be found in any other field in which researchers have to face the problem of the whole vs. its parts. Unlike psychoacousticians who exclusively work with highly controlled, synthesized sound stimuli, music researchers are particularly concerned with the unravelling of cross-modal mappings of real music, and a broadly accepted way to study these is to investigate music's constituent parts such as pitch, loudness or timbre. The problem with studying characteristics of musical sounds in isolation (e.g., change in pitch) is the creation of an ontological gap: we can never be sure that findings from studies using synthesized pure tones in order to investigate cross-modal mappings of pitch apply equally to situations in which we listen to the changing pitches of a musical performance. There are too many other factors involved in the latter that render generalizations problematic. On the other hand, the choice of real musical excerpts as experimental stimuli gives rise to a number of confounding variables since it is unavoidable that other musical qualities such as dynamics or articulation, or at least timbral qualities, will be co-varied with pitch. This makes it difficult, if not impossible, to study causal links. I therefore suggest that researchers should, whenever possible, include both types of stimuli in their experiments (e.g., Eitan & Timmers, 2010; Küssner & Leech-Wilkinson, 2014) in order to get a better idea of the extent to which findings from highly controlled psychoacoustical stimuli hold true for musical excerpts, but also to what extent findings from studies using musical excerpts can be replicated by manipulating the musical sound feature of interest in isolation.

7.4.2 Pure tones vs. synthesized musical sounds

A further option—which might be seen as an attempt to bridge the aforementioned ontological gap—may be to synthesize auditory stimuli that resemble real musical sounds, as can be achieved by the use of Musical Instrument Digital Interface (MIDI). For instance, Eitan and Granot (2006) used synthesized piano sounds to study cross-modal mappings of various musical features such as pitch, dynamics and speed. While such an approach has the advantage of presenting participants with more "natural" stimuli in comparison with pure tones, it leaves open the question of whether the same results would have been obtained with, say, guitar or trombone sounds. Any step towards a more ecological musical stimulus comes at the cost of introducing new variables that need to be controlled in an experiment aiming to uncover causal relationships. And while advances in music synthesizing software allow for features such

as “expression” or rubato to be switched on, the gap to human musical performance—though gradually shrinking—is still very much audible. When designing an experiment, researchers thus need to consider carefully the advantages and disadvantages of employing MIDI-based sound stimuli.

7.4.3 Isolated vs. concurrently varied

Although arguably being simplistic compared to real musical excerpts, pure tones can be synthesized with varying degrees of complexity. However, most studies so far—at least those concerned with music cognition—have included pure tones whose features were manipulated in isolation. There is scope for many more studies using controlled pure tones (or more naturally sounding ones, such as MIDI sounds) whose features are concurrently varied in a systematic manner (for recent examples see Eitan & Granot, 2011; Küssner, Tidhar, Prior, & Leech-Wilkinson, 2014). As mentioned above, in most cases music consists of the dynamic co-variation of several musical parameters. These co-variations may, to some extent, be recreated in the synthesis of pure tones, achieving more ecologically valid stimuli while keeping possible confounding variables at a minimum.

7.4.4 Musical excerpts vs. whole compositions

In almost all studies investigating cross-modal mappings of music, researchers have used relatively short excerpts from longer musical compositions. One notable exception is the study by Tan and Kelly (2004) in which musically trained and untrained participants were asked to depict graphically whole musical compositions. The authors raised the important issue that short musical excerpts, when taken out of its context within a piece, may lead to different visualizations and cross-modal mappings. I agree that the context plays an important role—perhaps not so much for basic mappings of sound features such as pitch and loudness—but for more elaborate (visual) representations of music that take into account instrumentation, texture, harmony, repetition and so on. Even though it is probably hardly ever feasible to include recordings of whole symphonies in an experiment, there is scope for studying the effects of shorter, yet complete musical compositions on people’s visual representations, and compare them with responses to shorter, out-of-context, musical excerpts.

7.4.5 Live vs. recorded

The “liveness” aspect of a musical performance has recently attracted increased attention, relating to topics such as audience engagement (Sloboda, 2013), performer-audience interaction (Whitney, 2013) and emotional responses in the listener (Egermann, Pearce, Wiggins, & McAdams, 2013). Being physically present at a concert might indeed give rise to quite different visualizations and representations of music than when listening to a recording in a laboratory setting. While one pioneering study (Hooper & Powell, 1970) revealed that pictorial representations of music in a live context led to more elaborate responses, there is certainly scope for more research of that kind. It should make intuitive sense that the visual presence of musicians, their body movements and instruments, as well as the presence of other audience members, may lead one to associate different shapes from those experienced during solitary listening.

7.4.6 Active vs. passive listening

Apart from the “liveness” aspect, there is evidence that individuals’ motor activity during listening affects their cross-modal mappings of music. For instance, it has been suggested that the motor behaviour during listening influences children’s visual representations of musical excerpts (Fung & Gromko, 2001). A group of children allowed to move with props or in sand while listening to the music produced visualizations that included more detailed representations of rhythm, beat and groupings of notes compared to a group of children who were asked to sit still. Hooper and Powell (1970) reported similar results for adults who were accompanying musical excerpts rhythmically: they showed more elaborate visual representations than groups of adults who were either told to listen carefully or for enjoyment. It is therefore plausible that the overt engagement in motor activities shifts our attention to rhythmic properties, which—during suppression of motor activity—might not have reached the threshold of consciousness.

7.4.7 Nature of task: spontaneous – mandatory – elaborate

The nature of the task, including experimental stimuli but also the exact wording of the instruction and participants’ interpretation thereof, determines what is being assessed during an experiment. As Rusconi and colleagues (2006) pointed out in a critique of some classic psychophysical experiments investigating pitch–height mappings, there is a crucial difference between spontaneous and mandatory mappings. Spontaneous cross-modal mappings are seen as occurring automatically, independent of the context and possibly without our being aware of

it, whereas mandatory mappings require our full consciousness and deliberate action. At best, the latter are used to refine some finding well supported by empirical evidence; at worst, they introduce highly artificial categories to an experiment, leading to meaningless responses.

Besides mandatory cross-modal mappings, which are restrained by a limited choice of response categories, there is also what might be called elaborate responses. Whether spontaneous or not,[70] they constitute free, unrestricted responses to some stimulus, for instance by drawing a sound or a piece of music. While such paradigms provide richer data than, for instance, reaction-time measures, they often also require some unstandardized analysis procedure, complicating the comparisons between studies. Whichever paradigm researchers apply after weighing advantages and disadvantages, it is important that they are aware of the kind(s) of cross-modal mappings they are measuring.

7.4.8 Synchronous vs. asynchronous

Thanks to the availability of adequate experimental tools such as electronic graphics tablets (Küssner et al., 2011), researchers investigating visualizations of sound and music have been able to study 'sound tracings' (Godøy et al., 2006a), or the process of visualizing sounds (see Chapter 3). This approach might offer an additional, different angle to what has been the focus in most previous studies, i.e. the final product of the sound/music visualization. Asking participants to draw or gesture along with the sound enables researchers to study not only how they map them cross-modally but also how well they are in synchrony with various sound/musical features. Particularly for researchers regarding perception as an active process based on action-perception cycles, it appears to be an overdue step to pay attention to the action that creates a certain cross-modal mapping. Having provided some (hopefully) useful considerations for future experiment, I will return once more to my empirical findings and attempt to situate them in a broader theoretical context.

7.5 Broadening the perspective

It is worthwhile considering how the results of my experiments may be explained within recent theories in cognitive science. One theory that has attracted heightened attention over the past few years is the theory of hierarchical predictive processing or predictive coding (Clark, 2013; Friston, 2005; Rao & Ballard, 1999). In this view, perception is a generative, hierarchical,

[70] They may be regarded as spontaneous if the presented stimulus is novel for the participant, or as unspontaneous if participants are familiar—due to prior exposure in the experiment or elsewhere—with the stimulus.

bidirectional process. On the one hand, there are top-down predictions and expectations about sensory states, and on the other hand, there is bottom-up sensory information from the outside world. The crucial point is that only prediction errors are passed forward to higher cognitive levels. That is, at each level of this hierarchical model, the brain's model of the outside world is compared to sensory input from the environment. Since the brain's model is based on probabilities—based on prior experiences of an agent—there will be some error variance involved in the model, i.e. a mismatch with sensory input from the "real world". And it is only this part of the sensory input—the error signal, not the whole representation of the outside world—that is processed further in higher cognitive levels.[71]

How does this apply to music cognition? Listening to one's favourite record would mean that the brain processes only very little of the physical signal of the sound on a higher cognitive level because it has already acquired a very accurate model of the unfolding sonic events during hundreds of times of repeated exposure. When listening to a new composition, on the other hand, there would be more processing of the actual physical signal—in the form of error signals—because although the listener might be familiar with the genre, instrumentation, performance style etc., the brain would not have had the chance to form an accurate model yet.

Explaining the results of my experiments in terms of hierarchical predictive processing is best illustrated with the effects of musical training. For example, the finding that musically trained compared to untrained participants were more accurate in representing pitch on the vertical axis could be explained as an effect of musical training on the power to predict auditory information and plan and execute motor programs accordingly. Thus, musically trained participants would exhibit fewer prediction errors in the coordination of auditory and proprioceptive input because their brains already have a sufficiently accurate model—on a higher-order, abstract level—of the sensory interplay between sounds and movement. In an attempt to refine Clark's (2013) account, Moore (2012) suggests that the most fundamental activity brains engage in should not be thought of as predicting errors but rather as testing actions in an environment. This might be interpreted as an effort to bring hierarchical predictive processing in line with the embodied cognition research programme. In such a view then, musically untrained participants have experienced significantly fewer instances of integrating sound and movement, an instance of

[71] I am slightly simplifying the model for the sake of the argument. Clark (2013) states that there are both error units and representational units at each level of the hierarchical processing model.

which therefore requires more processing effort by their brains, and is also prone to more mistakes.

Thus far, there have not been any systematic attempts of illuminating music perception and cognition from a predictive coding perspective. One of the first studies explaining a musical phenomenon—musical hallucinations in a single-case study—in terms of predictive coding has been carried out by Kumar and colleagues (2014). And also Schaefer, Overy and Nelson (2013)—in a response to Clark (2013)—emphasize the role prior experience plays for musical meaning formation and behaviour, seeing great potential for fitting empirical results from music cognition studies within the framework of hierarchical predictive processing if the role of musical emotions is considered seriously.[72] Indeed, the trend seems to be going towards unifying theories that are able to explain all aspects of human cognition and behaviour, such as the free-energy principle (Friston, 2010). It is true that visually or kinaesthetically representing the shapes of sound and music—whether considered from an embodied perspective, as cognition by shapes, or otherwise—cannot be meaningfully separated from other aspects of musical behaviour. The drawings and gestures by my participants inevitably involved affective responses of some extent, even though they were not the focus of the present investigation. As Van Dyck (2013, p. 136) states in the conclusion section of her thesis on music and movement:

> "Evidence that music perception incorporates tight linkages with other systems, such as the motor and emotion system was unveiled. […] [M]usic perception is considered as multimodal, involving auditory, motor, as well as emotional processes."

The challenge will thus be to incorporate all aspects of musical behaviour and cognition into a grand theory. While the theory of hierarchical predictive coding might be an appropriate place to start, the success of such an endeavour will crucially depend on taking on board insights gained from the full breadth of theories related to music cognition and perception, most notably perhaps embodied cognition approaches. Although the ultimate goal of a grand theory of music cognition will be to replace more specialized accounts, there is still plenty of scope for the development of the latter. After all, embodied music cognition has only just been formalized (Leman, 2007), and music cognition scholars will most likely engage in more fine tunings over the coming decade(s) (Schiavio & Menin, 2013). In fact, accounts of embodied music cognition may only be at the

[72] A more general account of emotions viewed from a predictive coding perspective has recently been put forward by Seth (2013).

start of a long journey—perhaps encompassing extensions to cross-cultural approaches—before they are replaced—if they are replaced—by a grand theory of human musical behaviour à la predictive coding.

As recently argued by Leman (2013), there is a need to extend approaches of empirical musicology beyond cultural borders. Cross-cultural comparisons of musicians' drawings (Athanasopoulos & Moran, 2013) as well as studies investigating musicians' gestures in a non-Western context such as Karnatak music (Pearson, 2013) have provided a glimpse of a still poorly understood research area, even though identifying regions untouched by Western culture is an increasingly difficult endeavour (Huron, 2008). The findings from Athanasopoulos and Moran's study suggest that there might be significant differences in the way people from different cultural backgrounds make sense of sound and music in terms of shape. Indeed, even one of the best-studied effects of cross-modal mappings of sound in Western societies probably requires reconsideration in a cross-cultural context. The observation that higher pitch leads to higher elevation (of one's arm or upwards movements on a graphics tablet) does not necessarily hold for musicians from Central Africa. Drawing on research by Kubik (1983), Ashley (2004) reports that African musicians would lower their hands with increasing pitch during the performance of a song, relating high pitch to the adjective "small" and indicating the (decreasing) size of an object with the distance of their hands to the ground (cf. Eitan & Timmers, 2010). This example reminds us that we need to be very cautious when generalizing even the seemingly most robust results from experiments with Western participants.

7.6 Conclusion

The goal of my thesis work was to investigate cross-modal mappings of sound and music within an embodied paradigm, conceptualizing music cognition as cognition by cross-modal sound shapes. I was able to demonstrate that visual and kinaesthetic representations of sound and music may be studied by asking individuals to draw or gesture along with auditory stimuli, providing insights into the process and the product of cross-modal mappings. The richness of my datasets allows for both qualitative and quantitative approaches, and extends previous studies that either investigated (children's) drawings and gestures mostly qualitatively or examined cross-modal correspondences mostly quantitatively in traditional experimental paradigms. The significant differences found between musically trained and untrained participants raise some important questions for music education and the way musicians are

trained in conservatoires. While some differences in the way musicians perceive and make sense of music are inevitable, we might want to ask whether those observed in this thesis—the relatively constrained and unimaginative representations of sound shapes—are all desirable and fit with our expectations of modern musicianship.

On a methodological level, I was able to show that the use of an electronic graphics tablet, as well as Microsoft® Kinect™ and Nintendo® Wii™ Remote Controller, has great potential for studying aspects of embodied music cognition. Clever use of new technologies—together with the development of appropriate analytical tools—may pave the way for genuinely new experimental protocols that allow music psychologists and empirical musicologists to tackle a broader spectrum of research questions related to musical behaviour. This should follow from situating cognitive processes not only in the brain but the whole body, which requires experimental settings in which individuals are able to engage in free bodily movements. Surely, the brain plays a pivotal role in all cognitive processes—and thus needs to be studied closely—but seen from an embodied perspective in which the interaction between the *whole* body and its physical environment is crucial for cognition, it is not acceptable to reduce cognition to neurophysiological processes alone. This would mean going back to Claxton's (1980) notion of the Frankensteinian preparation; it would mean going back to a world that—and the findings of my thesis hopefully emphasize this point—no longer seems an adequate place to study (musical) cognition.

In conclusion, there is thus perhaps only one thing that is clear. If multimodal perception of music is indeed essentially based on our bodies interacting with the physical environment, using appropriate body-centred experimental paradigms and analysis techniques to investigate cross-modal mappings of music will be a necessary step of our endeavour to capture the full breadth of human musical experience.

Bibliography

Antović, M., Bennett, A., & Turner, M. (2013). Running in circles or moving along lines: Conceptualization of musical elements in sighted and blind children. *Musicae Scientiae, 17*(2), 229-245. doi: 10.1177/1029864913481470

Ashley, R. (2004). *Musical pitch space across modalities: Spatial and other mappings through language and culture.* Paper presented at the 8th International Conference on Music Perception and Cognition, Evanston, IL.

Athanasopoulos, G., & Moran, N. (2013). Cross-cultural representations of musical shape. *Empirical Musicology Review, 8*(3-4), 185-199.

Athanasopoulos, G., Moran, N., & Frith, S. (2011). *Literacy makes a difference: A cross-cultural study on the graphic representation of music by communities in the United Kingdom, Japan and Papua New Guinea*. Paper presented at the Society for Music Perception and Cognition Conference, Rochester, NY.

Aziz-Zadeh, L., Iacoboni, M., Zaidel, E., Wilson, S., & Mazziotta, J. (2004). Left hemisphere motor facilitation in response to manual action sounds. *European Journal of Neuroscience, 19*(9), 2609-2612. doi: 10.1111/j.0953-816X.2004.03348.x

Bamberger, J. (1980). Cognitive structuring in the apprehension and description of simple rhythms. *Archives de Psychologie, 48*, 171-199.

Bamberger, J. (1982). Revisiting children's drawings of simple rhythms: A function for reflection-in-action. In S. Strauss (Ed.), *U-shaped behavioral growth* (pp. 191-226). New York: Academic Press.

Barrett, M. S. (2000). Windows, mirrors and reflections: A case study of adult constructions of children's musical thinking. *Bulletin of the Council for Research in Music Education, 145*, 43-61.

Barrett, M. S. (2005). Representation, cognition, and communication: Invented notation in children's musical communication. In D. Miell, R. MacDonald & D. Hargreaves (Eds.), *Musical communication*. Oxford: Oxford University Press.

Bastéa-Forte, M. (2011). Cellosoft JTablet Plugin. Retrieved 15th July 2011, from http://jtablet.cellosoft.com

Becking, G. (1928). *Der musikalische Rhythmus als Erkenntnisquelle*. Augsburg: B. Filser.

Bell-Berti, F., Raphael, L. J., Pisoni, D. B., & Sawusch, J. R. (1979). Some relationships between speech production and perception. *Phonetica, 36*(6), 373-383. doi: 10.1159/000259974

Ben-Artzi, E., & Marks, L. E. (1995). Visual-auditory interaction in speeded classification: Role of stimulus difference. *Perception & Psychophysics, 57*(8), 1151-1162. doi: 10.3758/BF03208371

Bernstein, I. H., & Edelstein, B. A. (1971). Effects of some variations in auditory input upon visual choice reaction time. *Journal of Experimental Psychology, 87*(2), 241-247. doi: 10.1037/h0030524

Boersma, P., & Weenink, D. (2012). Praat: Doing phonetics by computer (Version 5.3.15). Retrieved from http://www.praat.org/

Bond, B., & Stevens, S. (1969). Cross-modality matching of brightness to loudness by 5-year-olds. *Attention, Perception, & Psychophysics, 6*(6), 337-339. doi: 10.3758/BF03212787

Brass, M., & Heyes, C. (2005). Imitation: Is cognitive neuroscience solving the correspondence problem? *Trends in Cognitive Sciences, 9*(10), 489-495. doi: 10.1016/j.tics.2005.08.007

Bregman, A. S., & Steiger, H. (1980). Auditory streaming and vertical localization: Interdependence of "what" and "where" decisions in audition. *Perception & Psychophysics, 28*(6), 539-546. doi: 10.3758/BF03198822

Brochard, R., Dufour, A., & Després, O. (2004). Effect of musical expertise on visuospatial abilities: Evidence from reaction times and mental imagery. *Brain and Cognition, 54*(2), 103-109. doi: 10.1016/S0278-2626(03)00264-1

Burger, B. (2013). *Move the way you feel: Effects of musical features, perceived emotions, and personality on music-induced movement.* PhD thesis, University of Jyväskylä, Finland.

Burger, B., Thompson, M. R., Luck, G., Saarikallio, S., & Toiviainen, P. (2013). Influences of rhythm-and timbre-related musical features on characteristics of music-induced movement. *Frontiers in Psychology, 4*, Article 183. doi: 10.3389/fpsyg.2013.00183

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafo, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*(5), 365-376. doi: 10.1038/nrn3475

Cabrera, D., & Morimoto, M. (2007). Influence of fundamental frequency and source elevation on the vertical localization of complex tones and complex tone pairs. *The Journal of the Acoustical Society of America, 122*(1), 478-488. doi: 10.1121/1.2736782

Caramiaux, B., Bevilacqua, F., Bianco, T., Schnell, N., Houix, O., & Susini, P. (2014). The role of sound source perception in gestural sound description. *ACM Transactions on Applied Perception, 11*(1), Article 1. doi: 10.1145/2536811

Caramiaux, B., Bevilacqua, F., & Schnell, N. (2010). Towards a gesture-sound cross-modal analysis. In S. Kopp & I. Wachsmuth (Eds.), *Gesture in embodied communication and human-computer interaction* (Vol. 5934, pp. 158-170). Berlin: Springer.

Casasanto, D., & Bottini, R. (2010). Can mirror-reading reverse the flow of time? In C. Hölscher, T. F. Shipley, M. O. Belardinelli, J. A. Bateman & N. S. Newcombe (Eds.), *Spatial cognition VII* (pp. 335-345). Berlin: Springer.

Casasanto, D., & Jasmin, K. (2012). The hands of time: Temporal gestures in English speakers. *Cognitive Linguistics, 23*(4), 643-674. doi: 10.1515/cog-2012-0020

Casasanto, D., Phillips, W., & Boroditsky, L. (2003). *Do we think about music in terms of space? Metaphoric representation of musical pitch.* Paper presented at the 25th Annual Conference of the Cognitive Science Society, Boston, MA.

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. doi: 10.1162/jocn.2008.20018

Chiou, R., & Rich, A. N. (2012). Cross-modality correspondence between pitch and spatial location modulates attentional orienting. *Perception, 41*(3), 339-353. doi: 10.1068/p7161

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(03), 181-204. doi: 10.1017/S0140525X12000477

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis, 58*(1), 7-19. doi: 10.2307/3328150

Clarke, E. F. (2005). *Ways of listening: An ecological approach to the perception of musical meaning*. New York: Oxford University Press.

Claxton, G. (1980). Cognitive psychology: A suitable case for what sort of treatment? In G. Claxton (Ed.), *Cognitive psychology: New directions*. London: Routledge & Kegan Paul.

Collier, W. G., & Hubbard, T. L. (1998). Judgments of happiness, brightness, speed and tempo change of auditory stimuli varying in pitch and tempo. *Psychomusicology: Music, Mind & Brain, 17*(1), 36-55. doi: 10.1037/h0094060

Cook, N. (1990). *Music, imagination, and culture*. Oxford: Oxford University Press.

Cook, N. (2001). Between process and product: Music and/as performance. *Music Theory Online, 7*(2), 1-31.

Cooper, W. E. (1979). *Speech perception and production: Studies in selective adaptation*. Norwood, NJ: Ablex.

Cooperrider, K., & Núñez, R. (2009). Across time, across the body: Transversal temporal gestures. *Gesture, 9*(2), 181-206. doi: 10.1075/gest.9.2.02coo

Costa-Giomi, E. (2005). Does music instruction improve fine motor abilities? *Annals of the New York Academy of Sciences, 1060*(1), 262-264. doi: 10.1196/annals.1360.053

Cowart, M. (2004). Embodied cognition. *The Internet Encyclopedia of Philosophy.* Retrieved 8 January 2014, from http://www.iep.utm.edu/embodcog/

Cowles, J. T. (1935). An experimental study of the pairing of certain auditory and visual stimuli. *Journal of Experimental Psychology, 18*(4), 461-469. doi: 10.1037/h0062202

Cox, A. (2006). Hearing, feeling, grasping gestures. In A. Gritten & E. King (Eds.), *Music and gesture* (pp. 45-60). Aldershot: Ashgate.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*(1), 7-29. doi: 10.1177/0956797613504966

Davidson, L., & Colley, B. (1987). Children's rhythmic development from age 5 to 7: Performance, notation, and reading of rhythmic patterns. In J. C. Peery, I. W. Peery & T. W. Draper (Eds.), *Music and child development* (pp. 107-136). New York: Springer.

Davidson, L., & Scripp, L. (1988). Young children's musical representations: Windows on music cognition. In J. A. Sloboda (Ed.), *Generative processes in music: The psychology of performance, improvisation, and composition* (pp. 195-230). New York: Oxford University Press.

Davidson, L., Scripp, L., & Welsh, P. (1988). "Happy Birthday": Evidence for conflicts of perceptual knowledge and conceptual understanding. *Journal of Aesthetic Education, 22*(1), 65-74. doi: 10.2307/3332965

De Bruyn, L., Moelants, D., & Leman, M. (2012). An embodied approach to testing musical empathy in participants with an autism spectrum disorder. *Music and Medicine, 4*(1), 28-36. doi: 10.1177/1943862111415116

De Dreu, M., Van der Wilk, A., Poppe, E., Kwakkel, G., & Van Wegen, E. (2012). Rehabilitation, exercise therapy and music in patients with Parkinson's disease: A meta-analysis of the effects of music-based movement therapy on walking ability, balance and quality of life.

Parkinsonism & Related Disorders, 18(1), S114-S119. doi: 10.1016/S1353-8020(11)70036-0

Delalande, F. (1988). La gestique de Gould: Éléments pour une sémiologie du geste musical. In G. Guertin (Ed.), *Glenn Gould pluriel* (pp. 85-111). Québec: Louise Courteau.

Deroy, O., & Auvray, M. (2013). *A new Molyneux's problem: Sounds, shapes and arbitrary crossmodal correspondences*. Paper presented at the Second International Workshop The Shape of Things, Rio de Janeiro, Brazil.

Di Fede, D. (2011). Minim. Retrieved 15th July 2011 http://code.compartmental.net/tools/minim

Dixon, S., Goebl, W., & Widmer, G. (2002). *The Performance Worm: Real time visualisation of expression based on Langner's tempo-loudness animation.* Paper presented at the International Computer Music Conference (ICMC), Gothenburg, Sweden.

Dolscheid, S., Hunnius, S., Casasanto, D., & Majid, A. (2012). *The sound of thickness: Prelinguistic infants' associations of space and pitch.* Paper presented at the 34th Annual Meeting of the Cognitive Science Society, Austin, TX.

Dolscheid, S., Shayan, S., Majid, A., & Casasanto, D. (2013). The thickness of musical pitch: Psychophysical evidence for linguistic relativity. *Psychological Science, 24*(5), 613-621. doi: 10.1177/0956797612457374

Drost, U. C., Rieger, M., Brass, M., Gunter, T. C., & Prinz, W. (2005). When hearing turns into playing: Movement induction by auditory stimuli in pianists. *The Quarterly Journal of Experimental Psychology Section A, 58*(8), 1376-1389. doi: 10.1080/02724980443000610

Dunn, R. E. (1997). Creative thinking and music listening. *Research Studies in Music Education, 8*, 42-55. doi: 10.1177/1321103X9700800105

Egermann, H., Pearce, M., Wiggins, G., & McAdams, S. (2013). Probabilistic models of expectation violation predict psychophysiological emotional responses to live concert music. *Cognitive, Affective, & Behavioral Neuroscience, 13*(3), 533-553. doi: 10.3758/s13415-013-0161-y

Eitan, Z. (2013a). How pitch and loudness shape musical space and motion: New findings and persisting questions. In S.-L. Tan, A. Cohen, S. Lipscomb & R. Kendall (Eds.), *The psychology of music in multimedia* (pp. 161-187). Oxford: Oxford University Press.

Eitan, Z. (2013b). Musical objects, cross-domain correspondences, and cultural choice: Commentary on “Cross-cultural representations of musical shape” by George Athanasopoulos and Nikki Moran. *Empirical Musicology Review, 8*(3-4), 204-207.

Eitan, Z., & Granot, R. Y. (2006). How music moves: Musical parameters and listeners' images of motion. *Music Perception, 23*(3), 221-248. doi: 10.1525/mp.2006.23.3.221

Eitan, Z., & Granot, R. Y. (2007). Intensity changes and perceived similarity: Inter-parametric analogies. *Musicae Scientiae, 11*(1 suppl), 39-75. doi: 10.1177/1029864907011001031

Eitan, Z., & Granot, R. Y. (2011). *Listeners' images of motion and the interaction of musical parameters*. Paper presented at the 10th Conference of the Society for Music Perception and Cognition (SMPC), Rochester, NY.

Eitan, Z., Ornoy, E., & Granot, R. Y. (2012). Listening in the dark: Congenital and early blindness and cross-domain mappings in music. *Psychomusicology: Music, Mind & Brain, 22*(1), 33-45. doi: 10.1037/a0028939

Eitan, Z., Schupak, A., Gotler, A., & Marks, L. E. (2014). Lower pitch is larger, yet falling pitches shrink: Interaction of pitch change and size change in speeded discrimination. *Experimental Psychology, 61*(4), 273-284. doi: 10.1027/1618-3169/a000246

Eitan, Z., Schupak, A., & Marks, L. E. (2008). *Louder is higher: Cross-modal interaction of loudness change and vertical motion in speeded classification.* Paper presented at the 10th International Conference on Music Perception and Cognition, Sapporo, Japan.

Eitan, Z., & Timmers, R. (2010). Beethoven's last piano sonata and those who follow crocodiles: Cross-domain mappings of auditory pitch in a musical context. *Cognition, 114*(3), 405-422. doi: 10.1016/j.cognition.2009.10.013

Eitan, Z., & Tubul, N. (2010). Musical parameters and children's images of motion. *Musicae Scientiae, 14*(2), 89-112. doi: 10.1177/10298649100140S207

Elkoshi, R. (2002). An investigation into children's responses through drawing, to short musical fragments and complete compositions. *Music Education Research, 4*(2), 199-211. doi: 10.1080/1461380022000011911

Ernst, M. O. (2007). Learning to integrate arbitrary signals from vision and touch. *Journal of Vision, 7*(5), 1-14. doi: 10.1167/7.5.7

Espeland, M. (1987). Music in use: Responsive music listening in the primary school. *British Journal of Music Education, 4*(3), 283-297. doi: 10.1017/S026505170000615X

Evans, K. K., & Treisman, A. (2010). Natural cross-modal mappings between visual and auditory features. *Journal of Vision, 10*(1), 1-12. doi: 10.1167/10.1.6

Fadiga, L., Craighero, L., Buccino, G., & Rizzolatti, G. (2002). Speech listening specifically modulates the excitability of tongue muscles: A TMS study. *European Journal of Neuroscience, 15*(2), 399-402. doi: 10.1046/j.0953-816x.2001.01874.x

Fadiga, L., Fogassi, L., Pavesi, G., & Rizzolatti, G. (1995). Motor facilitation during action observation: A magnetic stimulation study. *Journal of Neurophysiology, 73*(6), 2608-2611.

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods, 17*(1), 120-128. doi: 10.1037/a0024445

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London: Sage.

Fitts, P. M., & Seeger, C. M. (1953). SR compatibility: Spatial characteristics of stimulus and response codes. *Journal of Experimental Psychology, 46*(3), 199-210. doi: 10.1037/h0062827

Fodor, J. A. (1983). *The modularity of mind: An essay on faculty psychology*. Cambridge, MA: MIT Press.

Françoise, J. (2013). *Gesture--sound mapping by demonstration in interactive music systems.* Paper presented at the 21st ACM International Conference on Multimedia, Barcelona, Spain.

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences, 360*(1456), 815-836. doi: 10.1098/rstb.2005.1622

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience, 11*(2), 127-138. doi: 10.1038/nrn2787

Fry, B., & Reas, C. (2011). Processing. Retrieved 15th July 2011, from http://processing.org

Fugiel, B. (2011). Waveform circularity from added sawtooth and square wave acoustical signals. *Music Perception, 28*(4), 415-424. doi: 10.1525/mp.2011.28.4.415

Fulford, R., & Ginsborg, J. (2013). The sign language of music: Musical Shaping Gestures (MSGs) in rehearsal talk by performers with hearing impairments. *Empirical Musicology Review, 8*(1), 53-67.

Fung, C. V., & Gromko, J. E. (2001). Effects of active versus passive listening on the quality of children's invented notations and preferences for two pieces from an unfamiliar culture. *Psychology of Music, 29*(2), 128-138. doi: 10.1177/0305735601292003

Gaab, N., Keenan, J. P., & Schlaug, G. (2003). The effects of gender on the neural substrates of pitch memory. *Journal of Cognitive Neuroscience, 15*(6), 810-820. doi: 10.1162/089892903322370735

Gadbury, G. L., & Allison, D. B. (2012). Inappropriate fiddling with statistical analyses to obtain a desirable p-value: Tests to detect its presence in published literature. *PloS ONE, 7*(10), e46363. doi: 10.1371/journal.pone.0046363

Galantucci, B., Fowler, C. A., & Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review, 13*(3), 361-377. doi: 10.3758/bf03193857

Gallace, A., & Spence, C. (2006). Multisensory synesthetic interactions in the speeded classification of visual size. *Attention, Perception, & Psychophysics, 68*(7), 1191-1203. doi: 10.3758/BF03193720

Gallese, V., & Lakoff, G. (2005). The brain's concepts: The role of the sensory-motor system in conceptual knowledge. *Cognitive Neuropsychology, 22*(3-4), 455-479. doi: 10.1080/02643290442000310

George, E. M., & Coch, D. (2011). Music training and working memory: An ERP study. *Neuropsychologia, 49*(5), 1083-1094. doi: 10.1016/j.neuropsychologia.2011.02.001

Godøy, R. I. (1997). Knowledge in music theory by shapes of musical objects and sound-producing actions. In M. Leman (Ed.), *Music, gestalt, and computing* (pp. 89-102). Berlin: Springer.

Godøy, R. I. (2003). Motor-mimetic music cognition. *Leonardo, 36*(4), 317-319. doi: 10.1162/002409403322258781

Godøy, R. I. (2006). Gestural-sonorous objects: Embodied extensions of Schaeffer's conceptual apparatus. *Organised Sound, 11*(2), 149-157. doi: 10.1017/S1355771806001439

Godøy, R. I. (2010a). Gestural affordances of musical sound. In R. I. Godøy & M. Leman (Eds.), *Musical gestures: Sound, movement, and meaning*. New York: Routledge.

Godøy, R. I. (2010b). Images of sonic objects. *Organised Sound, 15*(1), 54-62. doi: 10.1017/S1355771809990264

Godøy, R. I., Haga, E., & Jensenius, A. R. (2006a). *Exploring music-related gestures by sound-tracing: A preliminary study*. Paper presented at the 2nd ConGAS International Symposium on Gesture Interfaces for Multimedia Systems, Leeds, UK.

Godøy, R. I., Haga, E., & Jensenius, A. R. (2006b). Playing "air instruments": Mimicry of sound-producing gestures by novices and experts. In S. Gibet, N. Courty & J.-F. Kamp (Eds.), *Gesture in human-computer interaction and simulation* (pp. 256-267). Berlin: Springer.

Godøy, R. I., & Leman, M. (Eds.). (2010). *Musical gestures: Sound, movement, and meaning*. New York: Routledge.

Gold, N. E. (2011). *Knitting music and programming.* Paper presented at the 11th IEEE International Conference on Source Code Analysis and Manipulation, Williamsburg, VA.

Goldin-Meadow, S. (1999). The role of gesture in communication and thinking. *Trends in Cognitive Sciences, 3*(11), 419-429. doi: 10.1016/S1364-6613(99)01397-2

Gritten, A., & King, E. (Eds.). (2006). *Music and gesture*. Aldershot: Ashgate.

Gritten, A., & King, E. (Eds.). (2011). *New perspectives on music and gesture*. Aldershot: Ashgate.

Gromko, J. E. (1994). Children's invented notations as measures of musical understanding. *Psychology of Music, 22*(2), 136-147. doi: 10.1177/0305735694222003

Gromko, J. E. (1995). Invented iconographic and verbal representations of musical sound: Their information content and usefulness in retrieval tasks. *The Quarterly Journal of Music Teaching and Learning, 6*, 32-43.

Haga, E. (2008). *Correspondences between music and body movement.* PhD thesis, University of Oslo, Norway.

Hair, H. I. (1993/1994). Children's descriptions and representations of music. *Bulletin of the Council for Research in Music Education, 119*, 41-48.

Hallam, S., Rogers, L., & Creech, A. (2008). Gender differences in musical instrument choice. *International Journal of Music Education, 26*(1), 7-19. doi: 10.1177/0255761407085646

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology, 8*(1), 23-34.

Hargreaves, D. J. (1978). Psychological studies of children's drawing. *Educational Review, 30*(3), 247-254. doi: 10.1080/0013191780300306

Hartshorne, J., & Schachner, A. (2012). Tracking replicability as a method of post-publication open evaluation. *Frontiers in Computational Neuroscience, 6*, 8. doi: 10.3389/fncom.2012.00008

Hatten, R. S. (2006). A theory of musical gesture and its application to Beethoven and Schubert. In A. Gritten & E. King (Eds.), *Music and gesture* (pp. 1-23). Aldershot: Ashgate.

Haueisen, J., & Knösche, T. R. (2001). Involuntary motor activity in pianists evoked by music perception. *Journal of Cognitive Neuroscience, 13*(6), 786-792. doi: 10.1162/08989290152541449

Heyes, C. (2010). Where do mirror neurons come from? *Neuroscience & Biobehavioral Reviews, 34*(4), 575-583. doi: 10.1016/j.neubiorev.2009.11.007

Heylen, E., Moelants, D., & Leman, M. (2006). *Singing along with music to explore tonality.* Paper presented at the 9th International Conference on Music Perception and Cognition / 6th Triennial Conference of the European Society for the Cognitive Sciences of Music, Bologna, Italy.

Hickok, G., Buchsbaum, B., Humphries, C., & Muftuler, T. (2003). Auditory–motor interaction revealed by fMRI: Speech, music, and working memory in area Spt. *Journal of Cognitive Neuroscience, 15*(5), 673-682. doi: 10.1162/jocn.2003.15.5.673

Himberg, T., & Thompson, M. (2011). Learning and synchronising dance movements in South African songs – Cross-cultural motion-capture study. *Dance Research, 29*(2), 305-328. doi: 10.3366/drs.2011.0022

Hooper, P. P., & Powell, E. R. (1970). Influences of musical variables on pictorial connotations. *Journal of Psychology, 76*(1), 125-128. doi: 10.1080/00223980.1970.9916829

Huron, D. (1996). The melodic arch in Western folksongs. *Computing in Musicology, 10*, 3-23.

Huron, D. (2008). Science & music: Lost in music. *Nature, 453*(7194), 456-457. doi: 10.1038/453456a

Hyde, K. L., Lerch, J., Norton, A., Forgeard, M., Winner, E., Evans, A. C., & Schlaug, G. (2009). Musical training shapes structural brain development. *The Journal of Neuroscience, 29*(10), 3019-3025. doi: 10.1523/jneurosci.5118-08.2009

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), e124. doi: 10.1371/journal.pmed.0020124

Jensenius, A. R., Wanderley, M. M., Godøy, R. I., & Leman, M. (2010). Musical gestures: Concepts and methods in research. In R. I. Godøy & M. Leman (Eds.), *Musical gestures: Sound, movement, and meaning* (pp. 12-35). New York: Routledge.

Johnson, M. (2007). *The meaning of the body: Aesthetics of human understanding*. Chicago: University of Chicago Press.

Johnson, M., & Larson, S. (2003). "Something in the way she moves"-metaphors of musical motion. *Metaphor and Symbol, 18*(2), 63-84. doi: 10.1207/S15327868MS1802_1

Juslin, P. N. (2013). The value of a uniquely psychological approach to musical aesthetics: Reply to the commentaries on 'A unified theory of musical emotions'. *Physics of Life Reviews, 10*(3), 281-286. doi: 10.1016/j.plrev.2013.07.011

Karmiloff-Smith, A. (1992). *Beyond modularity: A developmental perspective on cognitive science*. Cambridge, MA: MIT press.

Keehner, M., & Fischer, M. H. (2012). Unusual bodies, uncommon behaviors: Individual and group differences in embodied cognition in spatial tasks. *Spatial Cognition & Computation, 12*(2-3), 71-82. doi: 10.1080/13875868.2012.659303

Kendon, A. (1982). The study of gesture: Some remarks on its history. *Recherches Sémiotiques/Semiotic Inquiry, 2*, 45-62.

Kerchner, J. L. (2000). Children's verbal, visual, and kinesthetic responses: Insight into their music listening experience. *Bulletin of the Council for Research in Music Education, 146*, 31-50.

Kestenberg-Amighi, J., Loman, S., Lewis, P., & Sossin, K. M. (Eds.). (1999). *The meaning of movement: Developmental and clinical perspectives of the Kestenberg Movement Profile*. New York: Brunner-Routledge.

Kilner, J. M., & Lemon, R. N. (2013). What we know currently about mirror neurons. *Current Biology, 23*(23), R1057-R1062. doi: 10.1016/j.cub.2013.10.051

Kohler, E., Keysers, C., Umiltà, M. A., Fogassi, L., Gallese, V., & Rizzolatti, G. (2002). Hearing sounds, understanding actions: Action representation in mirror neurons. *Science, 297*(5582), 846-848. doi: 10.1126/science.1070311

Kohn, D., & Eitan, Z. (2009). *Musical parameters and children's movement responses.* Paper presented at the 7th Triennial Conference of the European Society for the Cognitive Sciences of Music, Jyväskylä, Finland.

Kohn, D., & Eitan, Z. (2012). *Seeing sound moving: Congruence of pitch and loudness with human movement and visual shape.* Paper presented at the 12th International Conference on Music Perception and Cognition / 8th Triennial Conference of the European Society for the Cognitive Sciences of Music, Thessaloniki, Greece.

Kozak, M., Nymoen, K., & Godøy, R. I. (2012). Effects of spectral features of sound on gesture type and timing. In E. Efthimiou, G. Kouroupetroglou & S.-E. Fotinea (Eds.), *Gesture and sign language in human-computer interaction and embodied communication* (pp. 69-80). Berlin: Springer.

Kraus, N., & Chandrasekaran, B. (2010). Music training for the development of auditory skills. *Nature Reviews Neuroscience, 11*(8), 599-605. doi: 10.1038/nrn2882

Krumhansl, C. L., & Kessler, E. J. (1982). Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys. *Psychological Review, 89*(4), 334-368. doi: 10.1037/0033-295X.89.4.334

Kubik, G. (1983). Kognitive Grundlagen afrikanischer Musik. In A. Simon (Ed.), *Musik in Afrika* (pp. 327-400). Berlin: Staatliche Museen Preußischer Kulturbesitz.

Kuhn, T. S. (2012[1962]). *The structure of scientific revolutions* (50th Anniversary ed.). Chicago: University of Chicago Press.

Kumar, S., Sedley, W., Barnes, G. R., Teki, S., Friston, K. J., & Griffiths, T. D. (2014). A brain basis for musical hallucinations. *Cortex, 52*, 86-97. doi: 10.1016/j.cortex.2013.12.002

Küssner, M. B. (2013). Music and shape. *Literary and Linguistic Computing, 28*(3), 472-479. doi: 10.1093/llc/fqs071

Küssner, M. B., Gold, N., Tidhar, D., Prior, H. M., & Leech-Wilkinson, D. (2011). *Synaesthetic traces: Digital acquisition of musical shapes.* Paper presented at the Supporting Digital Humanities Conference: Answering the unaskable, Copenhagen, Denmark.

Küssner, M. B., & Leech-Wilkinson, D. (2014). Investigating the influence of musical training on cross-modal correspondences and sensorimotor skills in a real-time drawing paradigm. *Psychology of Music, 42*(3), 448-469. doi: 10.1177/0305735613482022

Küssner, M. B., Tidhar, D., Prior, H. M., & Leech-Wilkinson, D. (2014). Musicians are more consistent: Gestural cross-modal mappings of pitch, loudness and tempo in real-time. *Frontiers in Psychology, 5*, Article 789. doi: 10.3389/fpsyg.2014.00789

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago Press.

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics, 33*(2), 363-374. doi: 10.2307/2529786

Lartillot, O., Eerola, T., Toiviainen, P., & Fornari, J. (2008). *Multi-feature modeling of pulse clarity: Design, validation and optimization.* Paper presented at the International Conference on Music Information Retrieval (ISMIR), Philadelphia, PA.

Lartillot, O., & Toiviainen, P. (2007). *A Matlab toolbox for musical feature extraction from audio.* Paper presented at the International Conference on Digital Audio Effects, Bordeaux, France.

Laukka, P., Eerola, T., Thingujam, N. S., Yamasaki, T., & Beller, G. (2013). Universal and culture-specific factors in the recognition and performance of musical affect expressions. *Emotion, 13*(3), 434-449. doi: 10.1037/a0031388

Leech-Wilkinson, D. (2011). Making music with Alfred Cortot: Ontology, data, analysis. In H. von Loesch & S. Weinzierl (Eds.), *Gemessene Interpretation - Computergestützte Aufführungsanalyse im Kreuzverhör der Disziplinen*. Mainz: Schott.

Leech-Wilkinson, D. (2013). The emotional power of musical performance. In T. Cochrane, B. Fantini & K. R. Scherer (Eds.), *The emotional power of music*. Oxford: Oxford University Press.

Leech-Wilkinson, D. (forthcoming). Musical shape and feeling. In D. Leech-Wilkinson & H. M. Prior (Eds.), *Music and shape*. Oxford: Oxford University Press.

Leman, M. (2007). *Embodied music cognition and mediation technology*. Cambridge, MA: MIT Press.

Leman, M. (2013). The need for a cross-cultural empirical musicology. *Empirical Musicology Review, 8*(1), 19-22.

Leman, M., Desmet, F., Styns, F., Van Noorden, L., & Moelants, D. (2009). Sharing musical expression through embodied listening: A case study based on Chinese Guqin music. *Music Perception, 26*(3), 263-278. doi: 10.1525/mp.2009.26.3.263

Leman, M., & Godøy, R. I. (2010). Why study musical gestures? In R. I. Godøy & M. Leman (Eds.), *Musical gestures: Sound, movement, and meaning* (pp. 3-11). New York: Routledge.

Levitin, D. J., Nuzzo, R. L., Vines, B. W., & Ramsay, J. O. (2007). Introduction to functional data analysis. *Canadian Psychology/Psychologie canadienne, 48*(3), 135-155. doi: 10.1037/cp2007014

Lewkowicz, D. J., & Turkewitz, G. (1980). Cross-modal equivalence in early infancy: Auditory-visual intensity matching. *Developmental Psychology, 16*(6), 597-607. doi: 10.1037/0012-1649.16.6.597

Liberman, A. M., Delattre, P., & Cooper, F. S. (1952). The role of selected stimulus-variables in the perception of the unvoiced stop consonants. *The American Journal of Psychology, 65*(4), 497-516. doi: 10.2307/1418032

Liberman, A. M., Delattre, P. C., Cooper, F. S., & Gerstman, L. J. (1954). The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs: General and Applied, 68*(8), 1-13. doi: 10.1037/h0093673

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21*(1), 1-36. doi: 10.1016/0010-0277(85)90021-6

Lidji, P., Kolinsky, R., Lochy, A., & Morais, J. (2007). Spatial associations for musical stimuli: A piano in the head? *Journal of Experimental Psychology: Human Perception and Performance, 33*(5), 1189-1207. doi: 10.1037/0096-1523.33.5.1189

Lipps, T. (1903). *Ästhetik: Psychologie des Schönen und der Kunst*. Hamburg und Leipzig: Leopold Voss.

Lipscomb, S. D., & Kim, E. M. (2004). *Perceived match between visual parameters and auditory correlates: An experimental multimedia investigation.* Paper presented at the 8th International Conference on Music Perception and Cognition, Evanston, IL.

Ludwig, V. U., Adachi, I., & Matsuzawa, T. (2011). Visuoauditory mappings between high luminance and high pitch are shared by chimpanzees (Pan troglodytes) and humans. *Proceedings of the National Academy of Sciences, 108*(51), 20661-20665. doi: 10.1073/pnas.1112605108

Lundqvist, L.-O., Carlsson, F., Hilmersson, P., & Juslin, P. N. (2009). Emotional responses to music: Experience, expression, and physiology. *Psychology of Music, 37*(1), 61-90. doi: 10.1177/0305735607086048

Maeda, F., Kanai, R., & Shimojo, S. (2004). Changing pitch induced visual motion illusion. *Current Biology, 14*(23), R990-R991. doi: 10.1016/j.cub.2004.11.018

Maes, P.-J., & Leman, M. (2013). The influence of body movements on children's perception of music with an ambiguous expressive character. *PloS ONE, 8*(1), e54682. doi: 10.1371/journal.pone.0054682

Maes, P.-J., Leman, M., Palmer, C., & Wanderley, M. M. (2014). Action-based effects on music perception. *Frontiers in Psychology, 4*, Article 1008. doi: 10.3389/fpsyg.2013.01008

Maes, P.-J., Van Dyck, E., Lesaffre, M., Leman, M., & Kroonenberg, P. M. (2014). The coupling of action and perception in musical meaning formation. *Music Perception, 32*(1), 67-84. doi: 10.1525/mp.2014.32.1.67

Mancini, M., Glowinski, D., & Massari, A. (2012). Realtime expressive movement detection using the EyesWeb XMI platform. In A. Camurri & C. Costa (Eds.), *Intelligent technologies for interactive entertainment* (Vol. 78, pp. 221-222). Berlin: Springer.

Marks, L. E. (1974). On associations of light and sound: The mediation of brightness, pitch, and loudness. *The American Journal of Psychology, 87*(1-2), 173-188. doi: 10.2307/1422011

Marks, L. E. (1982). Bright sneezes and dark coughs, loud sunlight and soft moonlight. *Journal of Experimental Psychology: Human Perception and Performance, 8*(2), 177-193. doi: 10.1037/0096-1523.8.2.177

Marks, L. E. (2004). Cross-modal interactions in speeded classification. In G. A. Calvert, C. Spence & B. E. Stein (Eds.), *Handbook of multisensory processes* (pp. 85-105). Cambridge, MA: MIT Press.

Marks, L. E., Hammeal, R. J., & Bornstein, M. H. (1987). Perceiving similarity and comprehending metaphor. *Monographs of the Society for Research in Child Development, 52*(1), 1-102. doi: 10.2307/1166084

Martino, G., & Marks, L. E. (1999). Perceptual and linguistic interactions in speeded classification: Tests of the semantic coding hypothesis. *Perception, 28*(7), 903-923. doi: 10.1068/p2866

Matthay, T. (1945[1913]). *Musical interpretation: Its laws and principles, and their application in teaching and performing*. London: Joseph Williams Ltd.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*(5588), 746-748. doi: 10.1038/264746a0

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago: University of Chicago Press.

McNeill, D. (2000). *Language and gesture*. Cambridge: Cambridge University Press.

McNeill, D. (2005). *Gesture and thought*. Chicago: University of Chicago Press.

Melara, R. D., & O'Brien, T. P. (1987). Interaction between synesthetically corresponding dimensions. *Journal of Experimental Psychology: General, 116*(4), 323-336. doi: 10.1037/0096-3445.116.4.323

Micheyl, C., Delhommeau, K., Perrot, X., & Oxenham, A. J. (2006). Influence of musical and psychoacoustical training on pitch discrimination. *Hearing Research, 219*(1), 36-47. doi: 10.1016/j.heares.2006.05.004

Mikutta, C., Maissen, G., Altorfer, A., Strik, W., & Koenig, T. (2014). Professional musicians listen differently to music. *Neuroscience, 268*, 102-111. doi: 10.1016/j.neuroscience.2014.03.007

Miller, A., Werner, H., & Wapner, S. (1958). Studies in physiognomic perception: V. Effect of ascending and descending gliding tones on autokinetic motion. *The Journal of Psychology, 46*(1), 101-105. doi: 10.1080/00223980.1958.9916273

Miller, J. (1991). Channel interaction and the redundant-targets effect in bimodal divided attention. *Journal of Experimental Psychology: Human Perception and Performance, 17*(1), 160-169. doi: 10.1037/0096-1523.17.1.160

Mondloch, C. J., & Maurer, D. (2004). Do small white balls squeak? Pitch-object correspondences in young children. *Cognitive, Affective, & Behavioral Neuroscience, 4*(2), 133-136. doi: 10.3758/CABN.4.2.133

Moore, K. (2012). Brains don't predict; they trial actions. *Frontiers in Psychology, 3*, Article 417. doi: 10.3389/fpsyg.2012.00417

Mossbridge, J. A., Grabowecky, M., & Suzuki, S. (2011). Changes in auditory frequency guide visual–spatial attention. *Cognition, 121*(1), 133-139. doi: 10.1016/j.cognition.2011.06.003

Mudd, S. A. (1963). Spatial stereotypes of four dimensions of pure tone. *Journal of Experimental Psychology, 66*(4), 347-352. doi: 10.1037/h0040045

Müllensiefen, D. (2009). Statistical techniques in music psychology: An update. In R. Bader, C. Neuhaus & C. Morgenstern (Eds.), *Concepts, experiments, and field-work: Studies in systematic musicology* (pp. 193-215). Frankfurt: Peter Lang.

Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The musicality of non-musicians: An index for assessing musical sophistication in the general population. *PloS ONE, 9*(2), e89642. doi: 10.1371/journal.pone.0089642

Musacchia, G., Sams, M., Skoe, E., & Kraus, N. (2007). Musicians have enhanced subcortical auditory and audiovisual processing of speech and music. *Proceedings of the National Academy of Sciences, 104*(40), 15894-15898. doi: 10.1073/pnas.0701498104

Neuhoff, J. G. (2001). An adaptive bias in the perception of looming auditory motion. *Ecological Psychology, 13*(2), 87-110. doi: 10.1207/S15326969ECO1302_2

Noyce, G. L., Küssner, M. B., & Sollich, P. (2013). Quantifying shapes: Mathematical techniques for analysing visual representations of sound and music. *Empirical Musicology Review, 8*(2), 128-154.

Núñez, R., & Cooperrider, K. (2013). The tangle of space and time in human cognition. *Trends in Cognitive Sciences, 17*(5), 220-229. doi: 10.1016/j.tics.2013.03.008

Nuzzo, R. L. (2014). Scientific method: Statistical errors. *Nature, 506*(7487), 150-152. doi: 10.1038/506150a

Nygaard, L. C., Herold, D. S., & Namy, L. L. (2009). The semantics of prosody: Acoustic and perceptual evidence of prosodic correlates to word meaning. *Cognitive Science, 33*(1), 127-146. doi: 10.1111/j.1551-6709.2008.01007.x

Nymoen, K., Caramiaux, B., Kozak, M., & Torresen, J. (2011, November). *Analyzing sound tracings - A multimodal approach to music information retrieval.* Paper presented at the 1st International ACM Workshop on Music Information Retrieval with User-Centered and Multimodal Strategies (MIRUM), Scottsdale, AZ.

Nymoen, K., Godøy, R. I., Jensenius, A. R., & Torresen, J. (2013). Analyzing correspondence between sound objects and body motion. *ACM Transactions on Applied Perception, 10*(2), Article 9. doi: 10.1145/2465780.2465783

Nymoen, K., Torresen, J., Godøy, R. I., & Jensenius, A. (2012). A statistical approach to analyzing sound tracings. In S. Ystad, M. Aramaki, R. Kronland-Martinet, K. Jensen & S. Mohanty (Eds.), *Speech, sound and music processing: Embracing research in India* (pp. 120-145). Berlin: Springer.

Olson, D. R. (1970). *Cognitive development: The child's acquisition of diagonality*. New York: Academic Press.

OpenNI™. (2011). OpenNI™. Retrieved 14th October 2011, from http://www.openni.org/About

Patching, G. R., & Quinlan, P. T. (2002). Garner and congruence effects in the speeded classification of bimodal signals. *Journal of Experimental Psychology: Human Perception and Performance, 28*(4), 755-775. doi: 10.1037/0096-1523.28.4.755

Pearson, L. (2013). Gesture and the sonic event in Karnatak music. *Empirical Musicology Review, 8*(1), 2-14.

Pedley, P. E., & Harper, R. S. (1959). Pitch and the vertical localization of sound. *The American Journal of Psychology, 72*(3), 447-449. doi: 10.2307/1420051

Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (1992). Understanding motor events: A neurophysiological study. *Experimental Brain Research, 91*(1), 176-180. doi: 10.1007/bf00230027

Phillips, M., & Cross, I. (2011). About musical time – Effect of age, enjoyment, and practical musical experience on retrospective estimate of elapsed duration during music listening. In A. Vatakis, A. Esposito, M. Giagkou, F. Cummins & G. Papadelis (Eds.), *Multidisciplinary aspects of time and time perception* (pp. 125-136). Berlin: Springer.

Piaget, J., & Inhelder, B. (1973). *Memory and intelligence*. London: Routledge & Kegan Paul.

Pratt, C. C. (1930). The spatial character of high and low tones. *Journal of Experimental Psychology, 13*(3), 278-285. doi: 10.1037/h0072651

Prinz, W. (1990). A common coding approach to perception and action. In O. Neumann & W. Prinz (Eds.), *Relationships between perception and action* (pp. 167-201). Berlin: Springer.

Prinz, W., & Hommel, B. (Eds.). (2002). *Common mechanisms in perception and action: Attention and performance XIX*. Oxford: Oxford University Press.

Prior, H. M. (2011a). *Links between music and shape: Style-specific; language-specific; or universal?* Paper presented at the Topics in Musical Universals: 1st International Colloquium, Aix-en-Provence, France.

Prior, H. M. (2011b). Report for questionnaire participants. Retrieved from http://www.cmpcp.ac.uk/Prior_Report.pdf

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience, 2*(1), 79-87. doi: 10.1038/4580

Rasmussen, C., & Williams, C. (2006). *Gaussian processes for machine learning*. Cambridge, MA: MIT Press.

Repp, B. H. (1993a). Music as motion: A synopsis of Alexander Truslit's (1938) Gestaltung und Bewegung in der Musik. *Psychology of Music, 21*(1), 48-72. doi: 10.1177/030573569302100104

Repp, B. H. (1993b). *Musical motion: Some historical and contemporary perspectives.* Paper presented at the Stockholm Music Acoustics Conference (SMAC93), Stockholm, Sweden.

Repp, B. H. (2010). Sensorimotor synchronization and perception of timing: Effects of music training and task experience. *Human Movement Science, 29*(2), 200-213. doi: 10.1016/j.humov.2009.08.002

Repp, B. H., & Doggett, R. (2007). Tapping to a very slow beat: A comparison of musicians and nonmusicians. *Music Perception, 24*(4), 367-376. doi: 10.1525/mp.2007.24.4.367

Reybrouck, M., Verschaffel, L., & Lauwerier, S. (2009). Children's graphical notations as representational tools for musical sense-making in a music-listening task. *British Journal of Music Education, 26*(2), 189-211. doi: 10.1017/S0265051709008432

Ridley, J., Kolm, N., Freckelton, R. P., & Gage, M. J. G. (2007). An unexpected influence of widely used significance thresholds on the distribution of reported P-values. *Journal of Evolutionary Biology, 20*(3), 1082-1089. doi: 10.1111/j.1420-9101.2006.01291.x

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience, 27*(1), 169-192. doi: 10.1146/annurev.neuro.27.070203.144230

Rizzolatti, G., Fogassi, L., & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews Neuroscience, 2*(9), 661-670. doi: 10.1038/35090060

Rochester, L., Baker, K., Hetherington, V., Jones, D., Willems, A.-M., Kwakkel, G., . . . Nieuwboer, A. (2010). Evidence for motor learning in Parkinson's disease: Acquisition, automaticity and retention of cued gait performance after training with external rhythmical cues. *Brain Research, 1319*, 103-111. doi: 10.1016/j.brainres.2010.01.001

Roffler, S. K., & Butler, R. A. (1968a). Factors that influence the localization of sound in the vertical plane. *The Journal of the Acoustical Society of America, 43*(6), 1255-1259. doi: 10.1121/1.1910976

Roffler, S. K., & Butler, R. A. (1968b). Localization of tonal stimuli in the vertical plane. *The Journal of the Acoustical Society of America, 43*(6), 1260-1266. doi: 10.1121/1.1910977

Rusconi, E., Kwan, B., Giordano, B. L., Umiltà, C., & Butterworth, B. (2006). Spatial representation of pitch height: The SMARC effect. *Cognition, 99*(2), 113-129. doi: 10.1016/j.cognition.2005.01.004

Schaefer, R. S., Overy, K., & Nelson, P. (2013). Affect and non-uniform characteristics of predictive processing in musical behaviour. *Behavioral and Brain Sciences, 36*(3), 226-227. doi: 10.1017/S0140525X12002373

Schaeffer, P. (1966). *Traité des objets musicaux*. Paris: Editions du Seuil.

Schiavio, A., & Menin, D. (2013). Embodied music cognition and mediation technology: A critical review. *Psychology of Music, 41*(6), 804-814. doi: 10.1177/0305735613497169

Schubert, E. (2002). Correlation analysis of continuous emotional response to music: Correcting for the effects of serial correlation. *Musicae Scientiae, 6*(1 suppl), 213-236. doi: 10.1177/10298649020050S108

Schubert, E. (2004). Modeling perceived emotion with continuous musical features. *Music Perception, 21*(4), 561-585. doi: 10.1525/mp.2004.21.4.561

Schubert, E., & Dunsmuir, W. (1999). Regression modelling continuous data in music psychology. In S. W. Yi (Ed.), *Music, mind, and science* (pp. 298-352). Seoul: National University Press.

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences, 17*(11), 565-573. doi: 10.1016/j.tics.2013.09.007

Shapiro, L. (2007). The embodied cognition research programme. *Philosophy Compass, 2*(2), 338-346. doi: 10.1111/j.1747-9991.2007.00064.x

Shayan, S., Ozturk, O., Bowerman, M., & Majid, A. (2014). Spatial metaphor in language can promote the development of cross-modal mappings in children. *Developmental Science*, Advance online publication. doi: 10.1111/desc.12157

Shepard, R. N. (1982). Geometrical approximations to the structure of musical pitch. *Psychological Review, 89*(4), 305-333. doi: 10.1037/0033-295X.89.4.305

Shin, Y. K., Proctor, R. W., & Capaldi, E. J. (2010). A review of contemporary ideomotor theory. *Psychological Bulletin, 136*(6), 943-974. doi: 10.1037/a0020541

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359-1366. doi: 10.1177/0956797611417632

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2013). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, Advance online publication. doi: 10.1037/a0033242

Sloboda, J. A. (2013). *How does it strike you? Obtaining artist-directed feedback from the audience at a site-specific performance of a Monteverdi opera*. Paper presented at the Performance Studies Network Second International Conference, Cambridge, UK.

Smalley, D. (1997). Spectromorphology: Explaining sound-shapes. *Organised Sound, 2*(2), 107-126. doi: 10.1017/S1355771897009059

Smith, K. C., Cuddy, L. L., & Upitis, R. (1994). Figural and metric understanding of rhythm. *Psychology of Music, 22*(2), 117-135. doi: 10.1177/0305735694222002

Smith, L. B., & Sera, M. D. (1992). A developmental analysis of the polar structure of dimensions. *Cognitive Psychology, 24*(1), 99-142. doi: 10.1016/0010-0285(92)90004-L

Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics, 73*(4), 971-995. doi: 10.3758/s13414-010-0073-7

Spence, C., & Deroy, O. (2012). Crossmodal correspondences: Innate or learned? *i-Perception, 3*(5), 316-318. doi: 10.1068/i0526ic

Spilka, M. J., Steele, C. J., & Penhune, V. B. (2010). Gesture imitation in musicians and non-musicians. *Experimental Brain Research, 204*, 549-558. doi: 10.1007/s00221-010-2322-3

Stainer, J., & Barrett, W. A. (1888[1876]). *A dictionary of musical terms* (3rd thousand ed.). London: Novello, Ewer and Co.

Stein, B. E., & Meredith, M. A. (1993). *The merging of the senses*. Cambridge, MA: MIT Press.

Stern, D. N. (2004). *The present moment in psychotherapy and everyday life*. New York: W. W. Norton & Company.

Stern, D. N. (2010). *Forms of vitality: Exploring dynamic experience in psychology, the arts, psychotherapy, and development*. Oxford: Oxford University Press.

Stevens, J. C., & Marks, L. E. (1965). Cross-modality matching of brightness and loudness. *Proceedings of the National Academy of Sciences, 54*(2), 407-411.

Stewart, L., Walsh, V., & Frith, U. (2004). Reading music modifies spatial mapping in pianists. *Attention, Perception, & Psychophysics, 66*(2), 183-195. doi: 10.3758/BF03194871

Stumpf, C. (1883). *Tonpsychologie*. Leipzig: S. Hirzel.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America, 26*(2), 212-215. doi: 10.1121/1.1907309

Suzuki, S., Mills, E., & Murphy, T. C. (1973). *The Suzuki concept: An introduction to a successful method for early music education*. Berkeley: Diablo Press.

Suzuki, Y., & Takeshima, H. (2004). Equal-loudness-level contours for pure tones. *The Journal of the Acoustical Society of America, 116*(2), 918-933. doi: 10.1121/1.1763601

Tan, S.-L., Cohen, A. J., Lipscomb, S. D., & Kendall, R. A. (2013). Future research directions for music and sound in multimedia. In S.-L. Tan, A. J. Cohen, S. D. Lipscomb & R. A. Kendall (Eds.), *The psychology of music in multimedia*. Oxford: Oxford University Press.

Tan, S.-L., & Kelly, M. E. (2004). Graphic representations of short musical compositions. *Psychology of Music, 32*(2), 191-212. doi: 10.1177/0305735604041494

Tan, S.-L., Wakefield, E. M., & Jeffries, P. W. (2009). Musically untrained college students' interpretations of musical notation: Sound, silence, loudness, duration, and temporal order. *Psychology of Music, 37*(1), 5-24. doi: 10.1177/0305735608090845

Tervaniemi, M., Just, V., Koelsch, S., Widmann, A., & Schröger, E. (2005). Pitch discrimination accuracy in musicians vs nonmusicians: An event-related potential and behavioral study. *Experimental Brain Research, 161*(1), 1-10. doi: 10.1007/s00221-004-2044-5

Thelen, E., & Smith, L. B. (1994). *A dynamic systems approach to the development of cognition and action*. Cambridge, MA: MIT Press.

Thompson, M. (2012). *The application of motion capture to embodied music cognition research.* PhD thesis, University of Jyväskylä, Finland.

Trimble, O. C. (1934). Localization of sound in the anterior-posterior and vertical dimensions of "auditory" space. *British Journal of Psychology. General Section, 24*(3), 320-334. doi: 10.1111/j.2044-8295.1934.tb00706.x

Truslit, A. (1938). *Gestaltung und Bewegung in der Musik*. Berlin-Lichterfelde: Chr. Friedrich Vieweg.

Upitis, R. (1987). Children's understanding of rhythm: The relationship between development and music training. *Psychomusicology: Music, Mind & Brain, 7*(1), 41-60. doi: 10.1037/h0094187

Upitis, R. (1990). Children's invented notations of familiar and unfamiliar melodies. *Psychomusicology: A Journal of Research in Music Cognition, 9*(1), 89-106. doi: 10.1037/h0094156

Upitis, R. (1992). *Can I play you my song? The compositions and invented notations of children*. Portsmouth, NH: Heinemann.

Van Dyck, E. (2013). *The influence of music and emotion on dance movement.* PhD thesis, Ghent University, Belgium.

Van Dyck, E., Moelants, D., Demey, M., Deweppe, A., Coussement, P., & Leman, M. (2013). The impact of the bass drum on human dance movement. *Music Perception, 30*(4), 349-359. doi: 10.1525/mp.2013.30.4.349

van Wijck, F., Knox, D., Dodds, C., Cassidy, G., Alexander, G., & MacDonald, R. (2012). Making music after stroke: Using musical activities to enhance arm function. *Annals of the New York Academy of Sciences, 1252*(1), 305-311. doi: 10.1111/j.1749-6632.2011.06403.x

Verschaffel, L., Reybrouck, M., Janssens, M., & Van Dooren, W. (2010). Using graphical notations to assess children's experiencing of simple and complex musical fragments. *Psychology of Music, 38*(3), 259-284. doi: 10.1177/0305735609336054

Vines, B. W., Krumhansl, C. L., Wanderley, M. M., & Levitin, D. J. (2006). Cross-modal interactions in the perception of musical performance. *Cognition, 101*(1), 80-113. doi: 10.1016/j.cognition.2005.09.003

Vines, B. W., Nuzzo, R. L., & Levitin, D. J. (2005a). Analyzing temporal dynamics in music. *Music Perception, 23*(2), 137-152. doi: 10.1525/mp.2005.23.2.137

Vines, B. W., Nuzzo, R. L., & Levitin, D. J. (2005b). Analyzing temporal dynamics in music: Differential calculus, physics, and functional data analysis techniques. *Music Perception, 23*(2), 137-152. doi: 10.1525/mp.2005.23.2.137

von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing, 17*(4), 395-416. doi: 10.1007/s11222-007-9033-z

Wacom™. (2011). Intuos4 Series Product Brochure. Retrieved 18th July 2011, from http://www.wacom.eu/_bib_user/dealer/bro_int4_en.pdf

Wagner, K., & Dobkins, K. R. (2011). Synaesthetic associations decrease during infancy. *Psychological Science, 22*(8), 1067-1072. doi: 10.1177/0956797611416250

Wagner, S., Winner, E., Cicchetti, D., & Gardner, H. (1981). "Metaphorical" mapping in human infants. *Child Development, 52*(2), 728-731. doi: 10.2307/1129200

Walker, P., Bremner, J. G., Mason, U., Spring, J., Mattock, K., Slater, A., & Johnson, S. P. (2010). Preverbal infants' sensitivity to synaesthetic cross-modality correspondences. *Psychological Science, 21*(1), 21-25. doi: 10.1177/0956797609354734

Walker, P., & Smith, S. (1986). The basis of Stroop interference involving the multimodal correlates of auditory pitch. *Perception, 15*(4), 491-496. doi: 10.1068/p150491

Walker, R. (1985). Mental imagery and musical concepts: Some evidence from the congenitally blind. *Bulletin of the Council for Research in Music Education, 85*, 229-237.

Walker, R. (1987). The effects of culture, environment, age, and musical training on choices of visual metaphors for sound. *Perception & Psychophysics, 42*(5), 491-502. doi: 10.3758/BF03209757

Watanabe, D., Savion-Lemieux, T., & Penhune, V. B. (2007). The effect of early musical training on adult motor performance: Evidence for a sensitive period in motor learning. *Experimental Brain Research, 176*(2), 332-340. doi: 10.1007/s00221-006-0619-z

Welch, G. F. (1991). Visual metaphors for sound: A study of mental imagery, language and pitch perception in the congenitally blind. *Canadian Journal of Research in Music Education, 33*, 215-222.

Werner, H. (1980). *Comparative psychology of mental development* (3rd ed.). New York: International Universities Press.

Whalen, D. H. (1984). Subcategorical phonetic mismatches slow phonetic judgments. *Perception & Psychophysics, 35*(1), 49-64. doi: 10.3758/bf03205924

Whitney, K. (2013). *Singing in duet with the listener's voice: A dynamic model of the joint shaping of musical content in live concert performance*. Paper presented at the Performance Studies Network Second International Conference, Cambridge, UK.

Wicker, F. W. (1968). Mapping the intersensory regions of perceptual space. *The American Journal of Psychology, 81*(2), 178-188. doi: 10.2307/1421262

Widmann, A., Kujala, T., Tervaniemi, M., Kujala, A., & Schröger, E. (2004). From symbols to sounds: Visual symbolic information activates sound representations. *Psychophysiology, 41*(5), 709-715. doi: 10.1111/j.1469-8986.2004.00208.x

Williamon, A. (Ed.). (2004). *Musical excellence: Strategies and techniques to enhance performance*. Oxford: Oxford University Press.

Wilson, R. A., & Foglia, L. (2011). Embodied cognition. *The Stanford Encyclopedia of Philosophy* Fall 2011. Retrieved 8 January 2014, from http://plato.stanford.edu/archives/fall2011/entries/embodied-cognition/

Wöllner, C., & Deconinck, F. J. A. (2013). Gender recognition depends on type of movement and motor skill. Analyzing and perceiving biological motion in musical and nonmusical tasks. *Acta Psychologica, 143*(1), 79-87. doi: 10.1016/j.actpsy.2013.02.012

Zatorre, R. J., Chen, J. L., & Penhune, V. B. (2007). When the brain plays music: Auditory-motor interactions in music perception and production. *Nature Reviews Neuroscience, 8*(7), 547-558. doi: 10.1038/nrn2152

Zbikowski, L. M. (2002). *Conceptualizing music: Cognitive structure, theory, and analysis*. New York: Oxford University Press.

Zhao, L. (2001). *Synthesis and acquisition of Laban movement analysis qualitative parameters for communicative gestures.* PhD thesis, University of Philadelphia, PA.

Appendix

Chapter 3

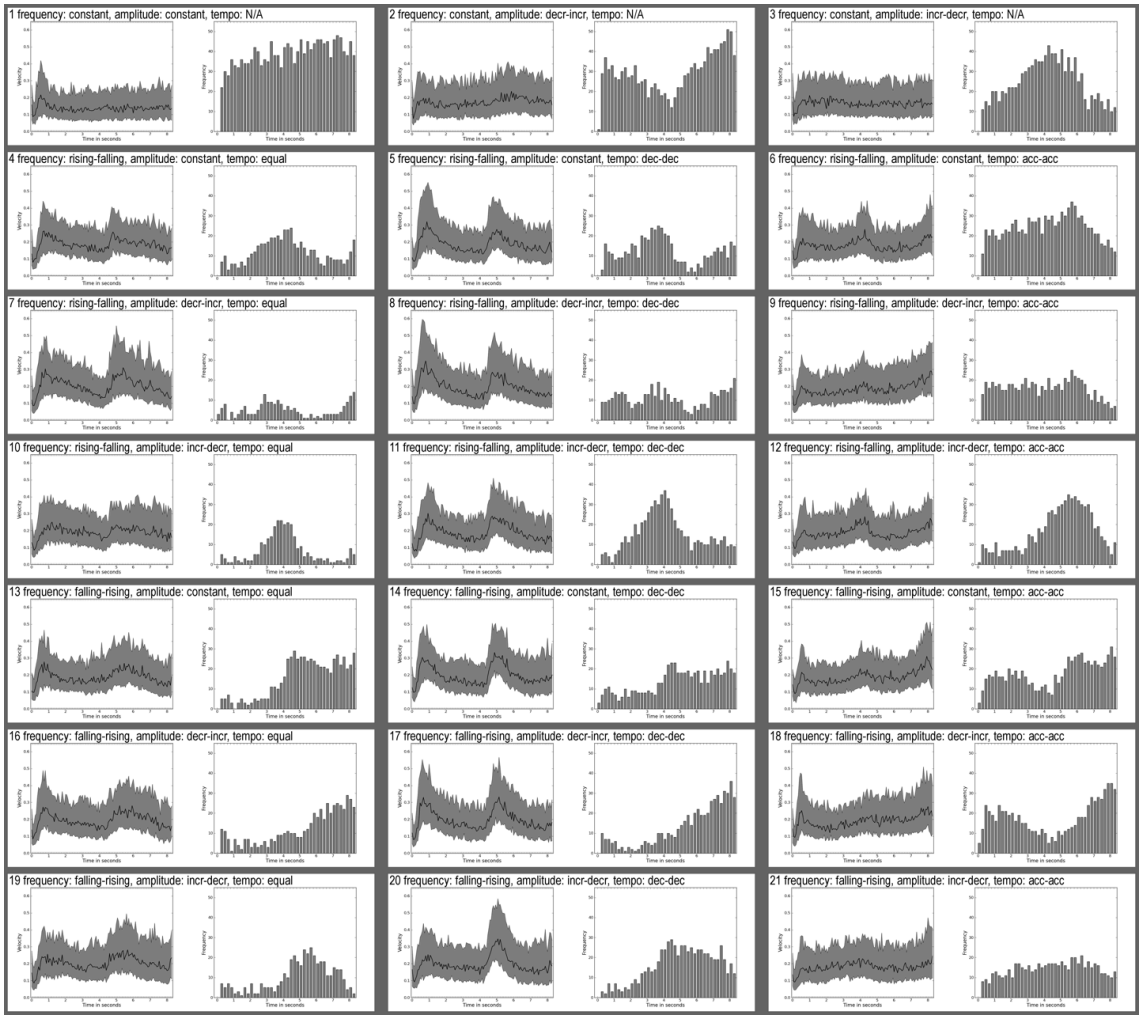
3.1 Musical recordings used in the experiment:

Martha Argerich: Chopin, Prelude in B minor, Op. 28, No. 6 (recorded October 1975). Deutsche Grammophone 1977.

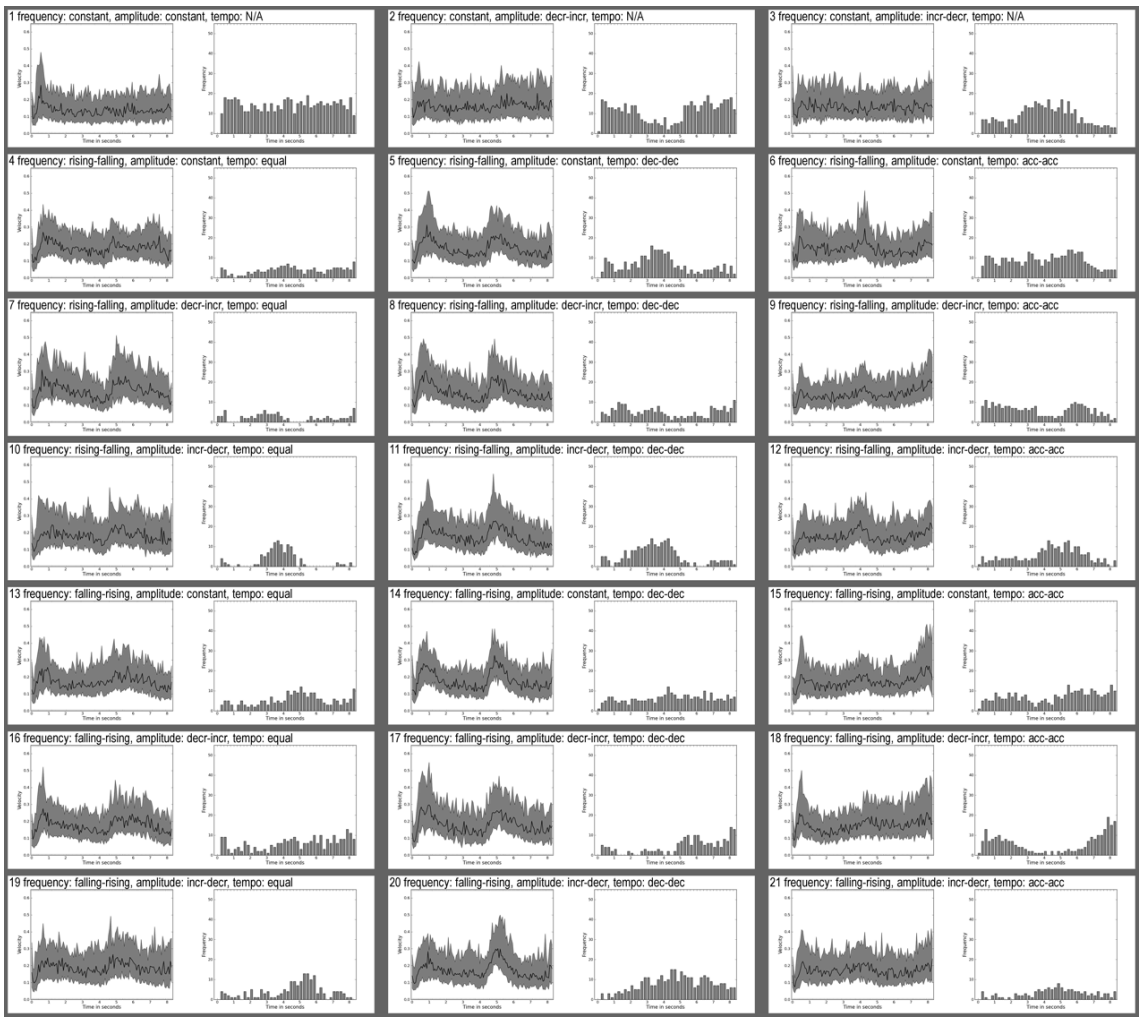
Alfred Cortot: Chopin, Prelude in B minor, Op. 28, No. 6. HMV matrix Cc-8157-3 (recorded 23 March 1926), originally issued on HMV DB 957. Digital transfer © King's College London 2007.

Chapter 5

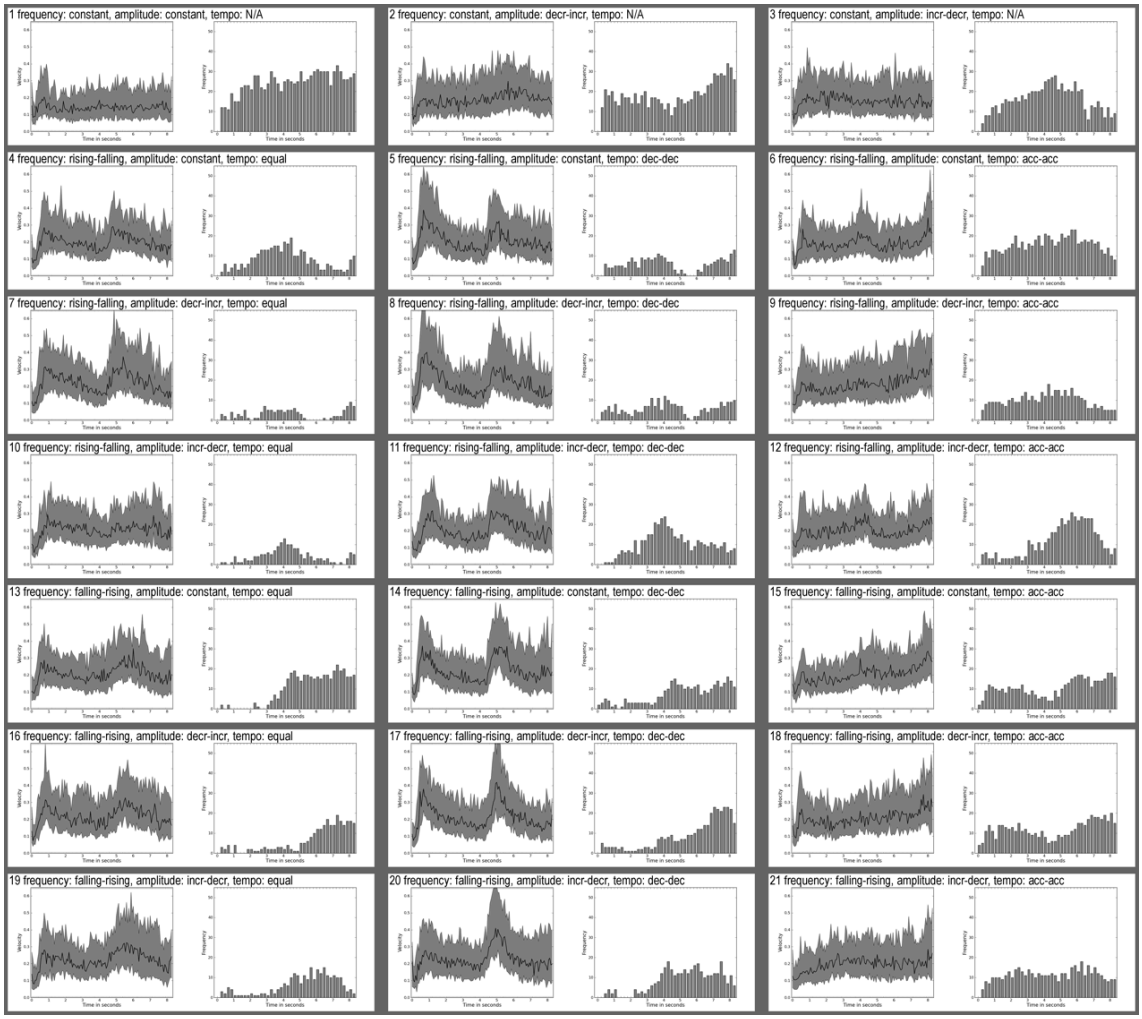
5.1 Velocity and muscular energy profiles overall



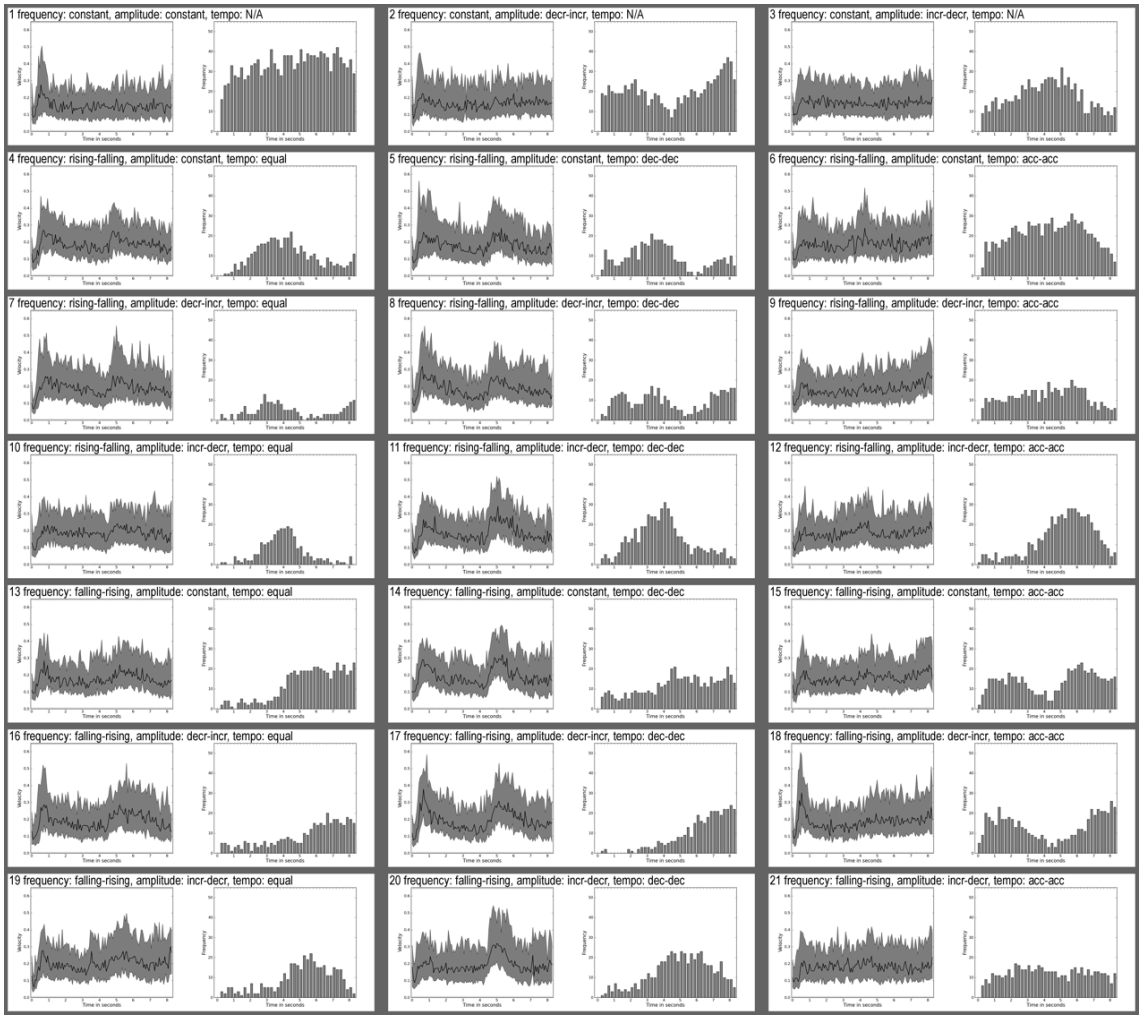
5.2 Velocity and muscular energy profiles for the non-visual condition



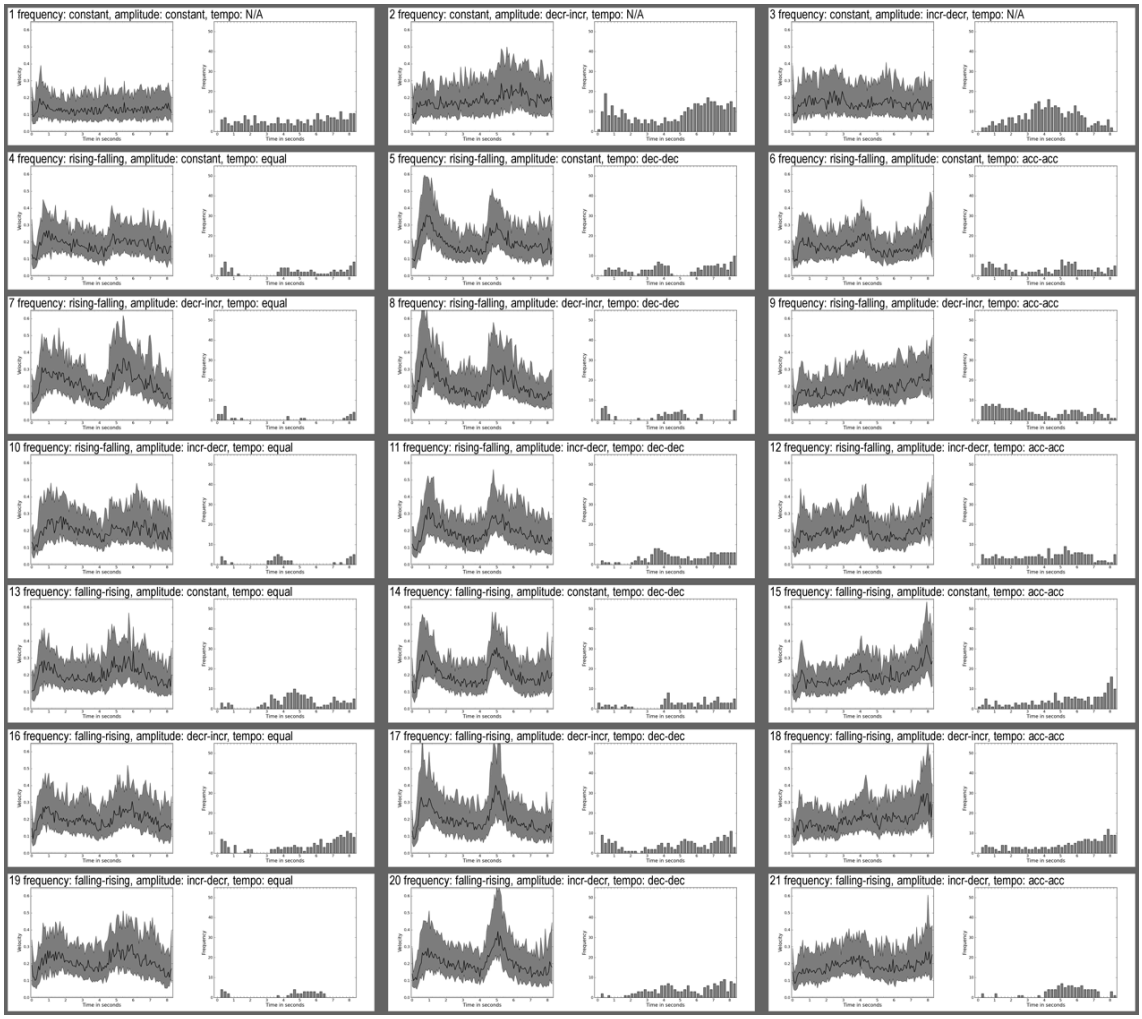
5.3 Velocity and muscular energy profiles for the visual condition



5.4 Velocity and muscular energy profiles for musically untrained participants



5.5 Velocity and muscular energy profiles for musically trained participants

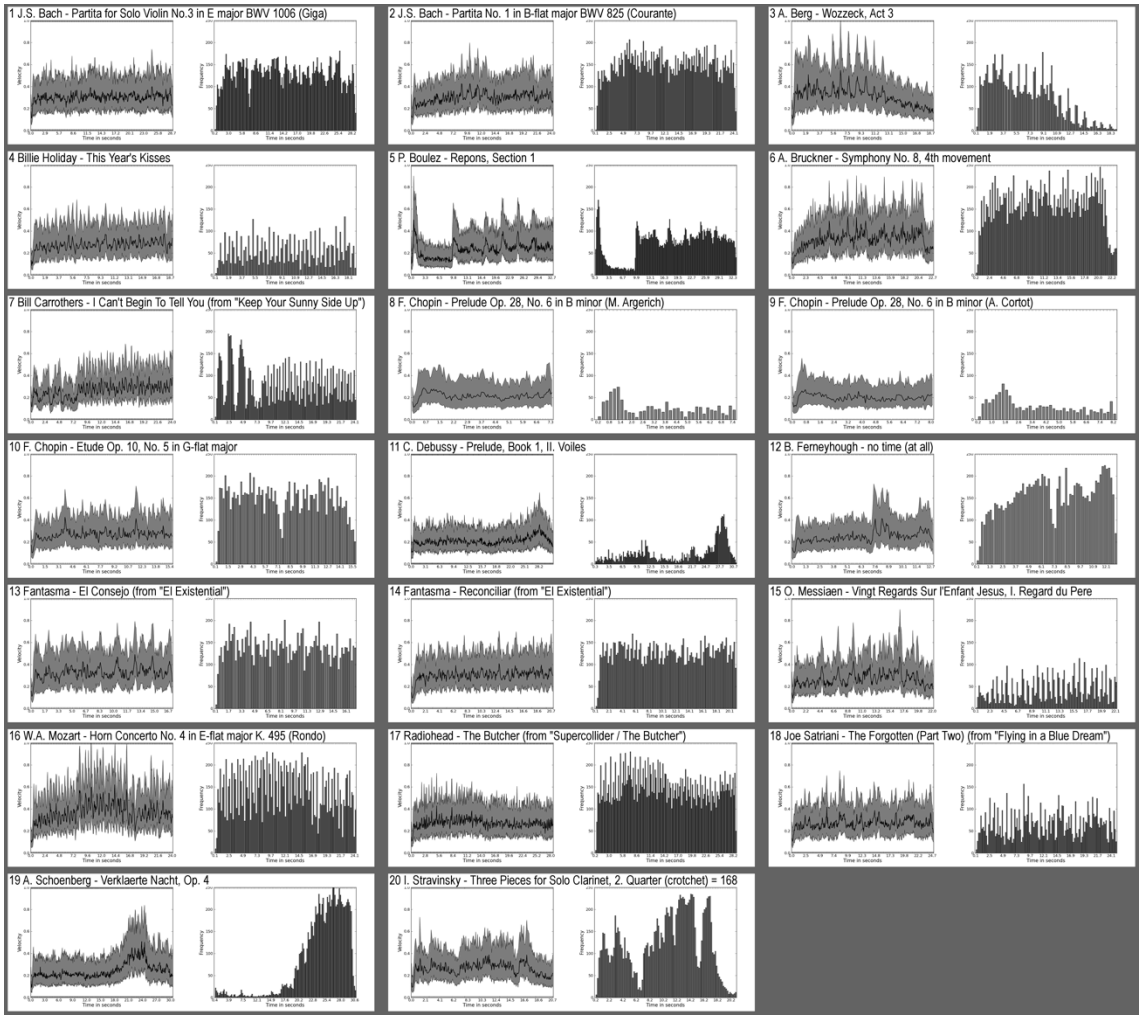


Chapter 6

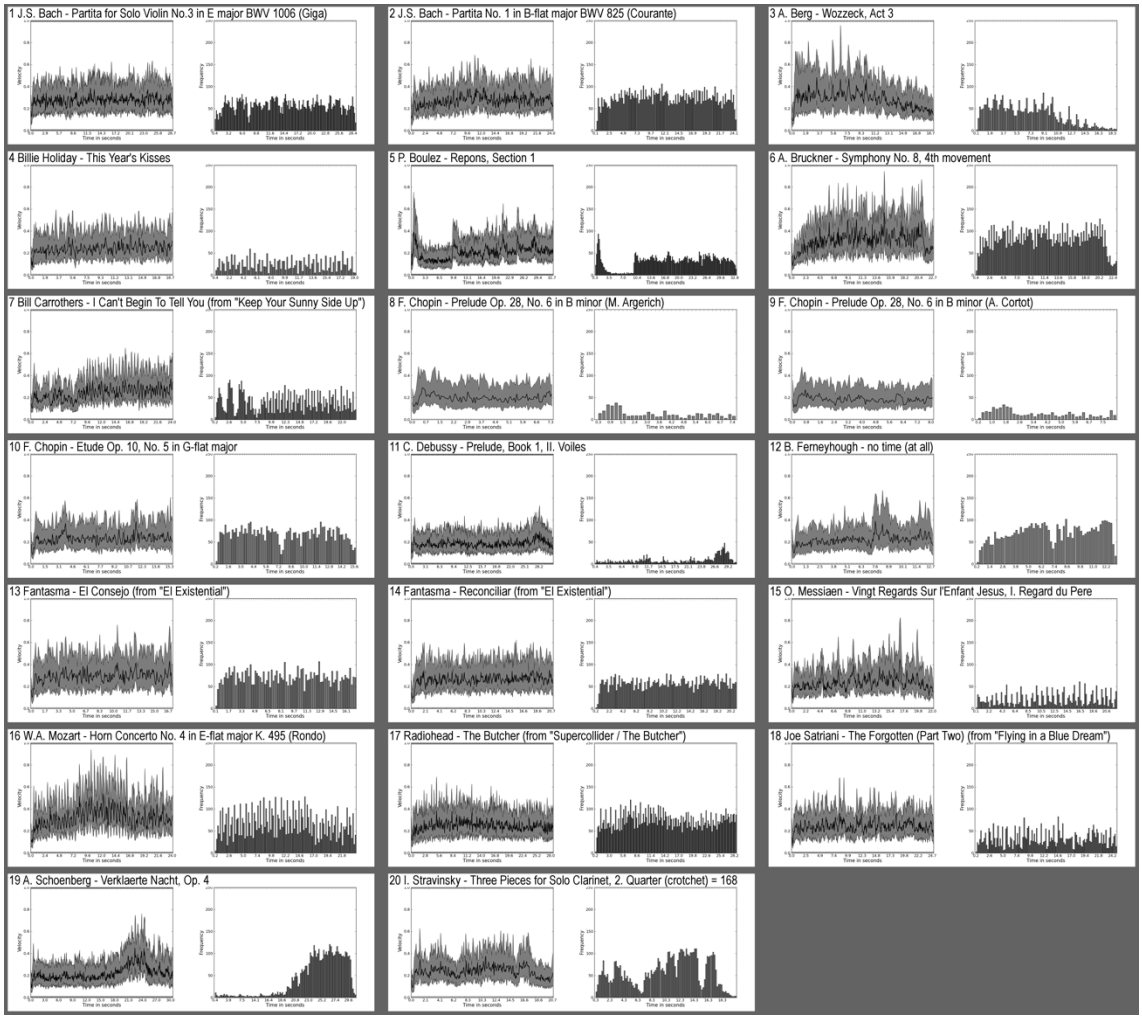
6.1 Detailed overview of musical excerpts

Composer	Performer(s)	Title	Recording	Excerpt starts at:	Length (in s)	Year released
Bach, Johann Sebastian	Rachel Podger	Partita for Solo Violin No. 3 in E major BWV 1006 (Giga)	Channel Classics CCS 14498, track 6	00:00	28.78	2000
Bach, Johann Sebastian	Gustav Leonhardt	Partita No. 1 in B-flat major BWV 825 (Courante)	Virgin Veritas, CD 1, track 3	00:40	24	2004
Berg, Alban	Wiener Philharmoniker (Christoph von Dohnányi)	Wozzeck, Act 3, bb. 103-6	Decca D231 D2, side 4	00:00	18.74	1981
Berlin, Irving	Billie Holiday, Teddy Wilson And His Orchestra	This Year's Kisses	Columbia CXK 85470, CD 2, track 17	02:49	18.63	2001
Boulez, Pierre	Ensemble InterContemporain (Alain Damiens)	Répons, Section 1	Deutsche Grammophon 457 605-2, track 2	00:00	32.9	1998
Bruckner, Anton	Berliner Philharmoniker (Herbert von Karajan)	Symphony No. 8, 4th movement	Deutsche Grammophon 2707 085, side 4	00:00	22.8	1976
Carrothers, Bill	Bill Carrothers (piano), Ben Street (bass), Ari Hoenig (drums)	I Can't Begin To Tell You (from "Keep Your Sunny Side Up")	Pirouet Records, track 2	00:00	24.06	2007
Chopin, Frédéric	Martha Argerich	Prelude Op. 28, No. 6 in B minor	Deutsche Grammophon 2530 721, track 6	00:00	7.32	1977
Chopin, Frédéric	Alfred Cortot	Prelude Op. 28, No. 6 in B minor	HMV DB 957	00:00	8.09	1926
Chopin, Frédéric	Alfred Cortot	Etude Op. 10, No. 5 in G-flat major	HMV DB 2027	00:00	15.3	1933
Debussy, Claude	Arturo Benedetti Michelangeli	Prelude, Book 1, II. Voiles	Deutsche Grammophon 289 477 8382, track 2	01:53	31.06	2009
Ferneyhough, Brian	Geoffrey Morris, Ken Murray (guitars)	no time (at all)	Kairos KAI0013072, track 2	00:00	13.01	2010
Grupo Fantasma	Grupo Fantasma	El Consejo (from "El Existential")	Nat Geo Music, track 4	00:00	16.81	2010
Grupo Fantasma	Grupo Fantasma	Reconciliar (from "El Existential")	Nat Geo Music, track 10	00:00	20.53	2010
Messiaen, Olivier	Pierre-Laurent Aimard	Vingt Regards Sur l'Enfant Jésus, I. Regard du Père	Teldec Classics 3984-26868-2, CD 1, track 1	03:55	21.85	2000
Mozart, Wolfgang Amadeus	Barry Tuckwell, London Symphony Orchestra (Peter Maag)	Horn Concerto No. 4 in E-flat major K. 495 (III. Rondo)	Decca E4757463, CD 1, track 16	00:00	23.8	2006
Radiohead	Radiohead	The Butcher (from "Supercollider / The Butcher")	Ticker Tape Ltd., track 2	00:00	28.41	2011
Satriani, Joe	Joe Satriani	The Forgotten (Part Two) (from "Flying in a Blue Dream")	Relativity Records, track 15	00:00	24.59	1989
Schönberg, Arnold	Santa Fe Chamber Music Festival	Verklärte Nacht, Op. 4	Nonesuch D-79028, side 1	01:53	30.15	1982
Stravinsky, Igor	Charles Neidich	Three Pieces for Solo Clarinet, 2. Quarter (crotchet) = 168	Naxos 8557505, track 12	00:39	20.41	2007

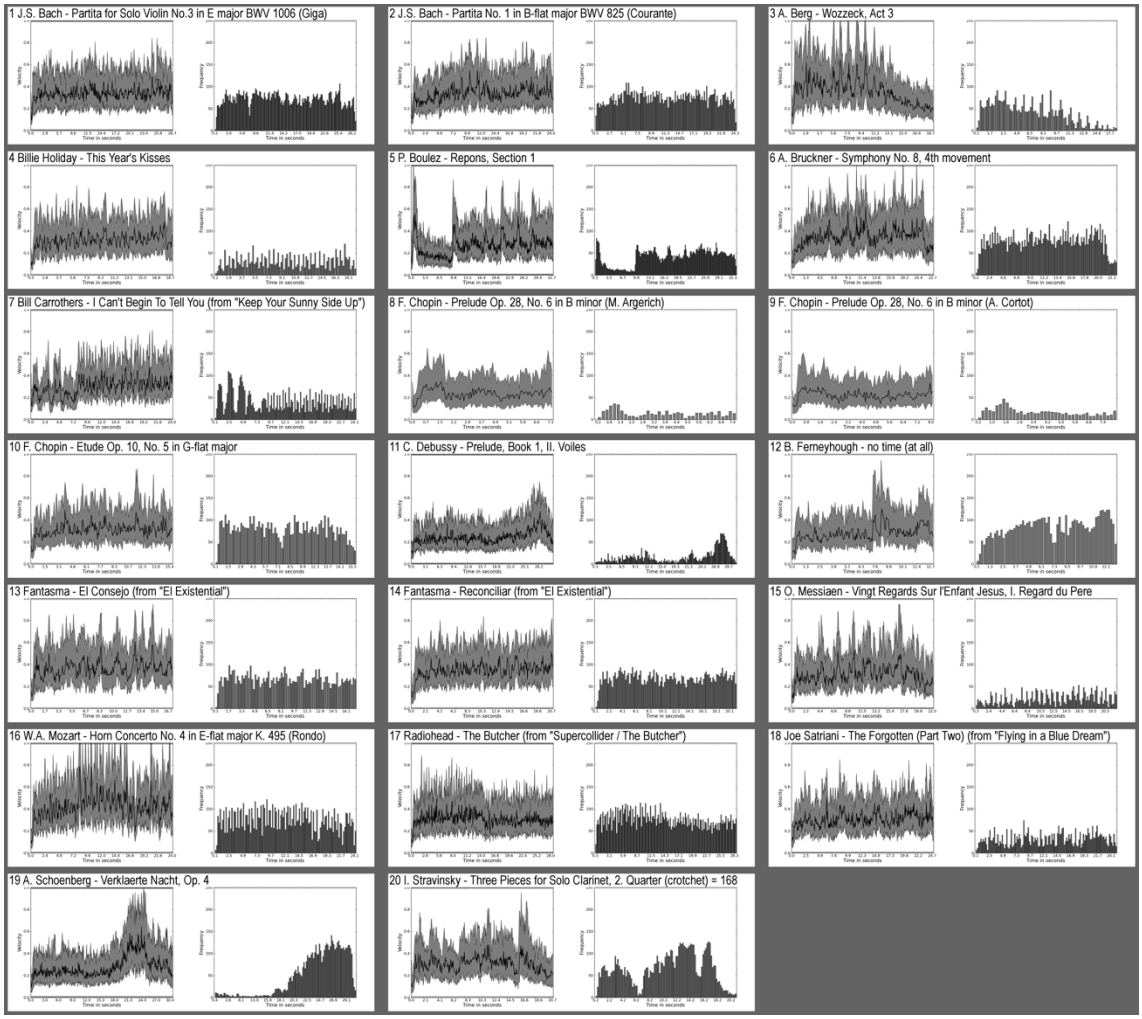
6.2 Velocity and muscular energy profiles overall



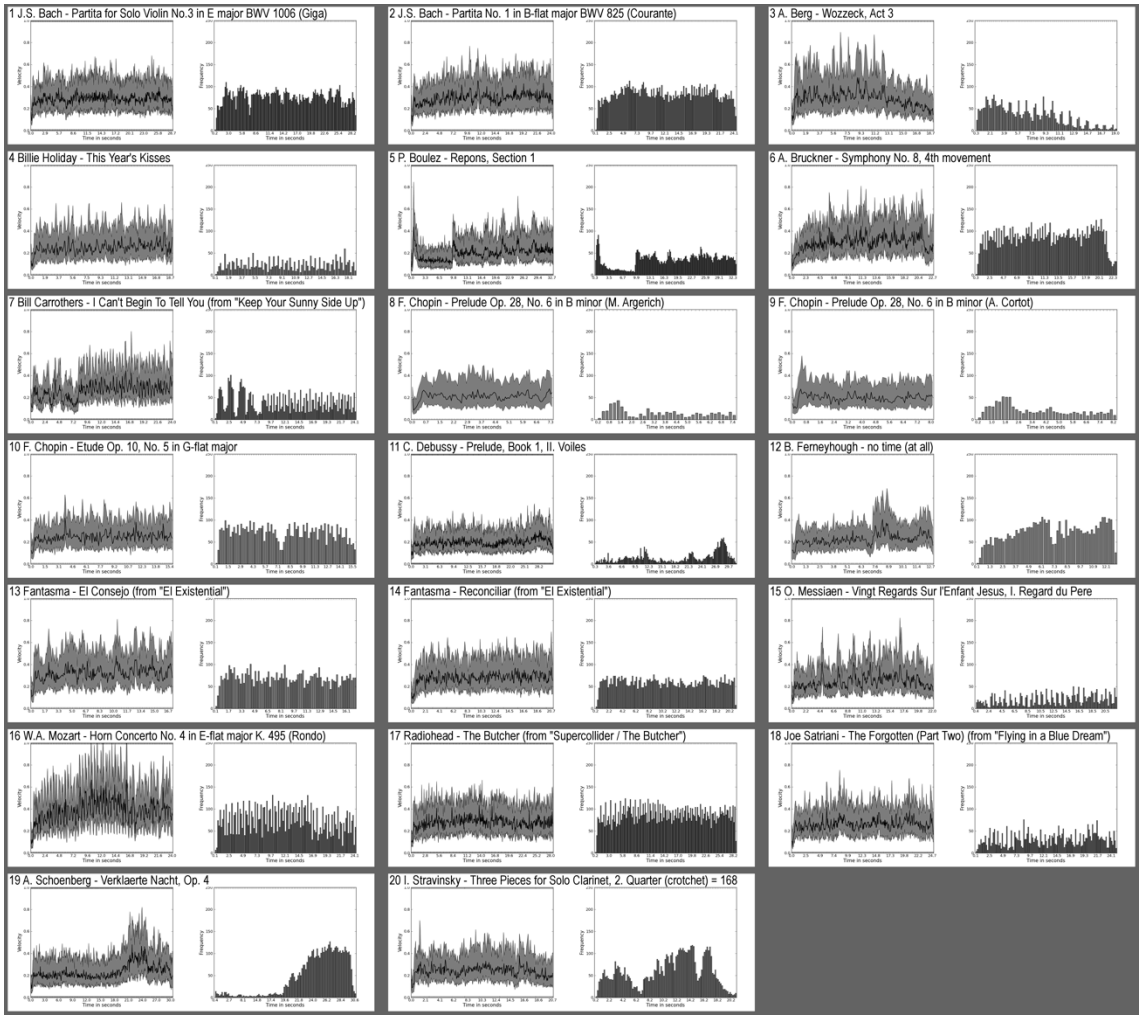
6.3 Velocity and muscular energy profiles for the non-visual condition



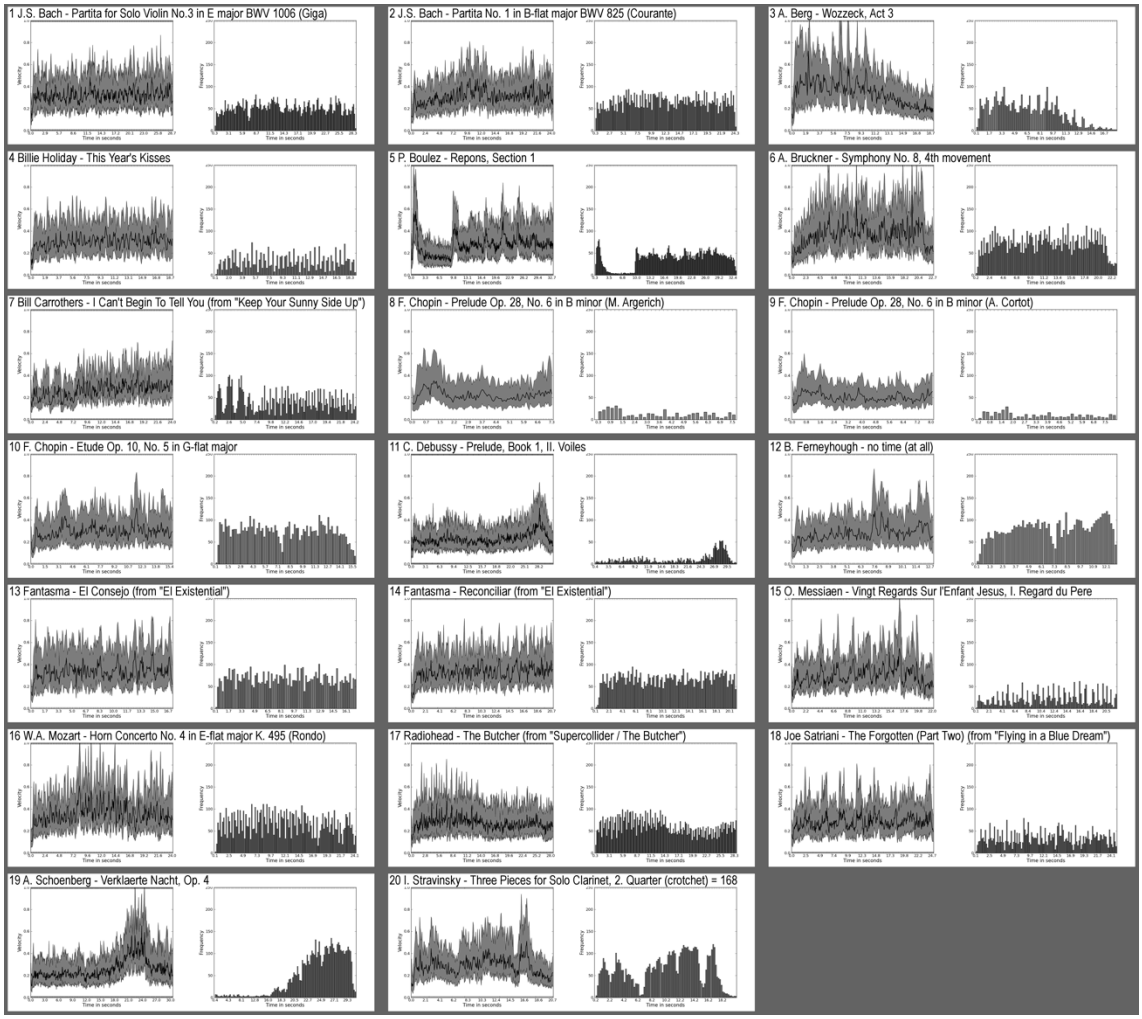
6.4 Velocity and muscular energy profiles for the visual condition



6.5 Velocity and muscular energy profiles for musically untrained participants



6.6 Velocity and muscular energy profiles for musically trained participants



6.7 Detailed overview of different types of gestures

Musical excerpt	Bach (violin)				Bach (keyb)				Berg				Boulez			
Types of gestures	UT	T	NV	V	UT	T	NV	V	UT	T	NV	V	UT	T	NV	V
up and down	X	X	X	X	X		X	X	X	X	X		X	X		X
forwards and backward	X			X	X			X	X			X				
from side to side		X		X	X	X		X	X		X		X		X	
shaking controller/wrist	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
U-shaped gestures	X		X	X	X		X		X		X	X	X		X	
figures of eight	X		X	X	X		X		X	X	X	X	X			X
zigzags	X			X												
lines	X	X	X	X	X	X	X	X		X	X	X		X	X	X
circles	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X
waves	X	X	X	X	X	X	X						X	X	X	X
spirals					X			X	X			X				
arcs/inverted U shapes									X			X				
outward pushes	X			X	X		X		X		X	X				
drawing Ws																
drawing Ms																
sweeping																
no movement		X		X												
number of gestures	10	7	6	12	11	5	8	7	10	5	8	9	7	5	6	6

Musical excerpt	Bruckner				Carrothers				Chopin (M.A.)				Chopin (A.C.)			
Types of gestures	UT	T	NV	V	UT	T	NV	V	UT	T	NV	V	UT	T	NV	V
up and down	X	X	X	X	X	X	X	X	X		X	X	X		X	X
forwards and backward	X		X	X	X		X	X	X		X		X		X	X
from side to side	X	X	X	X	X	X	X	X	X		X	X	X		X	
shaking controller/wrist	X		X	X	X	X	X	X	X		X	X	X		X	X
U-shaped gestures	X		X	X	X		X		X		X	X	X			X
figures of eight					X		X	X	X		X	X	X	X	X	X
zigzags													X		X	X
lines		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
circles	X	X		X					X	X	X	X		X		X
waves	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	
spirals	X		X		X			X	X	X		X	X		X	
arcs/inverted U shapes		X	X						X	X	X	X	X	X	X	X
outward pushes	X	X	X	X	X		X	X	X			X	X		X	
drawing Ws						X	X						X			X
drawing Ms														X	X	
sweeping																
no movement					X		X									
number of gestures	9	7	10	9	11	6	11	9	12	5	10	11	13	6	12	10

Musical excerpt	Ferneyhough				Gr. Fantasma				Messiaen				Mozart			
Types of gestures	UT	T	NV	V	UT	T	NV	V	UT	T	NV	V	UT	T	NV	V
up and down	X			X	X	X	X	X	X	X	X	X	X	X	X	X
forwards and backward	X			X					X		X	X	X		X	X
from side to side	X			X	X	X	X	X	X	X	X	X	X	X	X	X
shaking controller/wrist	X	X	X	X	X	X	X	X	X		X	X	X		X	X
U-shaped gestures					X		X	X	X	X	X	X	X	X	X	
figures of eight													X	X	X	X
zigzags																
lines		X	X	X	X	X		X	X	X	X	X		X	X	X
circles	X	X	X	X		X		X		X		X	X	X	X	X
waves					X		X	X	X	X	X	X	X		X	X
spirals					X		X		X			X	X		X	X
arcs/inverted U shapes														X	X	
outward pushes		X		X	X		X	X		X		X	X		X	X
drawing Ws									X		X		X		X	
drawing Ms						X		X								
sweeping																
no movement																
number of gestures	5	4	3	7	8	6	7	9	9	7	8	10	11	7	13	10

Musical excerpt	Radiohead				Satriani				Schönberg				Stravinsky			
Types of gestures	UT	T	NV	V	UT	T	NV	V	UT	T	NV	V	UT	T	NV	V
up and down	X	X	X	X	X	X	X	X		X		X				
forwards and backward	X	X	X	X	X		X	X								
from side to side	X	X	X	X	X	X	X		X	X	X	X				
shaking controller/wrist	X	X	X	X	X		X	X	X	X	X	X	X	X	X	X
U-shaped gestures	X			X	X		X	X	X	X	X	X	X		X	X
figures of eight					X	X	X		X	X	X	X				
zigzags																
lines	X	X	X	X		X	X	X	X	X	X	X		X	X	X
circles					X	X	X	X		X	X	X	X	X	X	X
waves	X		X		X	X	X	X	X	X	X	X	X			X
spirals	X			X	X			X					X			X
arcs/inverted U shapes	X			X					X			X				
outward pushes	X	X	X	X	X	X	X	X								
drawing Ws									X		X					
drawing Ms														X		X
sweeping	X			X					X		X					
no movement						X		X								
number of gestures	11	6	7	10	10	8	10	10	9	8	9	9	5	4	4	7