

JYU DISSERTATIONS 712

Juan Ignacio Mendoza Garay

Mimetic Relationships between Bodily Movement and Musical Structure

Theory, Measurement, and Application



UNIVERSITY OF JYVÄSKYLÄ
FACULTY OF HUMANITIES AND
SOCIAL SCIENCES

JYU DISSERTATIONS 712

Juan Ignacio Mendoza Garay

**Mimetic Relationships between Bodily
Movement and Musical Structure**
Theory, Measurement, and Application

sitetään Jyväskylän yliopiston humanistis-yhteiskuntatieteellisen tiedekunnan suostumuksella
julkisesti tarkastettavaksi yliopiston vanhassa juhlasalissa S212
marraskuun 13. päivänä 2023 kello 12.

Academic dissertation to be publicly discussed, by permission of
the Faculty of Humanities and Social Sciences of the University of Jyväskylä,
in building Seminarium, Old Festival Hall S212, on November 13, 2023, at 12 o'clock.



JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2023

Editors

Geoff Luck

Department of Music, Art and Culture Studies, University of Jyväskylä

Päivi Vuorio

Open Science Centre, University of Jyväskylä

Copyright © 2023, by the author and University of Jyväskylä

ISBN 978-951-39-9799-1 (PDF)

URN:ISBN:978-951-39-9799-1

ISSN 2489-9003

Permanent link to this publication: <http://urn.fi/URN:ISBN:978-951-39-9799-1>

ABSTRACT

Mendoza Garay, Juan Ignacio

Mimetic Relationships between Bodily Movement and Musical Structure:
Theory, Measurement, and Application

Jyväskylä: University of Jyväskylä, 2023, 107 p. + original articles

(JYU Dissertations

ISSN 2489-9003; 712)

ISBN 978-951-39-9799-1 (PDF)

Music moves us. It moves our body and our feelings. One needs to move the respiratory tract to sing, hands and fingers to play a musical instrument, perhaps the whole body to dance along with music. Even when one doesn't move, one might be emotionally moved by the music. These movements—actual or metaphorical—closely relate to the musical sound, as if one imitated the other. This coupling acts to communicate what is difficult to express otherwise: emotions. In effect, relations between music and bodily movement are numerous and diverse. The goal in this dissertation has been to examine specific aspects of such relations to gain a better understanding of them, as well as to use this understanding to devise novel digital musical instruments. First, a theoretical model was formulated to explain musical interaction. Here, people and musical instruments are regarded as agents that communicate by means of sensory signals organised in a hierarchical temporal structure of gestures at different timescales. This theory was utilised as a framework for the subsequent research, which dealt with measurement of different aspects of the framework. The first aspect related to modelling temporal segmentation of bodily motion. A method was developed and tested which is based on detection of change points, works in real time, and detects perceptually relevant gestures at different timescales. This method was applied to a novel gesturally controlled digital musical instrument, and to a system for musical sonification of daily activity to aid in reducing sedentarism. The second aspect of measurement provided new insights into the extent to which emotions may be conveyed by the body when playing an instrument or when dancing, and how this might be affected by perceptual sensory modalities and personality traits. The most salient factor was found to be not visual but auditory, with minor and major tonality being most strongly related to the perception of negative and positive emotions, respectively. Regarding movement, personality had a significant relationship with the way and extent that emotions were expressed in spontaneous dance, with Openness having the strongest relation, and Neuroticism and Conscientiousness the weakest. These contributions to knowledge serve to better understand musical phenomena and to advance innovation in the design of technologies for making music.

Keywords: music, body, movement, emotion, personality, segmentation, instruments, machine learning, sonification, embodied, agency, gesture, control.

TIIVISTELMÄ (ABSTRACT IN FINNISH)

Mendoza Garay, Juan Ignacio

Mimeettisten kehon liikkeiden ja musiikillisten rakenteiden suhteet: teoria, mittaus ja soveltaminen

Jyväskylä: Jyväskylän yliopisto, 2023, 107 s. + alkuperäiset artikkelit

(JYU Dissertations

ISSN 2489-9003; 712)

ISBN 978-951-39-9799-1 (PDF)

Musiikki liikuttaa meitä. Se liikuttaa kehoamme ja tunteitamme. Laulaakseen pitää liikuttaa hengityselimiä, soitinta soittaakseen käsiä sekä sormia ja musiikin mukana tanssiakseen ehkä koko kehoa. Vaikka ihminen ei liikukaan, musiikki saattaa liikuttaa häntä emotionaalisesti. Nämä liikkeet – todelliset tai metaforiset – liittyvät läheisesti musiikilliseen ääneen, ikään kuin toinen matkisi toista. Tämä kytkentä toimii välittäen asioita, joita on vaikea ilmaista muuten: tunteita. Musiikin ja kehon liikkeen väliset suhteet ovatkin lukuisia ja monelaisia. Tämän väitöskirjan tavoitteena on ollut tarkastella tällaisia suhteita tietyistä näkökulmista niiden ymmärtämiseksi paremmin sekä käyttää tätä ymmärrystä uusien digitaalisten soittimien kehittämiseen. Ensin kehitettiin teoreettinen malli musiikillisen vuorovaikutuksen selittämiseksi. Mallissa ihmiset ja soittimet nähdään toimijoina, jotka kommunikoivat aistisignaalien välityksellä. Aistisignaalit ovat järjestäytyneet hierarkkiseen ajalliseen rakenteeseen eri aikaskaaloilla. Tätä teoriaa käytettiin viitekehyksenä seuraavissa tutkimuksissa, joissa keskityttiin mittaamaan viitekehykseen liittyviä näkökulmia. Ensimmäinen näkökulma liittyi kehon liikkeen ajallisen segmentoinnin mallintamiseen. Tutkimuksessa kehitettiin ja testattiin menetelmä, joka perustuu muutospisteiden havaitsemiseen, toimii reaaliajassa ja havaitsee merkitykselliset eleet eri aikaskaaloilla. Tätä menetelmää sovellettiin uuteen eleohjattavaan digitaaliseen musiikki-instrumenttiin ja päivittäistä toimintaa kuvaavaan musiikilliseen sonifikaatiojärjestelmään, jonka avulla voidaan mahdollisesti vähentää sedentarismia. Mittauksen toinen näkökulma tarjosi uutta tietoa siitä, missä määrin keho voi välittää tunteita instrumenttia soitettaessa tai tanssittaessa, ja miten aistikanavat ja persoonallisuuden piirteet voivat vaikuttaa tähän. Keskeisimpänä tekijänä havaittiin olevan auditiivinen aistikanava visuaalisen sijaan, ja molli- ja duuri-tonaliteetti liittyi vahvimmin negatiivisten ja positiivisten tunteiden aistimiseen. Spontaanissa tanssissa taas persoonallisuudella oli merkittävä yhteys tapaan, jolla tunteet ilmaistaan. Persoonallisuuden piirteistä avoimuudella oli vahvin yhteys, neuroottisuudella ja tunnollisuudella oli heikoin. Nämä tulokset auttavat ymmärtämään musiikillisiä ilmiöitä paremmin. Lisäksi ne auttavat kehittämään innovaatioita, joita voidaan hyödyntää, kun suunnitellaan teknologioita musiikin tekemiseen.

Avainsanat: musiikki, keho, liike, tunne, persoonallisuus, segmentointi, instrumentit, koneoppiminen, sonifikaatio, ruumiillistuva, toimija, ele, ohjaus.

Author Juan Ignacio Mendoza Garay
Department of Music, Art and Culture Studies
University of Jyväskylä
ORCID: 0000-0003-3996-7537

Supervisor Geoff Luck
Department of Music, Art and Culture Studies
University of Jyväskylä

Reviewers Marcelo M. Wanderley
Schulich School of Music
McGill University

Tuomas Eerola
Department of Music
Durham University

Opponent Marcelo M. Wanderley
Schulich School of Music
McGill University

LIST OF ARTICLES

- I Mendoza, J. I., & Thompson, M. R. (2017). Gestural Agency in Human-Machine Musical Interaction. In *The Routledge Companion to Embodied Music Interaction* (pp. 412-419).
<https://doi.org/10.4324/9781315621364-45>
- II Mendoza, J. I., & Thompson, M. R. (2017). Modelling Perceived Segmentation of Bodily Gestures Induced by Music. In *Proceedings of the 25th Anniversary Conference of the European Society for the Cognitive Sciences of Music*. <http://urn.fi/URN:NBN:fi:jyu-201711024121>
- III Mendoza, J.I. (2022). Segmentation Boundaries in Accelerometer Data of Arm Motion Induced by Music: Online Computation and Perceptual Assessment. *Human Technology*, 18(3), 250–266.
<https://doi.org/10.14254/1795-6889.2022.18-3.4>
- IV Thompson, M.R., Mendoza, J.I., Luck, G., & Vuoskoski, J.K. (2023). Relationships between Audio and Movement Features, and Perceived Emotions in Musical Performance. *Music and Science*, 6.
<https://doi.org/10.1177/20592043231177871>
- V Mendoza, J.I., Burger, B., & Luck, G. (2022). Exploring Relations Between Big Five Personality Traits and Musical Emotions Embodied in Spontaneous Dance. *Psychology of Music*, 51(4).
<https://doi.org/10.1177/03057356221135355>
- VI Mendoza, J.I. (2023). The Rearranger Ball: Delayed Gestural Control of Musical Sound using Online Unsupervised Temporal Segmentation. To appear in *Proceedings of the Conference on New Interfaces for Musical Expression*.
- VII Mendoza, J.I., Danso, A., Luck, G., Rantalainen, T., Palmberg, L., & Chastin, S. (2022). Musification of Accelerometry Data Towards Raising Awareness of Physical Activity. In *Proceedings of the Conference on Sonification of Health and Environmental Data*.
<https://doi.org/10.5281/zenodo.7243875>

Author's contribution:

- I Original idea, development of concept, and writing (full).
- II, III Data collection, analysis, development of methods, and writing (full).
- IV Data analysis (correlation and regression), and writing (partial).
- V Methodology, data analysis, and writing (full).
- VI, VII Development of methods and writing (full).

FIGURES

FIGURE 1: Overview of the articles included in this dissertation.....	39
FIGURE 2: Human-machine musical interaction.....	42
FIGURE 3: Online temporal segmentation algorithm.....	46
FIGURE 4: Perceptual assessment of segmentation effectiveness.....	50
FIGURE 5: Effect sizes for perceived emotions.....	53
FIGURE 6: Relevant models of personality traits.	57
FIGURE 7: Proof of concept for musical gesture segmentation.....	59
FIGURE 8: Multigranular segmentation of daily activity accelerometry.....	63

TABLES

TABLE 1 : Best fitting models for each personality trait.....	58
TABLE 2 : Relevant models with all regressors of each subset.	58

CONTENTS

ABSTRACT

TIIVISTELMÄ (ABSTRACT IN FINNISH)

LIST OF ARTICLES

FIGURES AND TABLES

CONTENTS

1	INTRODUCTION	11
2	BACKGROUND	14
2.1	Music and the human body	14
2.1.1	Music moves us	14
2.1.2	Embodied musical interaction.....	15
2.1.3	Musical gestures.....	17
2.2	Segmentation of bodily motion	22
2.2.1	Human visio-temporal segmentation	22
2.2.2	Automatic temporal segmentation.....	23
2.3	Music and emotion	27
2.3.1	Musical emotions	27
2.3.2	Auditory and visual perception.....	28
2.3.3	Relation with personality and bodily motion.....	31
2.4	Technology for making music with broad bodily motion.....	34
2.4.1	Responsiveness of musical instruments	34
2.4.2	Machine learning of continuous gestures.....	35
2.4.3	Sonification of bodily motion for sports and healthcare	36
2.5	Opportunities for research	37
3	AIMS OF THE RESEARCH	39
4	METHODS AND RESULTS.....	42
4.1	Theory.....	42
4.1.1	Article I	42
4.2	Measurement.....	45
4.2.1	Articles II and III	45
4.2.2	Article IV	50
4.2.3	Article V.....	54
4.3	Application	59
4.3.1	Article VI	59
4.3.2	Article VII.....	62
5	DISCUSSION	66
5.1	Gestures and agency in musical interaction	66
5.2	Temporal segmentation of bodily motion.....	67
5.2.1	Online temporal segmentation.....	67

5.2.2	Delayed gestural control of musical sound.....	68
5.2.3	Musical sonification of daily activity	70
5.3	Embodied musical emotions.....	72
5.3.1	Contribution of sensory modality	72
5.3.2	Effect of personality	74
5.3.3	Prospect of practical application.....	76
5.4	A holistic model of mimetic musical interaction	77
5.5	Concluding remarks.....	81
YHTEENVETO (SUMMARY IN FINNISH).....		83
REFERENCES.....		86
ARTICLES		

1 INTRODUCTION

Imagine this:

It is a busy Saturday afternoon at the city. People are flocking towards the stadium, where a legendary rock band will be playing. While the scattered crowd moves towards the venue, a small group of concertgoers sings one of the band's hit songs and a driver passing by honks the car's horn to the rhythm of the singing. They smile to each other and continue their business. A few blocks from the stadium there is a small chapel where an exequy is taking place. A chamber trio plays a funeral march while the coffin is being taken out to the coach. The walking pace of those carrying the deceased matches that of the music. Yet a few blocks away there is a hall where a dance concert is about to be held. As with the concert in the stadium, many people arrive to the hall, but no one is singing or honking horns. Next to the concert hall there is a gymnasium where only one person is exercising. The lonely person could not get tickets for the concert at the stadium but listens to music of that band on a portable device with headphones. A song is starting with a tranquil riff by an acoustic guitar. The person stands in front of a heavily weighted bar, stretching neck, shoulders, and wrists. Distorted guitars and bass now enter the song playing heavy chords in staccato, as cymbals remark the rhythm's accents. The song is building up. At the stadium everyone expects the legends of rock to appear. At the concert hall people have tidily gone to seat at their reserved places. Although the atmosphere at the hall is calm, the audience is expectant. Many hear in their head the famous classical music soon to be played, they can anticipate the dancing that might go along. At the gym, the person grabs the bar, inhales, seemingly all the muscles of their body tighten. The cymbals chime, the drums start playing a powerful march and the full band explodes in heavy metal. The weighted bar comes off the floor as the person pulls it energetically in four beats. On the fifth beat the bar is laid down and every eight beats of the music the bar is lifted. At the stadium the band enters the stage amidst flashing lights and fireworks, immediately starting to play one of their classics. The crowd that before was scattered now moves together in unison, jumping to the beat of the music. The spotlight is on the guitarist, playing one of the most memorable

solos in the history of rock music. The long hair of the musician is blown by a hidden fan. The player, after a series of rapid melodies, holds and bends the last note with the left hand while moving the right hand in the air as if lifting something. When the hand reaches its maximum height and the guitar string reaches its maximum tension, the rest of the band stops playing. The note keeps ringing several seconds while the rockstar, immobile except for the waving hair, holds the invisible object in the air. The crowd is in ecstasy. At the concert hall the first movement of the concert has begun. The dancers sway, jump and twirl along with the music. In the pit, the conductor and the musicians of the orchestra also sway while playing the music, as if they were dancing too. The audience is apparently calm, but most of them move just a little bit sitting in place. Some slightly sway their whole body, some nod their head. Some gently tap their feet or their hands. Many feel an intense desire to move but stay put, they would be embarrassed otherwise. At the chapel the musicians had stopped playing as the coffin has been laid in the coach. A loved one is gone; it is time to move on. In an apartment in the building in front of the chapel a baby starts crying. The mother holds the baby and sings a lullaby. She gently rocks the baby to the rhythm of the song, then the baby stops crying.

Music moves us. It literally moves us. It moves our body. It moves our feelings. But it does not always move us. It does not move everyone in the same way. Music *may* move us. How? Music *can* move us. Why? Could those questions or their answers help us to explore new ways or perhaps better ways, of making music?

It seems that an attempt to answer may start by the simple observation that music is mimetic, in the sense that it involves mimicry, a close correspondence between the musical sound and the movement and posture of the human body. As in the story narrated above, correspondences exist between music (i.e., musical sound) and movement of the human body. Such movement may be that which produces a sound (e.g., playing an instrument, singing, honking a car's horn), that which moves along with the sound (e.g., dancers dancing, rockstars gesticulating, mourners marching, lifters lifting weights) or that which is internal (i.e., imagining to move, being affected by the musical performance).

The paragraphs above represent the origins of my motivations to start the research project described in this dissertation. Those are, however, similar to the motivations for much research already conducted. Therefore, my research was planned to focus on specific problems and was arranged to proceed in three sequential areas: theory, measurement, and application. The area concerned with theory was intended to provide a general framework. The area concerned with measurement was intended to quantitatively evaluate aspects of the theoretical framework. The area concerned with application was intended to put into practice the outcomes of the foregoing theory and measurement. For each of these areas, surveys of previous research allowed me to identify the specific problems to investigate. To wit, the area of theory inquired on the structure of musical interaction among people and their musical instruments (in

the broadest sense, including novel and future digital musical instruments), where the role of the human body and its gestures are at the core. Towards the application of this theoretical framework, the area of measurement investigated the temporal segmentation of musical movement of the body (e.g., dancing) and the embodiment of musical emotions (i.e., how emotions may be conveyed by the body when playing an instrument or when dancing). The area of application comprised the development of novel technological applications for making music, based on the measurement of temporal segmentation of bodily movement. These explored the idea that broad movement of the human body may be used to make music, opposed to the capability of fine control that is usually expected from musical instruments.

This dissertation is written around seven research articles and is organised in five sections. The next section (BACKGROUND) is an overview of previous research on the topics covered by the articles, providing background knowledge, context, and definitions. Section 3 (AIMS OF THE RESEARCH) succinctly explains motivations for, and goals of the studies reported in the articles, describing how they relate. Section 4 (METHODS AND RESULTS) is composed of summaries of the methods and results reported in the articles. Section 5 (DISCUSSION) examines the results of the studies, considering the linkages between them and offering suggestions for future research.

2 BACKGROUND

2.1 Music and the human body

2.1.1 Music moves us

The idea that we can be *moved by music* is widespread and often a cliché. Why are we moved by music? Before attempting to answer that question, we need to understand what it means—the definition of—to be *moved* by music. But the idea is so prevalent that it seems not to need clarification. In specialised literature and popular science articles, the question “Why music moves us?” suffices to introduce the inquiry on the powerful effects of music (Garrido et al., 2019; Hodges & Wilkins, 2015; Levitin, 2010; Schrock, 2009). Why might we take for granted that music moves us? What is in music so powerful to have such a strong and universal effect on us, individually and collectively?

“To be moved” by something is generally understood as to be affected emotionally by something. The word “emotion” has historically been related to the “movement of the soul” and a “physical disturbance and bodily movement” (Dixon, 2012). Examples of early investigation of the matter may be found in ancient Greek philosophy. Schoen-Nazzaro (1978) analysed Plato’s and Aristotle’s writings on the purpose of music. Plato’s writings in his books *The Laws* and *Republic*, are quoted respectively: “music education should measure and order the movements of the soul” and “we can recognise in music different types of emotional movements”. Schoen-Nazzaro infers the following:

For Plato music's power over emotional states is founded on its force as an imitation of emotion. When someone listens to a piece, he picks up its emotional movement and begins to move in the open way. To paraphrase Plato, musical movement, containing an expression of emotion, conveys this emotion to the listener.

Aristotle’s sayings in his books *Problemata* and *De Anima*, also emphasise that music and emotions are movements, as they can move faster and slower, up

and down. Schoen-Nazzaro interpreted this correspondence as facilitating the induction of movement into the listener's soul. These ideas converge in the essential notion that bodily processes (feeling emotions and moving the body) resemble music and vice versa, the essence of the concept "mimesis", appearing throughout this dissertation.

Beyond semantics, analogies and speculation, an emotion is indeed a bodily process. It involves motion being covert as in electrochemical processes in the nervous system. It also may involve more noticeable motion such as bodily movements or posture indicating the feeling of such emotion, for example gesticulation with hands or facial expression. Hence, scholars have rightfully asserted that "music moves us to tears" when researching music (Weber, 1891), and "music moves us almost without our effort" when researching dance (Wilkinson, 1869). Being "moved by music" is an idea also present in more recent research dealing with the emotional effects of music (Margulis, 2007; Sievers et al., 2012; Ter Bogt et al., 2011; Zeiner, 2010). Moreover, "moved" (as in being emotionally affected or *touched*) has also been used as a quantitative measure for the emotional effect of music (Eerola et al., 2016; Eerola et al., 2021; Juslin & Laukka, 2004; Vuoskoski & Eerola, 2017; Vuoskoski et al. 2022). Likewise, "being moved" as paronomasia, has been used in the description of research that has quantitatively measured bodily motion when listening to music (Burger et al., 2012; Demos & Chaffin, 2018; Solberg, 2015; Swarbrick et al., 2019; Zelechowska et al., 2020) and the relations between musicians' movement and their experienced emotions (Van Zijl & Luck, 2012).

The characteristic capacity of music to affect the human body, including feeling of emotions and bodily movement, has been appealing as an object of scientific inquiry, people's curiosity, and sheer fascination, for the whole of the history of humanity. It has not been, however, until the past century, possible to conduct rigorous research aided by technologies that measure bodily motion, brain activity, characteristics of sound, and so forth. This research has led to vigorous scholarly debate and formulation of explanatory theories, of which the following paragraphs present an overview.

2.1.2 Embodied musical interaction

Embodied Cognition is a research program based on the observation that the body is involved, or may even be essential, in cognitive processes (Lakoff & Johnson, 1980, 1999; Shapiro, 2011; Varela et al., 1991). The Embodied Cognition program is strongly rooted in the idea that the mind is shaped by enactive interactions with the environment, including other living beings. Likewise, "Embodied Music Cognition" (Leman, 2008) was proposed as a research paradigm based on evidence rather than speculation, facilitated by technologies to record human motion. The central idea of this paradigm is that an understanding of music perception needs to consider interaction, in which the role of the human body in an environment is essential.

According to the theory of Embodied Music Cognition, a person's spontaneous movement when listening to music can reflect the person's

perception of the music. The empirical study of this theory observes the correspondences between bodily movement and music based on qualitative and quantitative analysis (Haga, 2008). Qualitative investigation has observed, for example, that music teachers explain musical sound with bodily movements, especially with their hands (Fatone et al., 2011). Quantitative investigation has shown that bodily movement induced by music relates to features of the music, such as periodicity and kinetic energy (Toiviainen et al., 2010) or tonality (MacRitchie et al., 2013). It has been argued that moving the body (as in dancing) enhances the experience of music, and that the movement of a musician conveys information that complements or supplements the sound of the musical instrument. An example of such information is the communication of emotions (Leman & Godøy, 2010). In a review of theoretical and empirical research related to Embodied Music Cognition, Leman and Maes (2014a) propose that the human body is a mediator for meaning formation, linking performance and perception via mirroring processes. While the role of the human body in musical activity had been studied before (see Schneider, 2010), Leman's work proposed a comprehensive line of research which since his initial coinage of the term Embodied Music Cognition in 2007 has been embraced, challenged, and built upon.

Schiavio and Menin (2013) complement Leman's (Leman, 2008) arguments proposing that understanding of musical activity may be improved by considering the motor knowledge of the body, rather than considering the body as only a mediator. Geeves and Sutton (2014) also complement Leman and Maes' work (Leman & Maes, 2014a) pointing out that considering the body as a mediator separates perception from action, going against the principles of the theory (i.e., in its claim to be different from viewing body and mind as separated). Also, they challenge the dichotomy between perception and performance expressed in Leman and Mae's work, presenting some empirical evidence of the exchange of information between performers and listeners. Nonetheless, later Leman and Maes (2014b) state that "the mechanisms behind music perception are the same for music performance". Matyja (2016) made further criticism to the Embodied Music Cognition theory (Leman, 2008; Leman & Maes, 2014a), in particular the hypothesis that "embodied sensorimotor engagement is essential to both production and perception of music", arguing that it doesn't rigorously connect with empirical research and advocating for more investigation.

However, Leman, in the preface to his book where the denomination "Embodied Music Cognition" firstly appears in the literature, (Leman, 2008) commented that the book is an essay with ideas acknowledging its incompleteness. These ideas originated from the observation of the state of the art at the time, of systematic musicology, being highly interdisciplinary, strongly empirical, taking into account physiological and biomechanical processes (e.g., the brain, feeling of emotions, bodily motion), and facilitated by technologies to measure those processes (e.g., brain activity measurement and imaging, motion capture). Although the initial ideas expressed by Leman and

colleagues about Embodied Music Cognition as a theory and hypothesis have been questioned, there are general points of consensus. Firstly, that interaction is essential to music cognition and that the human body plays an important role in such interaction. Second, such interaction is enactive by necessity.

Leman, in his work, uses the phrase *interaction with music*, a concept connected to the idea that music is perceived by means of senses, that is, mediated by the human body. However, *interaction with music*, may fall short to describe the richness of musical phenomena, because when engaging in musical activity we don't interact only with music, be it the score, the sound, or the meaning. When engaging in musical activity we may also interact with other beings (living or inanimate) and in general with the environment. I contend that a better conceptualisation of music as an embodied phenomenon, is *musical interaction*. This term opens the possibility of viewing music as a process of communication, rather than solely an abstract object within the process of interaction. In this view the term *embodied musical interaction* may be applied to various modalities of musical experiences. For example, listening to musical sound (e.g., recorded music, a bird singing, a storm), observing a musical performance (which can be dancing along with music), playing a musical instrument, singing, and dancing to music, alone or with others. In sum, participating in one way or another in a musical activity.

2.1.3 Musical gestures

A useful construct to describe how musical interaction takes place is “musical gesture”. The term is closely connected with the more basic understanding of gesture as an action, usually a pose or movement of the human body, intended to express an idea or sentiment (Cambridge University Press & Assessment, n.d.; Merriam-Webster, n.d.). For example, waving a hand to greet or smiling to agree. From this it is possible to infer that a gesture is finite (i.e., has a starting and an end), and that it is a linguistic object as it implies coding and decoding of meaning. These characteristics are connected to the various ways in which the term gesture has been used in musical contexts, which may be grouped in two categories (Leman & Godøy, 2010, pp. 5-10; Schneider, 2010, p. 71).

The first category considers gestures by the literal definition, as movements or poses of the human body with the intention to express something. They can act as sources of information that supplement or complement musical sound. For example, a singer may move their hands to portray what they are singing; a guitarist playing heavy metal rock music may shake their head energetically to the beat. This notion is at least as early as classical Greek philosophy, in which gestures were seen as a means to realise mimesis, which literally meant a representation of mental and emotional states (Schneider, 2010, p. 77). The second category considers the properties of gestures applied to musical sound. This notion arguably derives from the observation that the intention in the production of music and the resulting musical sound are linked by bodily motion or lack thereof. For instance, production of vocal music requires the activation of the vocal organs, but it also

may involve the movement of other parts of the body to produce the singing utterance. Likewise, movement is required to play a musical instrument, for example blowing a tube, plucking a string, hitting an object, or pressing a key. The movement when singing or playing an instrument (including the human voice) normally corresponds directly to the sound, suggesting that the movement and therefore the gesture is present in the music itself. One could say that the sound has a shape. This implies as well that the sounding gesture has the gestural properties of being finite and of carrying information, often expressing a sentiment. Take for instance a sung melody rising in pitch. This melody is finite, obviously it cannot be produced beyond the lung capacity of the singer. The melody rising in pitch requires an effort, therefore a stress, which in turn communicates the corresponding sentiment. A similar process occurs when a person is excited and yells. These analogies suggest the mimetic nature of musical gestures.

There are many and notable examples in published literature about the use of the phrase and concept “musical gesture”, in the description and analysis of music (Schneider, 2010; Mazzola et al., 2016). Numerous writings use the first category definition, but during the twentieth century the usage of the second category progressively gained more strength. Possibly so because its metaphorical nature served first to the speculative explanation of how music communicates and later to scientific scrutiny of those speculations.

For example, Parker (1894, p. 229) considered gesture as complementary to music, but not a part of music:

All changes of the body which manifest mental states –changes not included in any of the other modes of expression- may be classified under the general name of gesture.

Other early examples may be found in writings of several authors mentioning composer and choir conductor J.F. Bridge’s method for teaching music notation to his singers based in “musical gestures”. The method consisted of figures made with arms and fingers depicting notes and rests (Novello et al., 1894; Simpson et al., 1894; Wurm, 1896). Similarly, Coomaraswamy regarded as “musical gesture” the hand motion of Indian singers, that may be expressive (as in acting) and may follow the music (Devi et al., 1913, pp. 4-5). Also, in a lecture on music and other arts, British composer H. W. Davies discussed the resemblance of music to gesture, remarking that they both develop in time (Music in Relation to Other Arts, 1910).

In an analysis of Cherubino, the character in Mozart’s opera “The Marriage of Figaro”, Lee (1881, p. 227) begins considering the gesture of the performer and the music as separate, both working towards expressing the state of mind of the character. Then, both facets are merged in the following comment:

What, then, can music give us, with all its powers of suggestion and feeling, if it cannot give us this? It can give us one thing, not another: it can give us emotion, but

it cannot give us the individual whom the emotion possesses. With its determined relations between the audible movement and the psychic movement, it can give us only musical gesture, but never musical portrait; the gesture of composure or of violence, the solemn tread of self-possessed melody, the scuffling of frantically rushing up and down, of throbbing, quivering, gasping, passion-broken musical phrases; it can give us the rhythm which prances and tosses in victory, and the rhythm which droops, and languishes and barely drags itself along for utter despair.

Developing that line of thought, Leichtentritt (1924) in an analysis of contemporary German music, used the term “musical gesture” to refer to the quality of music by which it expresses emotion, however much detached from the bodily component:

Thus we find in Erdmann's short piano pieces all the characteristic traits of Schönberg's manner: the conciseness, aphoristic brevity, the so-called *atonality*, the sudden jumps from high to low notes in the melodic line, the absence of regular, periodic construction, the abruptness of the musical "gesture", expressive of sighs, passionate outbursts of frenzy, sadness, etc. in rapid alternation.

In an analysis of a Duo for Violin and Piano of 1942 composed by Roger Sessions, Schubart (1946) wrote that it “begins with a musical gesture similar to that of the Piano Sonata, and creates a mode of quiet lyricism”. Correspondingly, a review of “A Dictionary of Musical Themes” by Harold Barlow (E.B., 1949, p. 273) uses the term “opening gesture” solely to refer to the melody at the beginning of Schubert’s D minor Quartet.

Sessions himself, in his essay “The Musical Experience of Composer, Performer, Listener” (Sessions, 1950), states that “A melodic motif or phrase is indeed a vocal gesture” (p.19), and elaborates:

Music is significant for us as human beings principally because it embodies movement of a specifically human type that goes to the roots of our being and takes shape in the inner gestures which embody our deepest and most intimate responses (p.19).

He comments further that “each musical phrase is a unique gesture and through the cumulative effect of such gestures we gain a clear sense of a quality of feeling behind them” (p.24). He suggests that the gesture may be communicated by a different means than the human body or the musical sound, the score: “the composer has attempted to indicate (I can find no better word) by means of a vastly complex system of symbols the essentials of what I have called a musical gesture” (p. .77). In contrast Waldbauer in an analysis of Bartók’s “Four Pieces” for two pianos (Waldbauer, 1960), uses the terms gesture and musical gesture solely to indicate intention, when writing that “Bartók in 1920 managed to shorten this section [the fugue] and yet at the same time made it convey a larger gesture” and “The pervading percussiveness [...] modifies the original musical gesture; it narrows the range of possible musical meaning by reinforcing one single aspect”.

The historical development of the idea of gesture in music suggests an interest in inherent qualities of musical phenomena that extend beyond the musical sound, into the human psyche and culture. The concept of gesture had been useful to describe in a single word a musical unit of meaning. With advancements of audio technologies and the possibilities they offered, the concept of gesture remained applicable to the new aesthetics of electroacoustic music. In early publications on the aesthetics of electroacoustic music, gesture is at the core of analyses and discussions. In 1985, Wishart defined gesture as an articulation, by an agent, of a continuum whose dimensions are pitch and timbre (essentially spectrum), and space (Wishart, 1996, p. 17, pp. 109-115). An example of this provided by Wishart is the description of how gestural information may be encoded to control Morton Subotnick's "Ghost Box", a voltage-controlled electronic musical instrument (p. 105). The gesture would be encoded by manipulating knobs in real time, controlling voltages that were recorded as audio signals in a tape. Afterwards, those signals could be decoded by the system and mapped to sound. This paradigm is still being used for the gestural control of audio workstations and electronic musical instruments, albeit taking advantage of digital technologies. The system that allows to record a gesture that can then be decoded and translated into sound, is analogous to the composer imprinting the gesture in the musical score, which can be then executed by a musician. This is consistent with the ideas expressed by Sessions. However, the remarkable difference is that the new technology allowed to literally, not metaphorically, encode the movement of their own hand.

The concept of articulation of spectrum and space in time as suggested by Wishart, was elaborated by Smalley (1986), hinting a definition of gestures as the spectral shapes (p. 62) or spatial trajectories (p. 91) of sounds that mirror instrumental and vocal sounds (p. 62), suggesting causality from human activity (p. 82), specifically that of the human body (p. 83). Smalley also touches upon, albeit tangentially related to gestures, the mimesis occurring in electroacoustic music as "musical materials and structures find resemblances and echoes in the non-musical world" (p. 64). Emmerson (1986) elaborated on the notion of mimesis in electroacoustic music, as "the imitation not only of nature but also of aspects of human culture not usually associated directly with musical material" (p.17). In the same text, Emmerson remarks the subjectivity in the decisions made by the composer of electroacoustic music when combining sounds. He states that "Loose terms such as 'gesture' may abound, but it is to this area, combining psychology of music with investigation of deeper levels of symbolic representation and communication, that future research must urgently be addressed". Emmerson's concern echoes the struggle of contemporary and past authors in defining gesture in more concrete terms.

The works mentioned until this point had been phenomenological, speculative, and subjective. Not that they are less valuable, at least for the purpose presented here, in the quest for a historically grounded definition of musical gesture. One of the first attempts of systematic research on musical gesture may be found in the work of Fay (1974). That work describes an

experiment that was conducted to investigate memory and attention when listening to music. A group of students were asked to compare the time experienced when listening to silence and music. The task was done two times, the second time the group was asked to clap when presented with silence. The results showed that both times a majority experienced the music as being longer in duration, than the silence. Also, the spontaneous clapping had a frequency of about 60 BPM (i.e., one clap per second), a figure consistent with observations by later studies (Fraisse, 1982; Styns et al., 2007). These results were not conclusive but led to several hypotheses and speculation about the experience of musical time, considering the interaction of memory, attention, and expectation. This reasoning resulted in a rationale for music analysis based on segmentation of musical patterns. Hence, a musical gesture was defined as a meaningful unit derived from the segmentation process. Later, Schneider (2010) elaborated on these ideas concluding that “A musical gesture thereby may exhibit properties known from Gestalt theory (e.g., completeness, distinctiveness, conciseness) yet the aspect of ‘movement’, and of temporal-dynamic organisation is often of special importance”.

The central ideas that define the concept of musical gesture presented so far have remained more or less unaltered, being at the core of a wealth of research on Embodied Music Cognition. Still, a noteworthy theoretical development in the understanding of the perception and cognition of musical gestures, is the incorporation of “chunking”. This may be seen as an extension of the segmentation process noted by Fay, whereby musical–bodily and sounding–gestures are perceived in different timescales. Shorter-scale gestures are grouped or fused together into larger entities (i.e., chunks) in a process that has been called *coarticulation* (Godøy, 2011). Furthermore, this process may integrate different sensory modalities (e.g., auditory, visual) to produce the chunks. Godøy (2014) gives as examples of coarticulation “the fusion of a rapid succession of tones and finger motion into what we perceive holistically as an ornament, or [...] the fusion of drum sounds and associated mallets/hand/arm motion into a rhythmical groove pattern”. This notion implies that gestures may contain smaller gestures, in a hierarchical structure of nested gestures, *hypergestures* (Mazzola, 2012; Mazzola et al., 2016, p. 168).

Take for instance a musical phrase that overall goes from a low pitch to a high pitch. That is a gesture, as its abstract meaning may be “going up”. The phrase may be decomposed in smaller chunks such as motives and into its atomic units, the notes. Also, the phrase may be part of a bigger structure, for example succeeded by a phrase going to a low pitch, forming an overall “arc” gesture. It may be the case that these sounds are performed by a musician that moves along with the sound. For example, a clarinetist might move the instrument upwards when playing an ascending motive or describing an arc when the melody is also an arc (Wanderley et al., 2005). Perhaps a more obvious example is the case in which a dancer would stretch the body upwards when the sound rises in pitch and depict an arc when the sound (e.g., a melody) rises and then decays in pitch. However, these are simplistic examples. It is not

always the case that the correspondences between sound and motion are evident. On the contrary, often they are complex. Furthermore, the perception of gestures and its hierarchical grouping, in either or both sensory modalities, is not straightforward. It depends much on context, including the subject (e.g., the person) and all the intricacies of individuality. Since a prime focus of this dissertation is bodily movement, the following subsection presents a succinct survey on research conducted towards the understanding of segmentation of bodily motion.

2.2 Segmentation of bodily motion

2.2.1 Human visio-temporal segmentation

The cognition of musical gestures requires a process that parses signals of sensory organs into meaningful chunks. For the case of signals carrying information about bodily motion, temporal segmentation is defined as the perception and cognition of distinct successive chunks in time. These perceived units are semantically meaningful, they are essential to understand what an observed subject is doing. Studies on perceived visual temporal segmentation have consisted of experiments in which people manually annotate the timing of segment boundaries in video. Newton (1973) employed recordings of a subject performing a sequence of actions such as “seated writing”, “standing up”, “walking”, and “lighting a cigarette”. Zacks et al. (2001) used video recordings of a subject performing the activities “making a bed”, “doing the dishes”, “fertilising a houseplant”, and “assembling a saxophone”. Hard et al. (2006) used animations of abstract figures interacting with one another and with static figures, where moving figures performed the activities “chase” and “hide and seek”.

These studies have asked the annotators to indicate boundaries of segments at different timescales, as motions may be described as a whole or its constituent parts. This is often referred to as *granularity*, a relative measure of detail in the description. For instance, a motion picture shows a person that walks to a chair, then sits down. The segmentation of this picture at a *coarse granularity* will result in two segments: walking and sitting. Each of those segments may be decomposed in segments of shorter duration, at a *fine granularity*. For example, the walking segment may be further segmented into each of the steps. Likewise, the sitting segment may be decomposed into motion that, once walking has ended, brings the body down to the chair, and another segment where the subject stays still while sitting. However, the segment in which the subject sits down may be considered a transition. Furthermore, the steps in the walking segment may be grouped such that each segment has a step of the left leg and the right leg, constituting segments at an intermediate level of granularity.

The results of the studies cited above indicate moderate agreement among annotators. Also, they observed that coarse-granularity boundaries matched fine-granularity boundaries, showing that the latter are sub-divisions of the former and vice versa. These findings are consistent with the theories expounded in the previous subsection. Kahol et al. (2004) did a similar experiment, in which two choreographers were presented with videos showing a dance routine and were asked to indicate segmentation points. The points annotated by one choreographer were significantly different to the other, with most segments of one being significantly longer than those of the other. This suggests that each choreographer had their own strategies for segmentation, thus it was not possible to establish that either was better or correct. Following the same experimental paradigm, Bläsing (2015) used stimuli composed of dance movements and annotators were professional and amateur dancers, and non-dancers. The results indicated that previous knowledge of the movement patterns had an influence on the resulting segmentation. The segments indicated by professional and amateur dancers tended to be of coarser granularity than those indicated by non-dancers. This suggests that the initiated in the art could group movement patterns into a cohesive idea, while those not knowledgeable would focus on the motion patterns without identifying links to cluster them in chunks. Likewise, Zacks et al. (2009) observed that perceived segmentation boundaries depend on context information when available, and on kinematics (i.e., movement features such as velocity, acceleration, amount of motion, etc.) if no context is available. Also, kinematics were found to correlate more with fine-granularity segmentation than with coarse-granularity segmentation.

2.2.2 Automatic temporal segmentation

The findings that kinematic features correlate with perceived segmentation at fine granularity and that those segments can be grouped in larger chunks, are crucial for the automation of segmentation. This is so because kinematic features can be measured by sensing technologies such as accelerometers or video tracking. The data from these devices may be processed in such a way that the kinematic features are grouped, emulating human perception. This process may be incorporated in a wide variety of applications that require the identification of patterns in signals. In what follows, a brief review is presented of research on automatic temporal segmentation methods developed for technologies that can measure human motion, such as optical marker-based motion capture and accelerometry. One key advantage of the former is that it can measure the position of limbs and torso. These can represent posture and its changes over time, from which time-derivatives (i.e., velocity, acceleration), and other characteristics (i.e., features) of motion can be computed. However, it requires an expensive and bulky apparatus that usually can only be accommodated in a laboratory. Accelerometry only measures acceleration, meaning less information about motion, but it has the advantage that sensors (accelerometers) are small, cheap, and reliable. These capabilities make possible

their incorporation into portable devices such as mobile phones and wearable devices. Example applications are the detection of activities (Noor et al., 2017) and of falls (Redmond et al., 2010), both in the context of assisted living for the elders (Ni et al., 2015). Accelerometers have also seen extensive use in the development of gesturally controlled digital musical instruments. A cornucopia of examples may be found in the NIME Proceedings Archive (n.d.). Temporal segmentation is one of the key processes to identify gestural patterns in the accelerometers' signals, which may be used in various ways for the control of musical sound, a topic elaborated in the next subsection. However, the lack of a method for fully automatic segmentation for digital musical instruments has been noted as an open problem (Caramiaux & Tanaka, 2013).

The implementation of automatic temporal segmentation is affected, among other factors, by the variability in human perception (if the system is meant to match human perception). However, the findings on perceived segmentation mentioned above are robust and have served as principles to devise automatic systems. Bodily movements at a coarse level of granularity have been called *gestures* (Mitra & Acharya, 2007) or *activities* (Ahad et al., 2008; Lara & Labrador, 2013). At a finer level of granularity, the movements considered to be the shortest indivisible units have been called *primitives* (Lin et al., 2016). Following this logic, Krüger et al. (2007) proposed a framework in which movements with semantic meaning are composed by shorter coherent units, thus formulating a heuristic for segmentation of human motion in two levels of granularity. Other studies have proposed segmentation in three levels of granularity. Bernard et al. (2017) proposed a model ordered from finer to coarser granularity, consisting of kinematic features, single patterns and groups of patterns. "Kinematic features", as outlined in the previous subsection, refers to any of the univariate time series that can be derived from motion (e.g., position or angle of each articulation of the body, velocity, acceleration, etc.). Dreher et al. (2017) proposed a model composed of perceptual granularities. In that model, coarse granularity is composed of activities (e.g., jumping, walking), medium granularity is composed by actions (e.g., step with left foot, step with right foot) and fine granularity is composed by motion primitives (e.g., lift a foot for a step, return the foot to the floor). However, there are no published studies that have extensively tested the perceptual validity of a fixed number of granularity levels. In fact, the number of perceptual granularities, as the phenomenon at large, in all probability is highly dependent on context.

Another challenge that research has faced is the different ways in which motion segments may be concatenated. One possibility is that there might be a moment between the end of one meaningful segment and the beginning of the next, in which motion does not correspond to either segment. This has been deemed to be a transition. For example, the segmentation method described by Krüger et al. (2017) excludes segments that are transitions between semantically meaningful segments. Another possibility is that segments are coarticulated (Meier et al., 2011). In the context of that study, the term coarticulation refers to the overlap of a distinct motion with the previous or the next. This meaning of

the word coarticulation is different from the one used in the previous subsection, which refers to the perception of gestures by integrating different sensory modalities.

A transition or overlap may be short enough to be perceived as instantaneous. Most studies mentioned here dealing with automated systems for segmentation of human motion have aimed to find segmentation boundaries that are instantaneous, even if in fact there were transitions or overlaps. This treatment may result in transitions detected as proper segments. For the case of overlaps, a segmentation boundary might be placed in the middle of the overlapping section, or the overlapping section may be identified as a segment, or even both overlapping segments may be merged.

Automatic segmentation has been considered a machine learning task, and as such, methods for segmentation can be classified as supervised or unsupervised. Supervised methods require examples of motion patterns to compare with the data intended to segment (Lan & Sun, 2015; Lv & Nevatia, 2006; Müller et al., 2005; Patrona et al. 2018; Salamah et al., 2015; Santos et al., 2015). Unsupervised methods detect motion units that are not known in advance (Barbič et al., 2004; Krüger et al., 2017; Zhou et al., 2013). These methods can be further classified as offline or online. Offline methods find segments taking into account the characteristics of the whole data. Online methods perform linear search on data, segmenting according to the similarity of observations (e.g., data samples) within a neighbourhood range. Also, online systems may be suitable for real-time applications, intuitively as long as computation of results is faster or as fast as the sample-rate of real-time data. Several online and offline temporal segmentation algorithms have been tested with data from accelerometers, as these sensors are suitable for a wide range of practical applications.

An example of unsupervised temporal segmentation was proposed by Gharghabi et al. (2019). The method evaluates the similarity in shape—but not in statistical properties—between all fixed-length windows within a bigger window whose length has to be given by the user. A segmentation boundary is recorded where the similarity is minimal. This method assumes that each segment will be composed of at least two instances of a periodic motion. Another approach is to pose the task as a multivariate change-point detection problem (Endres et al., 2011; Gong et al., 2014; Krüger et al., 2017; Zhou et al., 2013). Essentially, a change-point indicates a difference in statistical properties of the data within a sliding window (Aminikhanghahi & Cook, 2017; Fathy et al., 2019; Liu et al., 2013; Patterson et al., 2017). The size of the sliding window is a free parameter that adjusts granularity (i.e., timescale). Depending on the method other free parameters may be required to be adjusted. Zameni et al. (2020) described a method that efficiently finds segmentation boundaries in signals that may be highly dimensional. This method has initialisation parameters but no parameters to explicitly adjust granularity (i.e., timescale) or relevance (i.e., discarding boundaries that may be product of noise). The cited systems that were tested with data from triaxial accelerometers, segmented

activities that take at least a few seconds to complete. However, segments of dancing motion may range from less than a second to more than a few seconds.

To measure the effectiveness of segmentation algorithms, most published studies have relied at least to some extent on the classic measures *precision*, *recall* and *accuracy*, by comparing boundaries annotated by one or more people with computed boundaries. These measures work well for classification problems in which the options are either “match” or “not a match” between a computed boundary and an annotated boundary. Dreher et al. (2017) note that a computed segmentation boundary being only slightly different to the ground truth (i.e., an annotated boundary) should be counted as a match. This has been often solved by establishing a window around each annotated boundary, to allow for tolerance. A computed point is deemed to be a true positive if it lies within that window. This approach was used in the study by Zamani et al. (2020), for example. Dreher et al. proposed a method that involves a window weighted with a normal distribution. However, the problem with this approach is that the window’s width is fixed while there is no certainty that any given width will correspond to the true probability distribution for the occurrence of a boundary, for all boundaries. Following the observations made by the perceptual studies cited in the previous subsection, it is not possible to generalise the temporal length of the transition from one motion pattern to another. In contrast, the evaluation method used by Gharghabi et al. (2019) consists of a score that measures the temporal distance between each computed boundary and the closest annotated boundary. All the distances are added and then divided by the total time. However, this score does not penalise extra or missing computed boundaries, which is problematic as there is no certainty that the number of annotated and computed boundaries will always be the same. Lin et al. (2016) describe another approach for evaluation of results, in which all frames in the sequence of annotated segments are labelled and the number of frames in the computed segments corresponding to the annotated segments’ labels constitute the measure of similarity. This last method might be appropriate for classification of segments, but it might be too restrictive for evaluating only the boundaries. This is because boundaries of short false-positive computed segments (e.g., transitions between motion patterns) will break the continuity of parallel labelling resulting in a very high dissimilarity score. Notwithstanding, in a previous work (Mendoza, 2014), I described a similarity score that measures the distance between annotated and computed boundaries as in the method by Gharghabi et al., but also penalise missing or extra computed boundaries.

The challenges for the implementation of automatic temporal segmentation are as many as the opportunities for real-world applications. As such, there is vast unexplored territory that may provide fertile ground for the research on and utilisation of, automatic segmentation of bodily motion in human-machine musical interaction.

2.3 Music and emotion

2.3.1 Musical emotions

Musical gestures, whether they are realised as sound or as bodily motion, may carry emotional signification. This statement, summarising beliefs held throughout the history of humankind, is supported by empirical research conducted since more or less the second half of the twentieth century. Such research has found that musical emotions can be expressed, induced, and perceived (Gabrielsson, 2002; Juslin & Laukka, 2004). To clarify the definitions, “expressed musical emotions” refers to the emotions a musical performer (e.g., musician, singer) portrays in their performance. “Induced musical emotions” refers to a change or enhancement of mood as an effect of listening to music. “Perceived musical emotions” refers to the emotions that an observer (e.g., a listener) understands as being portrayed by the music. These three aspects of musical emotions are independent. For example, a musician might perform a musical piece to convey happiness, while such performance doesn’t make a listener to feel happy or to recognise happiness in the performance. The listener might recognise or feel a different emotion. Or they might feel nothing.

Several paradigms have been adopted to measure musical emotions in the three aspects mentioned above. The devices to measure induced emotions can be questionnaires (Zentner et al., 2008) or physiological measures (Coutinho & Cangelosi, 2011; Hodges, 2010; Koelsch et al., 2010; Västfjäll, 2010). Expressed and perceived emotions may not be measured by physiological measures, but only by self-report (Zentner & Eerola, 2010). For example, by declaring the emotion perceived after listening to a piece of music, or by using some device that allows to indicate the presence of an emotion continuously while listening to the music (Schubert, 2010). Likewise, the intended emotional expression of a musical performance may be assessed by asking the musician. Also, this may be accomplished by instructing the musician to play with a certain expression, usually indicated in musical scores using Italian words such as “allegro” (happy), “lamentoso” (mournful), and many others.

Regardless of the aspect of musical emotions and the device to measure it, the measurement scales that have been used most often because of their demonstrated reliability may be classified in two models (Eerola & Vuoskoski, 2011). The first is the discrete model and consists in assessing basic and possibly universal emotions such as fear, anger, tenderness, sadness, and happiness. The second model, sometimes referred to as “dimensional model of affect”, consists in assessing emotions in terms of dimensions, often two: valence and arousal. Valence is a continuum that extends from very negative to very positive, while arousal extends from “not excited” to “very excited”. These dimensions may be measured with discrete numerical scales, for example a “Likert scale” from -2 to 2 for valence (including zero) and from 0 to 5 for arousal. Of course, more fine-grained scaling may be used.

2.3.2 Auditory and visual perception

Several studies have measured the expressed and perceived emotional content in music. These studies have looked at musical features, in other words, information contained in the musical performances. This information can be extracted from the auditory and the visual component of the performance. The visual component is essentially the body of the musician playing an instrument. Diverse approaches have been employed to qualitatively and quantitatively assess the contribution of auditory and visual features, to the perception of emotions in musical performances. The quantitative approaches can be divided in two groups, the first being experiments in which musical stimuli is manipulated and then collected ratings of perceived emotions are examined by analysis of variance. The second approach is to select a candidate set of features of the stimuli and examine how well they fit to the rated perceived emotions either individually or in combination, using statistical modelling. The latter has been used in music information retrieval to predict perceived emotions, seeing great advancement in the last couple of decades. However, quantifying the relation between the movement of musicians and the perceived emotion in the music they play, has not received as much research attention.

The relations between perception of emotions and auditory musical features in different kinds of music have been measured and studied from a variety of perspectives. For instance, the perception of happiness in music is associated with a fast tempo and major mode (Dalla Bella et al., 2001; Juslin, 2000; Peretz et al., 1998), as well as high pitch and increased sound level (Lange & Frieler, 2018), and soft timbre (Juslin & Lindström, 2010). The same studies found that sadness, in general, has an inverse association to the features associated with happiness. Anger has been identified as being associated with fast tempo, increased sound level, high-frequency content (Juslin, 2000), sharp timbre and minor mode (Juslin & Lindström, 2010; Lange & Frieler, 2018). Fear has been found to be related to reduced sound level, staccato articulation, large articulation variability, soft timbre (Juslin, 2000), and minor mode (Juslin & Lindström, 2010; Lange & Frieler, 2018). Tenderness has been found to be related to slow tempo and reduced sound level (Lange & Frieler, 2018), as well as low pitch, major mode, soft timbre (Juslin & Lindström, 2010), and reduced changes in dynamics (Eerola et al., 2009). In an experimental study Eerola et al. (2013) found that the most important feature was mode, and that the relations between ratings of perceived emotions and musical features is mostly linear. Additionally, Gabrielsson and Juslin (1996) found that different performers and instruments yield distinct ratings of perceived emotions. Battcock and Schutz (2019) observed that mode predicted the most variance for perceived valence.

The relations between the movement of a musician and the emotions perceived by an observer have been studied to see if there is an effect of the visual component and to see which parts of the body have significant effects. Dahl and Friberg (2004, 2007) did experiments in which musicians performed marimba, bassoon, and soprano saxophone, in such a way that they expressed happiness, sadness, anger and fear. These performances were presented to

participants as video with audio, only audio and only video. The participants rated the perceived emotions in each performance. The video image was filtered to preserve the contours of the body and remove facial expressions. This is a paradigm that has been used to study communication and expression through movements of the body without the influence of facial expressions. One of the variants of this paradigm is the use of point-light displays, which is a visualisation of optical motion-capture markers often joined by lines forming a figure that resembles an anthropomorphic skeleton (Burger, Thompson, et al., 2013; Eaves et al., 2020; Vuoskoski et al., 2014; Wöllner & Deconick, 2012).

The analysis of responses in the experiment by Dahl and Friberg (2007) indicated that all emotions expressed by the musicians were recognised, except fear, and that the ratings for strength of expressed emotions were similar when performances were presented with or without audio. This suggests a strong effect of the visual component. For the performances of marimba, the head was found to play an important role in the communication of emotion. Also, participants gave subjective ratings of movement features such as amount of movement, speed, fluency, and regularity. The reported significant relations between emotions and movement features are as follows: happiness was associated with slow speed (bassoon), and large amount of movement (marimba and saxophone); sadness with little amount of movement (marimba), slow speed (all), and smooth fluency (marimba); anger with large amount of movement (marimba), fast speed (marimba), and fluency (all); fear with little amount of movement (marimba and saxophone).

It is worth to notice that the cited studies have analysed averaged data from participants' self-responses. In other words, participants normally would respond to a questionnaire asking the emotions they perceived while listening to music, then those responses would be averaged to get a rating of perceived emotions representative of the group. However, the variance within and among groups of responders has been a concern, albeit more pronounced for the measurement of felt emotions than for perceived emotions (Gabrielsson, 2002; Juslin, 2008; Peretz et al., 1998). Indeed, Vuoskoski et al. (2014) and Vuoskoski, Thompson, et al. (2016), found evidence for this by observing that the effect size of felt emotions was greater than that of perceived emotions. Lange and Frieler (2018) remarked that "means are only a crude approximation of the full distributions for the rating variables" after observing low to moderate agreement amongst participants rating perceived emotions in a range of musical stimuli. These and other studies (e.g., Hodges, 2010; Abeles & Chung, 1996) have suggested that this variability may be attributed to individual factors such as age, gender, musical training, music preference, current mood, race, social status, and personality.

Conversely, some studies have reported high agreement among participants rating perceived emotions. For example, Eerola et al. (2009) used film music to minimise inter-rater variability, assuming that this genre is intended to express clear emotional content to a large audience. Furthermore, they devised a selection protocol in which a panel of experts selected 360

excerpts of soundtracks representative of distinct discrete emotions and quadrants of dimensional affect. Then the 110 excerpts with highest ratings were selected to be used in a perceptual experiment with non-experts, achieving high inter-rater consistency measured by Cronbach's alpha (over 0.99). The same stimuli were used by Eerola and Vuoskoski (2011), resulting in Cronbach's alpha over 0.88 for the rating of perceived emotions. In another study measuring perceived emotions, Eerola et al., (2013) used composed stimuli, yielding Cronbach's alpha over 0.92.

Lange and Frieler (2018) also observed that the measurement of inter-rater agreement may be confusing. For example, Schedl et al. (2016) used classical music as stimuli for responses on perceived emotions, and deemed the agreement was low as Krippendorff's alpha (another popular measure for inter-rater agreement) was less than 0.4. Friberg et al. (2014) measured perceived energy and valence in ringtones, obtaining Cronbach's alpha over 0.9 but a mean correlation ranging from 0.42 to 0.57. Cronbach's alpha, Krippendorff's alpha and the mean inter-rater correlation measure different aspects of the broader concept "inter-rater agreement" or "consistency". The interpretation of them is not trivial and to date there is no systematic study on the details of these measures when applied to measure the variability of perceived emotions.

Beyond the context of music perception, the measurement of correspondences between movement and perceived emotions has been widely studied for the purpose of automatic emotion recognition (Ahmed et al., 2019; Kleinsmith & Bianchi-Berthouze, 2012; Noroozi et al., 2018; Saganowski et al., 2020; Sapiński et al., 2019). Among these studies there is also wide discrepancy, and the description of movement features that correspond to perceived emotions tend to be imprecise. It has been found and discussed that main factors for variability may be culture and gender (Noroozi et al., 2018). The cited studies reviewed research that found correspondences between movement descriptors in non-musical contexts and the emotions investigated by Dahl and Friberg (2007). These are summarised as follows: Happiness has been associated with arms open and moving, legs open or in parallel; sadness with low energy, head forward or trunk forward; anger with high energy and limbs spread; fear with head straight or bent back, and breath held. Although these findings may be useful in general contexts, most of these features are not relevant in the context of performing musical instruments due to constraints imposed by the playing techniques. In general, a musician uses the hands and arms mostly to perform the movements required to produce music and secondarily for those not required to produce sound (often referred to as "ancillary gestures"). Therefore, the arms may be not used for expressive intention if that prevents proper execution of the instrument. The expressive movement from which an observer may perceive emotional content should be from other parts of the body or somehow coincident with the movements used to play the instrument. For example, a pianist or a violinist cannot extend the arms to express anger, as the hands are required to be near the keyboard in the case of the pianist and holding the violin and moving the bow in the case of the

violinist. Nonetheless, it may be argued, for example, that lifting the elbows to enhance body width can be a surrogate for expressing anger. Other movements of the whole body, torso, head, and hands may convey emotional information visually. A pianist that plays sitting in front of the piano has considerable freedom for moving the torso and head. For example, arching the back and leaning close to the keyboard may convey a sense of intimacy. A violinist that plays in standing position has considerable freedom for using legs and torso. For example, in a simultaneous movement to the beat, the legs can be bent, and the torso can be swayed to express joy.

2.3.3 Relation with personality and bodily motion

The studies that have investigated perception in and induction of musical emotions, both auditorily and visually, have examined responses summarised as averages and it has been acknowledged that this approach is a rough representation of a sampled population. Several causes have been suggested for inter-subject variability when measuring expressed, perceived, and felt emotions (Abeles & Chung, 1996; Gabrielsson, 2002; Hodges, 2010; Juslin, 2008; Vuoskoski, Thompson, et al., 2016; Vuoskoski, Gatti, et al., 2016). Among these causes are individual characteristics of performers (e.g., musicians, dancers) and raters (i.e., those who report perceived or felt emotions).

The relationships between people's individual characteristics and musical emotions have been studied in various ways. Individual characteristics may be examined in terms of personality traits and measured with a questionnaire. Musical emotions have been observed in terms of perceived emotions in music and felt emotions when listening to music (also referred to as emotions induced by music). These can be evaluated with a questionnaire (i.e., self-report) or by measurement of physiological activity. For example, Gerra et al. (1998), described an experiment in which participants were presented with classical and electronic dance music, while several physiological and psychological measurements were recorded. Results showed that after listening to both kinds of music there was a change in emotional state. However, only after listening to electronic dance music, changes towards a negative mood and release of stress hormones had a positive correlation with "harm-avoidance" and a negative correlation with "novelty-seeking" temperaments of Cloninger's personality scales (Cloninger, 1987). Another study, conducted by Park et al. (2013), looked at how "Big Five" personality traits (Extraversion, Agreeableness, Conscientiousness, Neuroticism and Openness) modulate neural correlates of musical emotion processing. In that study, participants completed the NEO-FFI questionnaire of Big Five personality traits (McCrae & Costa, 2004) and, while being scanned by a Magnetic Resonance Imaging device, listened to music expressing different emotions. The results showed significant correlations between brain activity and both Neuroticism and Extraversion as a response to music expressing happiness and fear, respectively.

Other studies have evaluated musical emotions, perceived, or felt, solely by means of self-report. Vuoskoski and Eerola (2011a) conducted an experiment

in which participants completed the Big Five Inventory (BFI) personality questionnaire (John & Srivastava, 1999), the POMS-A questionnaire to evaluate mood (Terry et al., 2003), and rated music in terms of perceived discrete emotions (happiness, sadness, anger, fear, and tenderness). Ratings of perceived sadness correlated positively with Neuroticism and negatively with all other traits except Conscientiousness. Also, mood was associated with mood-congruent biases in perceived emotions, moderated by Extraversion. In another experimental study, Vuoskoski and Eerola (2011b) asked participants to complete the BFI and to rate emotions felt when listening to music. Ratings in terms of three-dimensional affect—valence, energy, and tension—yielded more consistent and differentiated responses than discrete emotions. However, the relation between personality and music-induced emotions was stronger for discrete emotions. In addition, Extraversion was significantly correlated with experienced happiness, sadness, and tenderness. In a similar vein, Liljeström et al. (2012) asked participants to listen to music and indicate if it was familiar, how much they liked it, which emotions they felt and how intensely. Participants also completed the NEO-PI-R questionnaire for Big Five traits (Costa & McCrae, 1992). A positive correlation was observed between Neuroticism and the experience of negative emotions, while for all other traits that correlation was negative. This is consistent with the results of Vuoskoski and Eerola (2011a). Furthermore, the correlation between personality traits and ratings of emotion intensity was moderately positive for Agreeableness, Extraversion and Openness, negligible for Conscientiousness, and weakly negative for Neuroticism.

The studies mentioned in the previous paragraphs reveal distinct relationships between personality traits and the perception and feeling of emotions in music. Trait Openness is a special case as it has been suggested to be related to transient emotional responses (colloquially referred to as “chills”) to music and other expressions facilitating aesthetic experiences (McCrae, 2007). Nusbaum and Silvia (2011) tested this hypothesis in an experiment and found that Openness was the only Big Five trait that significantly predicted such responses as an effect of music listening. Furthermore, Silvia et al. (2015), found a significant and moderate correlation between Openness and the feeling of a *profound experience* (also referred to as “awe”) when listening to music, while the correlation with the other traits was much lower.

While perception and experience of musical emotions may be observed by means of physiological measures and self-report questionnaires, it may also be observed by measuring characteristics of spontaneous movement to music, namely the embodiment of emotions. Burger, Saarikallio, et al. (2013) did an experiment in which participants were asked to spontaneously move to music (i.e., dance) while they were recorded with a motion-capture system. Bodily features were extracted from the motion-capture data, for example the torso’s tilt and rotation, floor area used, and acceleration of different body parts. Another group of participants rated the perceived emotional content of the same music in terms of both dimensional affect—arousal and valence—and

discrete emotions happiness, anger, sadness, and tenderness. A correlational analysis between bodily features and emotion ratings revealed significant relations between them, even though the two datasets were collected independently of each other and from different groups of participants. Using the same data, Burger, Polet, et al. (2013) found a mediation effect of emotion ratings on the relation between bodily features and features of the music, such as energy and activity in the low and high frequency ranges, attack time, and note density. That study also used Big Five personality scores of the dancing participants and found a moderation effect of Extraversion on the relation between head acceleration and the activity of low frequency audio. Furthermore, Conscientiousness was found to be a significant moderator of the relation between note density (i.e., notes played per unit of time) and movement fluidity.

Using the same motion-capture and personality data as Burger, Polet, et al. (2013), Luck et al. (2010) found that Extraversion was directly related to the level of overall acceleration. This was later confirmed in a study with different data by Carlson et al. (2016), which also found that responsiveness to changes in tempo correlated positively with Conscientiousness and negatively with Extraversion. This suggests that conscientious people were compelled to follow tempo accurately while extraverts preferred to divert and follow their own beat. Bamford and Davidson (2019) measured the time to entrainment (i.e., the alignment of the periodicity of the movement of the body to the beat of the music) of participants that had completed the BFAS Big Five questionnaire (DeYoung et al., 2007) and the Empathy Quotient questionnaire (Wakabayashi et al., 2006). Results showed that Empathy and Agreeableness correlated negatively with time to entrainment. In other words, the more empathic or agreeable a person is, the faster (and arguably more easily) they will align their dancing motion with the beat of the music.

While these studies have identified significant relations between dancing motion and personality, the predictive power of the produced models and correlations is at best modest. However, a later study by Agrawal et al. (2020) traded the interpretability of bodily features for greater prediction power. Instead of using bodily features extracted by manual selection (e.g., speed or acceleration of body parts, or the distance or angle between them) or by dimensionality reduction (e.g., vertical or lateral speed), they used the covariance among the speed of body parts. As a result, predictions for all Big Five personality traits were remarkably close to their scores as measured by a questionnaire. In summary, the cited studies provide evidence that embodied responses to music are related to personality traits and to musical emotions. However, none of these studies have examined the relation between personality traits and the extent that musical emotions may be embodied.

2.4 Technology for making music with broad bodily motion

2.4.1 Responsiveness of musical instruments

Musical instruments are usually designed to be controlled with fine movements of hands and fingers, as they afford precision and speed. These qualities are often described as the foundations of responsiveness, believed to be indispensable for musical expression. The instrument thus becomes an extension of the human body. Following these ideas, the capability of musical interaction is thought as uniquely human, despite the advancements in technology for automatic music composition, machine learning and in general of artificial intelligence. Such beliefs may be challenged, considering that newer technologies may provide a broad range of opportunities for musical interaction that cannot be achieved with non-electronic musical instruments. Can we engage in musical activity with broad movements of our body, without the need for precision or speed? Could a machine learn and understand these movements as gestures, in such a way that it can interact musically? Could this learning be a continuous process such that the machine learns the gestures by itself?

Two and a half decades ago, Moore (1998) used the term “Control Intimacy”, referring to the cohesion between a musical instrument’s output and the ability of the musician using it. He argued that this cohesion depends on the time of the interaction between instrument and musician, which facilitates the translation of subtle gestures of the musician into sound that is emotionally expressive. This argumentation was based on the observation that such properties are present in the human voice acting as an instrument, and in most common musical instruments. Overall, this concept was used to support the desirability of electronic musical instruments that have a low time of response.

Following the principles outlined by Moore, a response time approaching zero has been adopted as a goal by many designers and builders of electronic musical instruments (e.g., Bosi & Jordà, 2012; Jordà, 2002; Moro & McPherson, 2020; Trolland et al., 2022). In the same vein, Wessel and Wright (2002) observed that many of these systems had a response time of up to 7ms, which lead to propose 10ms as the maximum acceptable. Over the years this number has been held as the standard (Jack et al., 2018; McPherson et al., 2016). However, a case may be made against this by considering non-electronic instruments that have slow response time. This happens with bass instruments that necessitate time to resonate, such as large wind instruments, or instruments that have mechanisms that impose time between the action and the sound, such as the piano. Musicians that play these instruments learn to play ahead of time.

Dahl and Bresin (2001) did an experiment in which the response time of custom-made digital musical instruments was manipulated. Musicians were asked to play these instruments along with a metronome and the difference of the metronome and played onsets was measured. The results showed that

musicians can adapt to play ahead up to 55ms. This is close to the findings of Rasch (1981), who observed that the standard deviation of the time difference of notes that are to be played simultaneously by musicians in a small ensemble can be up to 50ms. An extreme case of measured response time corresponds to the grand piano, which can be up to about 200ms when played very softly (Goebel et al., 2005). In between there is a variety of response times that have been found to be adequate, depending on the musical context (Lago & Kon, 2004). Nonetheless, the lowest acceptable response times are found for percussive motion, such as when tapping to a steady beat, at around 4ms (Rubine & McAvinney, 1990).

Similar aspirations have been present when designing electronic musical instruments, posing a challenge for the case of digital musical instruments (DMI) that are made to recognise gestures “in the air” using machine learning techniques. For example, a musician wears, holds, or stands in front of, a device that may sense position or motion. The musician makes a gesture in free space, for example describes a circle with the head, or wiggles a hand, or stands in a particular pose. This is called “training”. The DMI would learn these gestures and recognise them when they are performed. The recognition of a gesture can be mapped to a musical action, such as triggering a sound, activating an effect, etc. (e.g., Gillian, 2011). There are several challenges with this paradigm, the most salient being timing. To recognise a gesture, usually a machine learning system must firstly observe the whole motion, then process the information and then output the result (i.e., which gesture is recognised, out of those that had been learned). Only then an action can be triggered.

Several advancements have been made to reduce the processing time of DMI, often referred to as “latency” (McPherson et al., 2016; Wang, 2021). However, it is more challenging to work around the fact that a gesture must be observed in its entirety to be reliably recognised. This might not be a substantial problem when the goal is to recognise static gestures, for example a bodily posture or a hand sign. The challenge becomes evident when the goal is to recognise and use musically, gestures that take some time to perform, such as the aforementioned circle or wiggle. These are often called “continuous gestures” (Gillian, 2011).

2.4.2 Machine learning of continuous gestures

The matter of timing in the recognition of continuous gestures by machine learning systems for musical applications has not been comprehensively studied, except towards the making of DMI that use percussive gestures such as tapping (Gillian & Paradiso, 2012) and “air drumsticks” (Dahl, 2015; Trolland et al., 2022). Nonetheless, machine learning of continuous gestures has been used in numerous musical applications (NIME Proceedings Archive, n.d.). Chiefly two algorithms and variations of them have been extensively used to recognise continuous gestures, regardless of the sensing technology: Dynamic Time Warping (DTW) (Gillian et al., 2011) and Hidden Markov Models (HMM) (Bevilacqua et al., 2010). Both can estimate the likelihood that a gesture being

performed corresponds to a gesture that has been learned in the training. Nonetheless, these algorithms need to be trained with individual gestures. To accomplish this, the beginning and ending of a gesture need to be explicit. This task, called segmentation (as described in the previous subsection), may be executed by an external mechanism dictating the change-point within a stream of data, for example the user pressing a button (e.g., Merrill & Paradiso, 2005) or with pauses between gestures (e.g., Bernier et al., 2013; Murad et al., 2017). Segmentation can also be performed in non-real-time (i.e., offline), for example by manually editing the boundaries with a graphical user interface.

Alternatively, a probabilistic model may predict the occurrence of a change-point and the following gesture, provided that these gestures have been previously learned (e.g., Martin et al., 2020). However, to date there is no known method to reliably predict the occurrence of a change-point given any previous history of performed change-points, or without any prior knowledge at all. While these constraints have not prevented the use of these algorithms in new interfaces for musical expression (e.g., using only static gestures), the ability of a machine to recognise and learn static and continuous gestures without explicit training, but rather as a discovery process, would allow a more fluid musical interaction between human and machine. Thus, unsupervised segmentation has been identified as one of the most important challenges in gesture recognition (Escalera et al., 2016). This, combined with the acknowledged importance of computational modelling of gestures as the vehicles of emotional content (Camurri et al., 2001), has provided motivation for much of the research described in this dissertation.

2.4.3 Sonification of bodily motion for sports and healthcare

Miniature sensors, wearable devices and mobile technologies can track daily activity of people, both in extent (i.e., amount of movement) and type (e.g., walking, sitting). This capability has been utilised as a behavioural change technique (Michie et al., 2013) in interventions to promote a healthier lifestyle, increase physical activity (i.e., net amount of movement) and reduce sedentary behaviour (i.e., time of inactivity) (O’Keeffe et al., 2020). These technologies may be effective aids in interventions to increase physical activity and reduce sedentary behaviour (Larsen et al., 2022), but only in the short-term. Long-term adherence is still a major challenge (Brickwood et al., 2019; Buckingham et al., 2019; Cajita et al., 2020; Creaser et al., 2021; Liu et al., 2020). Recent reviews suggest that more engaging methods are needed to effectively produce a change in behaviour (Wang et al., 2022). Sonification is a potential strategy to increase long term engagement and adherence, especially since it has been shown that the temporal dynamics of human motion and activity are similar to that of music (Chastin & Granat, 2010; Levitin et al., 2012).

Sonification is the representation of data with sounds. A system that produces sonification of data from motion sensors attached to the body may be seen as a special case of a gesturally controlled DMI. Several studies have explored the use of real-time sonification of movement to aid sports

performance and rehabilitation (Schaffert et al., 2019). Ley-Flores et al. (2021) conducted a small-scale study to explore how sonification of exercise with metaphorical sounds affects body perception. They found that some sounds do affect body perception such as feeling strong, further increasing the amount of physical activity of the participants. Other studies investigated presenting activity patterns as musical sound to raise people's awareness about their behaviour. For example, Krasnoskulov (2019) developed a system to present data as musical sound. In that study, data measured by an accelerometer and an optical heart-rate sensor were used in their raw form such as the number of steps per minute, heart-beat pulse, and after classification into events such as walking, running, and sleeping. These were mapped to musical parameters such as pitch, timbre, tempo, space, and loudness. This form of musical sonification is rather direct and may not result in a clear representation of discrete events such as those of normal daily life. Towards that end, some studies have considered segmentation of data so that the temporal relations of events in data are clearly reflected in the sonification.

Last and Usyskin (2015) developed a sonification paradigm that segments data into a user-defined number of segments. They tested the ability of the produced sonifications to convey the desired information and found that most users are able to decode the intended information. Along the same line, Vickers and Höldrich (2019) defined segments representing domain-relevant characteristics using zero-crossings of a one-dimensional data stream, which then were mapped to sound. These studies show that sonification is effective in conveying activity data, and that temporal segmentation may be a relevant part of the process, as it allows for mappings between data and sound that produce clear representations of data aligned with the temporal structure of motion behaviour. However, the temporal segmentation methods used by the mentioned studies have important limitations, as they are based on threshold, zero-crossings, or clustering. All these methods require careful calibration of input parameters and may not generalise well when patterns in data are multidimensional. Addressing these limitations would greatly benefit the design and implementation of systems for sonification of human movement, in essence, gesturally controlled DMI.

2.5 Opportunities for research

The preceding background has offered an overview of the topics covered in this dissertation, hinting at possible courses of action to expand the existing knowledge. These are recapitulated as follows:

Music is evidently a phenomenon that involves the human body immersed in an environment. Music, then, while having sound as its main component, is made up of multiple interactions among people by means of different sensory modalities (i.e., auditory, visual). The information exchanged in these interactions is organised in a hierarchical temporal structure of gestures

embedded in gestures. All these elements may be further scrutinised to formulate a model of musical interaction.

The temporal structure of visually perceived human motion is hierarchically organised such that short units of motion are cognised by perception of their kinematic properties. These are further grouped into larger entities. This fundamental quality can be utilised to model human temporal segmentation and emulate it by automatic systems. This may be accomplished by having a motion sensor to measure kinematic characteristics of bodily motion and applying some process that finds temporal regularities. These temporal regularities, in turn, may be subsequently grouped into larger chunks. However, the emulation of human perception presents many challenges, one of which is the variability depending on context. One seemingly promising avenue for research is the incorporation of automatic temporal segmentation of bodily motion in systems that make musical sound based on broad bodily movement, as this paradigm has not been extensively explored before. Such systems may see application in gesturally controlled digital musical instruments in the broad sense, including systems for sonification of bodily motion that may be useful in healthcare and sports.

The modelling of perception of musical emotions is also nontrivial and indispensable for a better understanding of musical interaction. In the first place, different sensory modalities (e.g., auditory, visual) may have different contributions. The detailed measurement and identification of which characteristics of auditory and visual perception contribute to the perception of emotions, is a possible direction for future research. Second, it is of interest to investigate the sources of variability in the perception of emotions portrayed by music. In that regard, personality traits are a measurable aspect of individuality that has been shown to be related to the perception and expression of emotions. In particular, the relations between personality and the extent that emotions are reflected in dancing motion, has not been measured before.

The following section describes how these opportunities for research have been addressed in this dissertation.

3 AIMS OF THE RESEARCH

The overall aims of the research reported in the articles included in this dissertation, were to advance the understanding of the relationships between the structure of music and the movement (including posture) of the human body, and to use this new understanding to devise novel technologies for making music. Towards that end, several studies were conducted, each focusing on a specific problem. The studies and their corresponding articles are grouped into three broad areas: theory, measurement, and application. The last two areas may be further regrouped into broad topics (FIGURE 1).

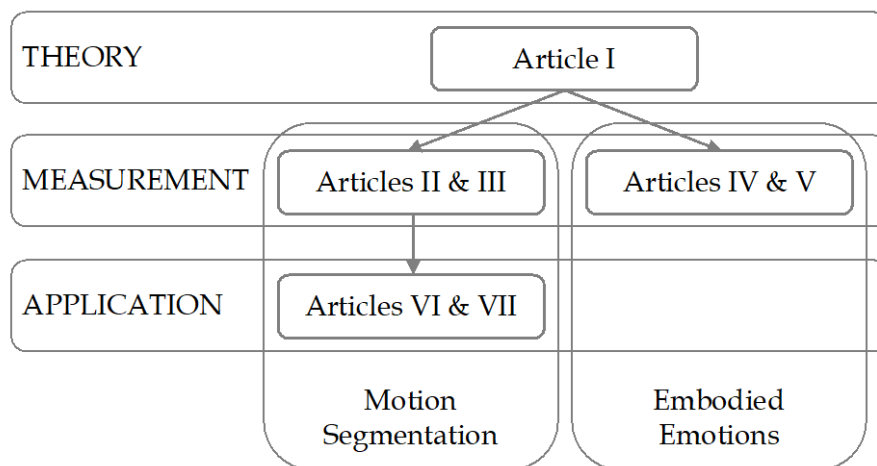


FIGURE 1: Overview of the articles included in this dissertation.

Theory:

The motivation for Article I was to establish a framework to investigate the correspondences between bodily movement and musical structure. However, since much work had already been done towards that purpose, it became necessary to focus into a less explored facet of the problem. Accordingly, it was decided to look at the broadest hierarchical level of musical structure, proposed to be musical interaction constituted by musical gestures. In addition, it was decided to incorporate human and machines as being the agents of the interaction. These decisions would make the resulting framework useful to approach the quantitative study of the phenomena (i.e., measurement), which would in turn inform the development of technology for music (i.e., application).

Measurement:

This part of the research is further divided in two parts. The first part of measurement was concerned with modelling the parsing of musical gestures, which are structural constituents of the musical interaction framework proposed in Article I. The goal of the study reported in Articles II and III was to develop a method for automatic segmentation of data from motion sensors. The method to be used was required to infer gestures from motion and posture, without the need to explicitly tell the system when the execution of those gestures begin and end. A further goal was to formulate the method to perform in real time (i.e., online). This capability would make it suitable for musical interaction systems such as digital musical instruments controlled by bodily gesture. A long-term vision for these capabilities was the design of systems for musical interaction that operate without supervision, enhancing their capabilities of gestural agency as proposed in Article I.

The second part of measurement is concerned with musical emotions. In light of the framework proposed in Article I, musical emotions are dimensions of the gestural information that is exchanged among agents of musical interaction. The study reported in Article IV was aimed to measure and model the relationships between the movement of musicians playing musical instruments, and the perceived emotions when observing them in three conditions: auditory, visual, and audiovisual. This facilitates the assessment of the contribution of the different perceptual modalities and parts of the body to the perception of emotions in musical performance. Article V corresponds to a study that aimed to measure and model the relations between personality traits and the extent to which people embody musical emotions when spontaneously dancing to music. This contributes to assess the effect of the individual characteristics of agents participating in musical interaction.

Application:

The studies reported in Articles VI and VII were aimed to develop applications within the conceptual framework outlined in Article I, and of the segmentation method described in Articles II and III. Concretely, the studies corresponding to Articles VI and VII focused on producing proof of concepts to demonstrate the feasibility of the applications. The study corresponding to Article VI aimed to develop a system to take data from a hand-held motion sensing device, segment the motion in real time, and use the segmentation data to control musical sound. This paradigm was thought of as a first approach to systems capable of fully unsupervised learning of musical gestures. These systems may be embedded into advanced musical instruments, or rather musical agents. The study reported in Article VII aimed to apply a multigranular version of the segmentation method, into a system that automatically produces musical sonifications (i.e., short pieces of music) representing daily activity recorded with sensors attached to the body. The motivation for the development of this system was its potential use as an aid in public health interventions towards a healthier lifestyle, by raising awareness of people's daily activity in an engaging way.

4 METHODS AND RESULTS

4.1 Theory

4.1.1 Article I

This article discusses the interaction between humans and machines with the purpose of making music. It first presents a generalised model of a musical instrument, simply referred to as a *machine*. Then an analogy is made between the model of the musical machine and a model of the human with whom it interacts. These models are composed by modules that represent processes, which communicate through signals. The models are also represented as greater modules that communicate with each other in a network of human-machine musical interaction (FIGURE 2).

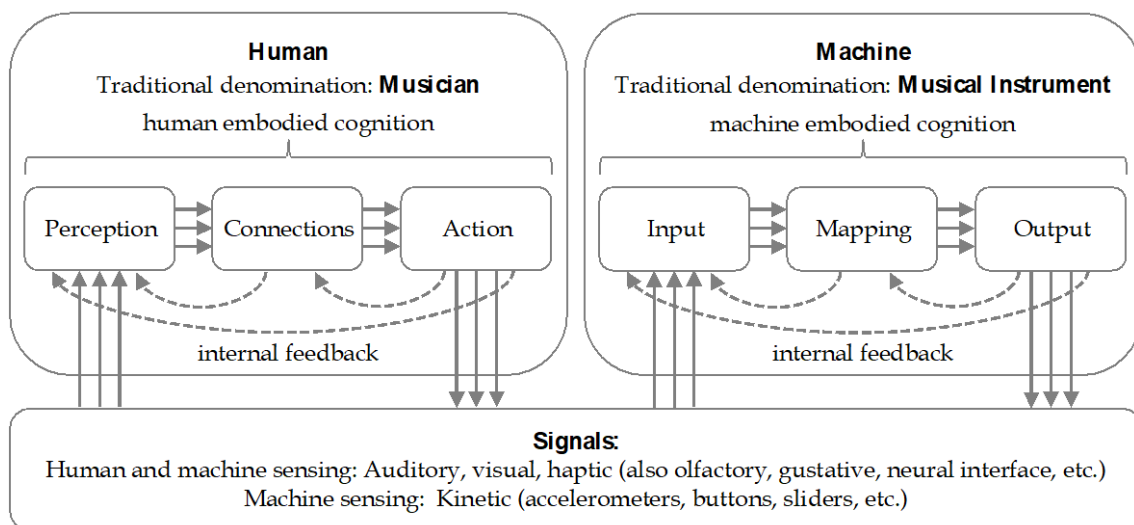


FIGURE 2: Human-machine musical interaction.

The discussion is formulated in light of scholarly literature mostly related to the development of electronic musical instruments, as they have seen great development and are a logical step after non-electronic musical instruments. Take, for instance, the guitar followed by the electric guitar and then the electric guitar as a controller of software. The design of electronic musical instruments has progressively diverted from the design of non-electronic musical instruments, and the interaction of human beings with these musical machines has changed in turn, however still retaining some of its fundamental characteristics.

Musical machines comprise non-electronic and electronic musical instruments. The generalised model of musical machines is composed by three modules. The first module is called "Input", and it receives signals into the machine. This module executes signal acquisition by means of one or more input devices, for example controllers such as a keyboard, knobs, and buttons. The second module, "Mapping", processes the input signals and its resulting signals flow to the last module, "Output", which produces the outgoing signals. The signals going into the input, emanating from the output, and communicating the modules, may be representative of different sensory modalities: auditory, visual, or haptic. An advanced electronic musical instrument may have a microphone at the Input, enabling the acquisition of auditory signals. Such an instrument may have a Mapping module equipped with a machine learning process whose output is connected to a synthesiser and a visual display at the Output. The human user (e.g., a musician) is connected to the first module by means of the controller. Using this interface, the human transmits an action, also called *gesture*, to the instrument. This action is a signal, which flows toward the final module being affected by the various processes. This conceptualisation works also for non-electronic musical instruments. For example, the Input for the violin is the action upon bow and strings; while Mapping is the action executed by the musician upon strings, fretboard, bridge, and body; Output is the result of these actions. In this scenario, not all signals flow from the Input to the Output. In a musical instrument, most possibly an electronic one, there might be internal feedback signals going from the Output module back to the Input module or into the Mapping module without leaving the Output module, enabling the musical machine to monitor its own internal behaviour.

The human being that interacts with a musical machine may be represented using the same logic as for the machine. The Input, Mapping, and Output modules that constitute a musical machine become Perception, Connections, and Action. This representation resembles the Cartesian model of the human mind composed by Perception, Cognition, and Action (Armstrong, 2006; Hurley, 2002). However, the proposed models follow a logic akin to the viewpoint of Embodied Cognition (Anderson, 2003) and in particular to the viewpoint of Embodied Music Cognition (Leman, 2008). Generally, the embodied cognition views consider cognition as an enactive process involving the body and its environment. For the case of the proposed models, body may

be understood as the triad of modules and the environment may be understood as the network of humans and machines connected by signals flowing among them. Furthermore, the intermediate modules in both models are actively modifying the signals flowing in an enactive direction: toward Action in the human and toward Output in the machine. Additionally, as in the model for the musical machine, the model for the human has internal feedback signals connecting the modules. These signals represent, for example, the modification of a perception organ's behaviour triggered by an action, with or without the mediation of the brain.

The generalised model of musical machines and the model of the musical human are highly resemblant of each other. When human and machine interact, the Action module of the human is connected to the Input module of the machine, and the Output module of the machine is connected to the Perception module of the human. In the resulting loop, the signals are continuously updated by the human and the machine. Thus, human and machine have the potential to become agents exerting influence upon each other to produce music. These agents' behaviour is affected by the conditions imposed, demanded, or proposed by other agents (Bown et al., 2009; Gurevich & Treviño, 2007).

The musical interaction process may start with exploration and discovery of the intentions of the other agents, gradually turning into an objective-based task as a musical aesthetic emerges (Caramiaux et al., 2014). In this way, the machine resembles an entity – more a musician than a musical instrument (Van Nort, 2011). The evolution of these agents requires some internal adaptation. In the model of the human, the Connections module is dynamically affected by reasoning and experience, to satisfy demands from the musical environment or to accomplish individual musical goals. The machine can also go through such a process, as the mapping module can be affected by generative algorithms and machine learning.

The influence that an agent exerts over other agents within a musical ecosystem, by means of the musical gestures carried by the multimodal signals, may be called *gestural agency*. The extent of this influence is a means of power that an agent has on shaping its musical environment, including the behaviour of other agents. A conservative view on this has the human in possession of most of the power (i.e., a musician uses a musical instrument as a tool), but we can see that in the system proposed here, the human is more a participant than a user (Kaipainen et al., 2011). In this way the whole musical ecosystem is enactive as the production of signals is linked to a function, a role of each agent in relation to the other agents (Matyja & Schiavio, 2013).

In sum, this article presents a generalised model for a musical machine that resembles a model for human embodied cognition. Both models depict enactive agents interacting in a musical ecosystem. These agents are connected by signals that carry gesture. Gesture is the means that an agent has to exert influence over other agents to produce a musical result. This understanding of musical interaction between humans and machines fits well to traditional and

newer technologies for making music. In practice this serves as a framework to analyse musical interactions that integrate humans, traditional musical instruments, and newer electronic musical instruments.

4.2 Measurement

4.2.1 Articles II and III

These articles describe the modelling of perceived segmentation of bodily movement induced by music. The modelling was based on the detection of change-points in bodily motion captured by sensing technology. The detection of change-points in motion data can be seen as equivalent to novelty detection, which is the identification of abrupt changes in data by a system, without training of the system (Markou & Singh, 2003). Foote (2000) described an algorithm suitable for finding segmentation boundaries in musical audio signals. This algorithm was chosen to be tested for the segmentation of motion data. The choice was based firstly because it is possible to implement the algorithm to be used in real time. This capability would be useful in the design of systems for control of sound or music, for example a gesturally controlled digital musical instrument. The second key capability of the algorithm is that it can be adjusted to detect segments at different timescales (*granularities*). This capability would allow to fine-tune the system to match the timescale of perceived segments.

The segmentation algorithm exploits the characteristic checkerboard patterns that can be observed in a self-similarity distance matrix (FIGURE 3b) of a motion signal (FIGURE 3a). Elementwise multiplication of this matrix with a checkerboard matrix of the same size, results in a novelty score that indicates the rate of change in the data (FIGURE 3c). The peaks of the novelty score (encircled data-point in FIGURE 3c) indicate change-points, equivalent to segment boundaries (segmented vertical grey line in FIGURE 3a). The granularity of the novelty score is adjusted with the width of the distance matrix and relevant peaks can be selected over a threshold (θ in FIGURE 3c).

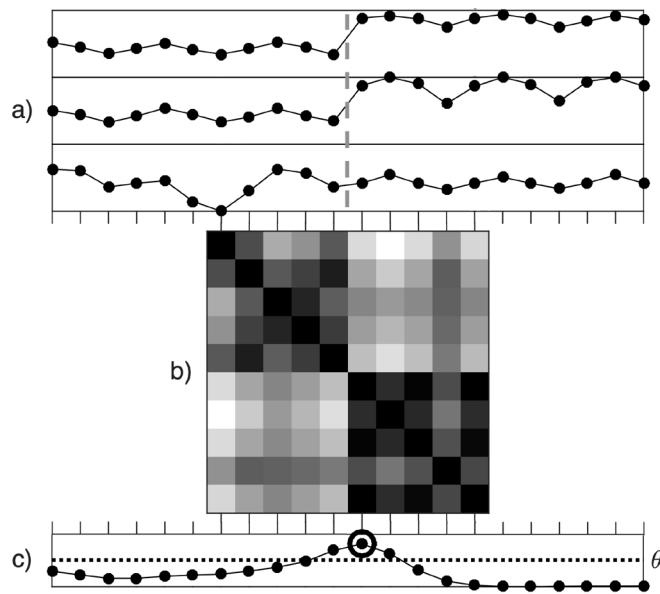


FIGURE 3: Online temporal segmentation algorithm.

The algorithm was originally formulated to work offline, meaning that it requires the full extension of recorded data (Foote, 2000). However, it has been shown that the algorithm may be implemented to work online (Schätti, 2007), meaning that computation is carried out as data enters the system and only a small portion of the data is needed. This portion of data is the one needed to produce the distance matrix as shown in FIGURE 3. The result of the algorithm—a segmentation boundary—will be produced at least after the change-point in the signal (i.e., the online data) reaches half the size of the distance matrix. This will produce a delay between the occurrence of the change-point and the novelty peak. Further delay may be caused by a smoothing filter to the novelty score, to eliminate noise, and the test for a peak. This delay time is further referred to as *lag* and should be considered in the practical application of the algorithm.

To investigate the suitability of the algorithm for the segmentation of perceptually relevant segments of bodily motion, multimodal data were collected at the motion capture laboratory of the Music, Art and Culture Studies department of the University of Jyväskylä. Adult participants ($n = 12$) were invited to the laboratory for an experiment about moving to music. These participants are further referred to as *dancers*, although it was not required from them to have any previous training or skill in dance. The music used for the experiment was the following:

“Bouzouki Hiphop - Rempetila” (Tetarto Hood, 2014) from the beginning to 45.7 s. with no fade-in or fade-out. This is Rembetiko instrumental music mixed with Hip-hop bass and drums. Tempo is 90 BPM and meter is 4/4. All participants declared to not know this piece.

“Minuet” in G Major (Petzold, ca. 1725). MIDI rendition with piano sound, from beginning to end (104 bars, duration 92.5 s.). Tempo is ca. 128 BPM and metre is 3/4. All participants declared to know this piece.

“Ciguri” (Otondo, 2008) from 56 to 183.7 s. (duration 122.7 s.) with fade-out the last 5 s. This is an electroacoustic piece that has no perceivable beat and therefore no metre. All participants declared to not know this piece.

“Stayin’ Alive” (Gibb et al., 1977) from the beginning to 108.5 s. with fade-out the last 2.3 s. Tempo is 104 BPM and metre is 4/4. All participants declared to know this piece.

Data recording was done with one dancer at a time. Each dancer was asked to wear a motion capture suit to collect position data with an optical motion capture system. They were asked to move spontaneously to the music when it started sounding through the loudspeakers. Optical motion capture data, accelerometry, and video were recorded. The recordings started and ended at the same time as the music excerpt.

Each music excerpt was presented twice. On the first presentation dancers were asked to move to the music without any constraint other than an area of approximately 4m². The second time they were asked to hold a *wiimote* (a device containing a triaxial accelerometer) with one hand and “dance” only with that arm. In this condition the dancers were asked to remain at the centre of the area facing to a corner of the room in order to get in the video recording the most complete visualisation of the arm’s movement. This procedure was repeated for each music excerpt.

After the multimodal data were collected, only videos of the dancing with one-arm holding the wiimote were used for the collection of perceived segmentation data, or *annotation*. This choice was made as it seemed more appropriate to evaluate the segmentation algorithm with portable sensing technology (accelerometry) instead of laboratory equipment (optical motion capture). This considered a possible subsequent use of the segmentation procedure in the design of a portable practical application such as a digital musical instrument.

The annotations were later to be used as *ground truth* to test the segmentation algorithm. Article II described a pilot experience with two participants providing annotations: one female, the other being the author of this dissertation. In the pilot experience only the video corresponding to single-arm movement to the “Stayin’ Alive” excerpt was annotated. Article III described the experience with six participants (3 male, 3 female) in which they annotated only two performances of dancers (one of a male, the other female), for each music excerpt of “Minuet”, “Ciguri” and “Stayin’ Alive”. This reduction was made to prevent the task being too long and to cause fatigue, while still retaining musical variety. The participants performing this task are referred to as *annotators*, to differentiate them from the dancers. The annotation was done in two conditions using a computer for presentation. The first condition was real-time annotation, in which videos with their corresponding audio are segmented as they are watched. The second condition was non-real-time annotation, in which videos without audio were segmented as they were watched, with the option of scrolling the video back and forth to refine the annotation. Only the latter was used as ground truth, as the participants

declared in a post-task interview, that it was not always possible to accurately identify segment boundaries in real time.

After the ground truth data collection was completed, it was used to test the effectiveness of the segmentation algorithm applied to the accelerometry data captured by the wiimote at the laboratory. In the first attempt, reported in Article II, annotations of the participants were summarised into a single compound sequence of segmentation boundaries for each video, using peaks of Kernel Density Estimation. This method had been used elsewhere for manual segmentation of audio (e.g., Hartmann et al., 2017). These boundaries were compared with the output of a segmentation procedure. This procedure combined several features extracted from the accelerometry, such as kurtosis, skewness, mean, root mean square, standard deviation, mean absolute deviation, interquartile range, and centred zero-crossings count. Then, these features were fed into the segmentation algorithm explained above.

A grid search was performed to obtain the highest similarity between perceived and computed boundaries. The search involved the manipulation of several free parameters of the procedure, such as whether the triaxial data or its magnitude was used, the size of the window over which a feature was computed or the void use of a feature, the size of a gaussian to smooth the feature, the size of the distance matrix, and the peak threshold. Since the dimensionality of parameters is high, also a genetic algorithm was used to perform the search and to optimise for the highest similarity between annotated and computed boundaries. The similarity measure used in the pilot (Article II) was improved, therefore only the improved version described in Article III is explained further down this text.

Visual inspection of the computed boundaries having the highest similarity with annotated boundaries reveals that while some boundaries are remarkably close, there are some computed boundaries that do not have any matching annotated boundary or are too far to be considered as matching. In general, it was possible to observe remarkable closeness between annotated and computed boundaries, but only within isolated regions.

The methods described in Article II for assessing the segmentation algorithm were deemed worthy of improvement. Firstly, only by visual inspection of the annotations of more than two participants it was evident that annotations were too dissimilar to summarise them in a single response. This is consistent with previous research observing that perceived annotations of dancing subjects are highly dissimilar amongst observers (e.g., Kahol et al., 2004). Second, the segmentation procedure had too many free parameters and optimising them lead only to find local optima and overfitting (i.e., works well only with a very specific combination of parameters). Third, the high number of parameters also contributed to make the process computationally expensive and thus unsuitable for real-time implementation. Following these observations, the methods were improved and reported along with results in Article III. The first improvement was to tailor the computed segmentation to the responses of each annotator separately. In a practical application, this would mimic the

adjustment that might be achieved manually by a user or automatically by a machine-learning procedure. The second improvement was to eschew features extracted from the raw acceleration data. Instead, the raw acceleration data was used as an input to the segmentation algorithm.

Additionally, the similarity measure was refined, and it works as follows: a and b are the time indexes of annotated and computed segmentation boundaries, respectively. Each element in a is paired to the closest element in b . Then, the time differences between each paired element are added. The resulting sum is divided by the total time of the corresponding music excerpt and the resulting value is subtracted from one. The result of this calculation represents closeness of the paired values. To penalise the difference in the number of boundaries, the rate of paired elements is computed by dividing the number of paired values (the number of pairs multiplied by two) by the sum of the unpaired values in a and b . Finally, the measure of similarity is closeness multiplied by the rate of paired elements. This measure will range from 0 to 1, where the maximum value indicates that a and b are equivalent and vice versa. A Monte Carlo simulation revealed that similarity equal to, or more than 0.66, will have a p -value of 0.05 or less.

A grid search was performed to maximise similarity between annotated and computed segmentation boundaries, by modification only of the size of the similarity matrix and the novelty peak threshold. The greatest mean values of similarity between computed and annotated segments ranged from 0.73 to 0.91 and were found for the musical excerpts of “Minuet” and “Stayin’ Alive”. These music excerpts have a clear beat and were familiar to the dancers. Conversely, similarity was lower for “Ciguri”, which is a piece that has no clear beat and was not familiar to the dancers. This suggests that the effectiveness of the method may be directly related to the presence of a clear beat and the familiarity of dancers with the music.

The computed boundaries having the greatest similarity with annotated boundaries were then assessed by the same annotators. This means that the assessment is for the ‘best case scenario’. The annotators were presented with the same videos used for the annotation. The videos had an embedded scrolling timeline with consecutive numbers for boundaries and were presented in three versions: computed boundaries, annotated boundaries, and random boundaries. The videos were presented in random order and the annotators were asked to confirm or reject each boundary. FIGURE 4 shows the result of the assessment. Full lines indicate confirmed and dotted lines indicate rejected. The music excerpts for annotations 1, 3, 5 and 6 is “Stayin’ Alive”; for annotations 2 and 4 is “Minuet”. This assessment resulted in maximum Precision¹ values ranging from 0.71 to 0.89, and Recall² ranging from 0.82 to 1, for computed boundaries.

¹ Number of confirmed computed boundaries divided by number of computed boundaries.

² Number of confirmed computed boundaries divided by the sum of confirmed computed boundaries and the difference between the number of confirmed annotated boundaries and the number of paired annotated and computed boundaries.

Additionally, Precision of annotated boundaries³ ranged from 0.67 to 1. The latter may be seen as measure for the reliability of annotations. The lag time corresponding to the evaluated sequences of computed boundaries was maximum 0.5 s., while the median was 0.35s.

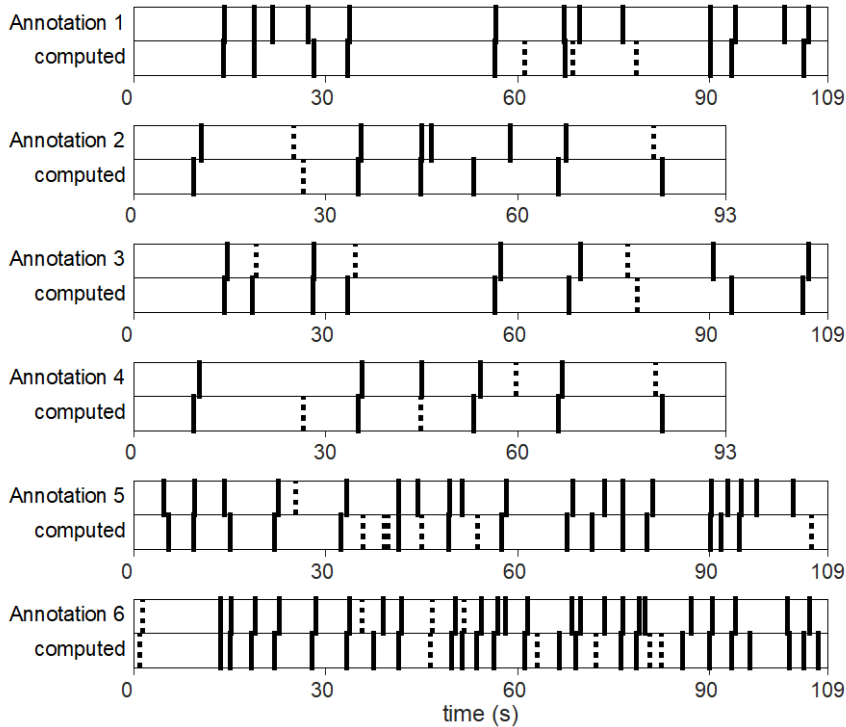


FIGURE 4: Perceptual assessment of segmentation effectiveness.

4.2.2 Article IV

The study reported in this article explored the relative contributions of auditory and visual information, as well as composed and performed emotional expression, to the perception of emotions in musical performances. The study builds on previous research and contributes with novel elements. Firstly, in this study musical performances of piano and violin were used, as these instruments had not been extensively used in previous similar research. Second, this study focuses on perception of emotions in the musical performances, complementing similar previous research that had focused on felt (i.e., induced) emotions (Camurri et al., 2004; Castellano et al., 2008; Vuoskoski et al., 2014; Vuoskoski, Thompson, et al., 2016). Third, the musical performances had emotional intentions independent from the composed musical emotions, yielding congruence and incongruence. This was expected to generate a richer variety of emotional content for the investigation. Fourth, to assess the contribution of auditory and visual sensory modalities, the perceived emotions were measured in the musical performances presented only auditorily, only

³ Number of confirmed annotated boundaries divided by the number of annotated boundaries.

visually, and audiovisually. Fifth, an analysis was carried out to quantify the relationships between perceived emotions in the musical performances and features from the performances. These features were auditory and kinematic, and allowed for detailed identification of parameters associated to perceived emotions. The data collection, analyses and results are explained in what follows.

A violinist and a pianist recorded solo performances of four short musical pieces. These musical pieces were taken from a previous study by Vieillard et al. (2008) and were composed to convey three discrete emotions: happiness, threat, and sadness. They were originally composed for the piano, having chords and bass in the left-hand part, and melody in the right-hand part. The melody of each piece was extracted and was performed by the violin. Additionally, a piece intended to convey peacefulness was generated by changing the modality of the piece conveying sadness, from minor to major. The musicians were asked to perform four versions of each piece, having different emotional expressions: Happy, Angry, Sad, and deadpan (i.e., no emotional expression). They also were instructed to avoid extreme variation of tempo between performances.

The musicians' performances were recorded at the motion capture laboratory of the Music, Art and Culture Studies department of the University of Jyväskylä. The posture and motion of the musicians was recorded with an optical marker-based motion capture system. Eighteen reflective markers were placed on the pianist's head, arms, wrists, hands, torso, and hips, with an additional two markers placed at each end of the keyboard. Twenty-six markers were attached to the violinist's head, shoulders, torso, arms, wrists, fingers, hips, knees, ankles, and feet. Additional markers were attached to the violin: one on the curl, two on the body, and one on each extreme of the bow. This resulted in 16 musical performances of each instrument.

Each performance was recorded as audio and as motion-capture data. Using the latter, a video file was generated for each performance, showing a skeleton produced by a procedure that reduces markers and draws lines between markers (Burger, Thompson, et al., 2013), plus the markers on the instruments. This was done to remove facial expression and retain only the bodily posture and motion.

The performances were then rated for their perceived emotional content by 90 participants which were university students. Three groups were made, each of which was presented with the musical performances in different conditions: Participants in group 1 ($n = 31$) rated only audio and only video presentations of the piano performances, while participants in group 2 ($n = 34$) rated only audio and only video presentations of the violin performances. Finally, participants in group 3 ($n = 25$) rated audiovisual presentations of both the piano and violin performances. The participants were asked to rate on a continuous scale, the extent of perceived Happiness, Anger, Sadness and Tenderness, in each presentation. The performances were presented in random order, on a computer screen, and the participants listened using headphones. The presentation was executed by a software that also showed horizontal

sliders on the screen that could be manipulated with a mouse, for the rating of perceived emotions.

The data of perceived emotions was scrutinised for inter-rater agreement. Each rated emotion for each presentation condition was assessed by means of Krippendorff's alpha (K_α) (Krippendorff, 2011) and two-way random Intra-Class Correlation (ICC) (Shrout & Fleiss, 1979). K_α was low to moderate (0.33 to 0.63), indicating high variance in the responses. However, ICC was high (0.86 to 0.98), indicating consistency for variation across responses even though their means may substantially differ. Regardless of the measure, inter-rater agreement was much lower for the ratings of only video.

A series of two-way repeated-measure analyses of variance (ANOVA) were computed to explore the relative contributions of Composition (i.e., composed emotions) and Expression (i.e., performers' expressive emotional intentions) on the ratings of mean perceived emotions, and how this might vary across the presentation conditions (only audio, only video, audiovisual). The two within-subjects factors were Expression (Happy, Angry, Sad, and deadpan) and Composition (happiness, threat, sadness and peacefulness). The analyses were carried out separately for each presentation condition, emotion, and instrument. FIGURE 5 shows the mean main effects for all emotions.

For the only audio condition, the main effect of Composition was larger than the main effect of Expression, suggesting that composed emotion accounted for more variance in participants' ratings of perceived emotion. Post-hoc tests revealed that the participants were, with some exceptions, successful in decoding the expressive emotional intentions of the musicians based on auditory cues alone. In the only video rating condition, the main effect of Expression was substantially larger than the main effect of Composition. In other words, in the absence of auditory information, the type of expressive intention accounted for substantially more of the variance in participants' ratings. Post-hoc tests of these responses revealed that the participants were, with some exceptions, successful in decoding the expressive emotional intentions of the musicians based on auditory cues alone and visual kinematic cues alone. In both cases, the Sad expression was rated as the most tender although Tenderness was not an emotion expressed by the musicians. In the audiovisual condition, the differences between the mean effect sizes of Expression and Composition were reduced, suggesting that, compared to the only audio condition, visual information enhanced the perceptual salience of expressive intentions in relation to the composed emotional intention.

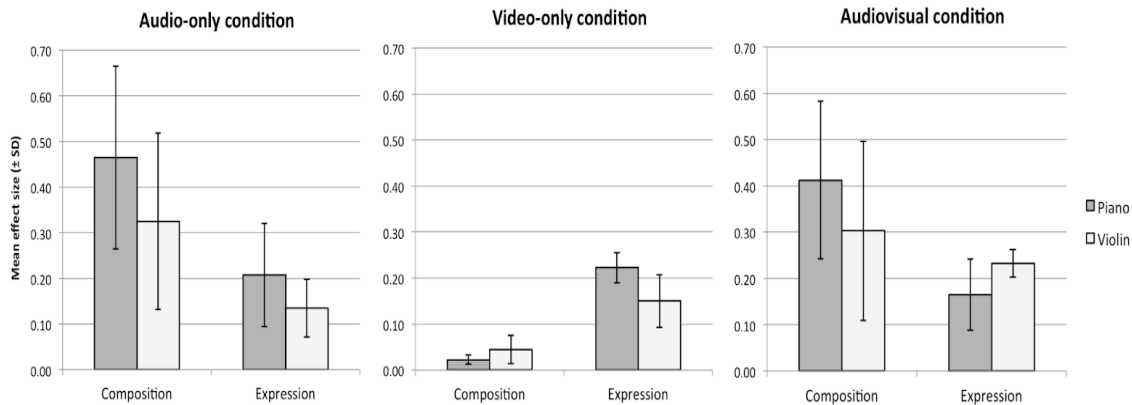


FIGURE 5: Effect sizes for perceived emotions.

Additional analyses looked at the relationships between perceived emotions and features of the performances. Several features were computed from the motion-capture data describing kinematic aspects of movement (Dahl & Friberg, 2007). Also, auditory features were annotated and computed from the audio recordings. The correlations amongst features are mostly obvious, for example the average time between notes onset was negatively correlated to all kinematic features. However, kinematic features were highly intercorrelated for the violin but not so for the piano. Interestingly, mode (major or minor) was the least correlated feature with other performance features, auditory or kinematic.

Linear correlations and simple regression models between each feature and mean ratings for each emotion showed that, for both instruments, time-derivatives of motion (i.e., speed and acceleration) have greater correlations with perceived emotions when audio was not present. Also, the time-derivatives, for both instruments, are negatively correlated with the ratings of Sadness and Tenderness. This effect is greater in the violin ratings, showing clear positive correlation between motion time-derivatives and the ratings of Happiness and Anger. The correlations between time-derivatives and ratings for violin performances are stronger than for piano performances when the presentation of performances was audiovisual. However, in the only audio condition, all features have very low or no correlation with the perceived emotions in violin performances. Conversely, for piano performances, the relations between emotion ratings and all features are remarkably similar in both conditions where audio is presented. Auditory features that are highly correlated with the motion derivatives also have high correlations with emotion ratings. This may be attributed to the obvious relation between the physical energy used to produce a sound and the energy of the resultant sound. Other auditory features such as mode and lower spectrum content are most correlated with perceived emotions when audio is included in the presentation.

Further analysis comprised multiple regression models for the mean perceived emotion ratings, having standardised audio and kinematic features as regressors (i.e., independent variables, predictors). Subsets of features with low collinearity were compiled having only auditory features, only kinematic features, and both. Models having all the possible permutations of features

within each subset were computed, including two to five features. None of the multiple regression models for subsets of only kinematic features had better fit than their simple regression counterparts. All the multiple regression models having subsets of only auditory features and having higher fit than their simple regression counterparts appeared in the models selected from the ones computed with the subset of audio and kinematic features, except one. The model for Sadness ratings of only audio piano performances made of only auditory features is slightly improved by adding the kinematic feature describing total amount of motion of the right hand and replacing variability of the lower part of the spectrum for variability of the higher part of the spectrum. This indicates a contribution of the coupling between movement of the right hand and the melody it is playing.

Notably, all the multiple regression models with highest fit corresponding to the presentation conditions with audio, have mode included. For positive valence (Happiness and Tenderness) mode is positive and vice versa. For the ratings of Sadness in audiovisual condition, the models are the same: lower spectrum and minor mode. A similar effect is observed for the ratings of Anger: high variability of lower spectrum and minor mode.

The combination of features in the models reveal some differences between the instruments. For example, in the case of the piano, models for Happiness and Anger in both conditions with audio are similar, meaning the same features with very close coefficient values. For ratings of Sadness there is an inverse relation with variation of lower spectrum, while for Anger the relation is positive. This might be because the pianist played chords with less dynamics in the pieces rating high in Sadness, while in the cases for higher perceived Anger, the pianist may have played the chords with more energy.

A few models were improved over the simple regression, by including either or both average bodily speed and amount of movement, which are closely related measurements. Some models corresponding to ratings for performances presented without video have relevant contributions of kinematic features. Also, for ratings of piano performances, the average speed of the right hand has the greatest contribution for ratings of Happiness and Sadness, and the average speed and amount of motion of the left hand for ratings of Anger. Presumably this is because the right hand played the melody noticeably fast. Likewise, the left hand played the chords and, as mentioned previously, they might have been played more energetically in the pieces with higher ratings for Anger. In the case of the violin, the model for Anger was improved with the inclusion of total amount of motion and variability of high spectrum, to the existing model including only mode and having very poor fit. This leaves without a significant model, only the ratings of Happiness for violin performances presented in only audio condition.

4.2.3 Article V

This article reports on a study that explored the hypothesis that musical emotions are embodied differentially by people according to their personality.

To accomplish that, a set of musical excerpts was rated for perceived emotions. A separate group of people was asked to spontaneously move to each musical excerpt. The correspondence of bodily movement to the emotions portrayed by the music is referred to as “embodied musical emotions”. The extent of embodied musical emotions was measured against the personality traits of each individual. The experimental procedure, analysis and results are explained in the following paragraphs.

The music used were 30 audio excerpts of different popular music genres, chosen to have a variety of rhythmic complexity and tempo. All excerpts were 28 seconds long, solely instrumental, and had a binary metre. They were further trimmed to 15 seconds by removing the first and last 6.5 seconds. A group of 34 participants, all musicology students at the University of Jyväskylä, familiar with research on music and emotions, and of Finnish nationality, rated perceived emotions in the trimmed excerpts. These shorter excerpts were used to abbreviate the duration of the rating task, thus reducing the risk of fatigue. They were asked to rate perceived emotions in the music, on seven-point scales for dimensional affect in terms of Arousal and Valence, and for discrete emotions Happiness, Anger, Sadness, and Tenderness. The music excerpts were presented in random order.

Another group of participants was asked to move to the untrimmed music excerpts. These were 60 participants selected from a pool of 952 individuals that had completed the Big Five personality inventory. The selection was made such that the responses evenly covered low, middle, and high scores for each personality trait. Each participant was recorded in a separate session in which they were asked to move to the music in a way that feels natural. A recording was made for each music excerpt, with an optical motion capture system tracking position of 28 reflective markers attached to the body. As in the rating for perceived emotions, in this task the music excerpts were also presented in random order.

Kinematic and non-kinematic bodily features were computed using the motion-capture data. They represented a variety of aspects of bodily motion and posture. Those features representing movement of individual bodily parts use subsets of markers locked to a local coordinate system defined by a reference plane. This reduces collinearity among features, which is desirable to use them as regressors in linear models (see below). Kinematic features were time-derivatives (speed, acceleration, and jerk) and the square of speed (speed²) of markers, amounting to 32 features. The square of speed was included as a supplemental measure for kinetic energy which is half the mass multiplied by the squared velocity. The mass can be omitted from the equation because it is constant. The Euclidean norm was computed for each kinematic feature, resulting in a single value corresponding to each motion-capture recording. The non-kinematic features were rotation of the torso, distance between hands, distance between elbows, distance between feet, and area. For these features, the median was used as a single value for each recording. An exception is the feature “rotation of the torso”, for which the standard deviation was used. The

feature “area” is defined as the smallest rectangular area of the centremost marker projected to the horizontal plane in a moving window of 4 seconds. The result was 38 bodily-feature values for each motion-capture recording.

The rank correlation between each feature and each of the six ratings of perceived emotions was computed, resulting in 228 values of embodied emotions for each participant. Then, two analyses were performed to assess the relations between embodied emotions and personality traits. Analysis 1 comprised the rank correlation between personality traits and embodied emotions. Additionally, the rank correlation was computed between personality traits and six aggregated embodied emotions. “Aggregated embodied emotions” is the sum of absolute values, of embodied emotions corresponding to the same perceived emotion.

The results of Analysis 1 reveal moderately weak correlations ($< \pm 0.25$) between each Big-Five personality trait and the embodiment of each rated emotion by each bodily feature. However, the number of bodily features and perceived emotions with correlations having p -value over 0.05, are distinct for each personality traits. This number is higher for Extraversion, followed by Openness and Agreeableness, then Conscientiousness, and Neuroticism only showing negative correlation between hands distance and Valence. Correlations between personality traits and aggregated embodied emotions were weak ($< \pm 0.2$) and their p -values were high, casting doubt of their significance. However, it is at least possible to observe negative correlations between trait Neuroticism and perceived Happiness and Valence, while the opposite was observed for the other traits. This suggests correctness of the measure as Neuroticism is by definition a negative trait compared to the other traits.

Analysis 2 consisted in linear regression models for each personality trait. The regressors (i.e., independent variables, predictors) of the models were embodied emotions, but only one bodily feature was allowed in each model, to examine the effect of each bodily part separately. All 63 possible combinations of perceived emotions for 38 embodied emotions (one for each bodily feature) resulted in 2394 models for each personality trait. Instead of selecting models by their statistical significance, *relevance* was assessed empirically by comparing the cross-validated error of a data model ($RMSE_{CV}$) and the error of a null model ($RMSE_{null}$). The relevance measure is $\Delta RMSE = RMSE_{null} - RMSE_{CV}$, where a positive value indicates that the model is relevant, as it performs better than the null model and vice versa. The regression models produced by Analysis 2 amounted to 11970.

FIGURE 6a shows the relevance measure $\Delta RMSE$ clustered by personality trait, for all models performing better than the null model. A quick visual inspection reveals that the highest values are for Openness, followed by Agreeableness and then Extraversion. The greater number of relevant models are for Agreeableness, followed by Openness and Extraversion. Conscientiousness and Neuroticism have both the weakest values and smallest number of models. FIGURE 6b shows only models whose regressors are correlations between a bodily feature and any combination of ratings for only

dimensional affect. Notably, none of these models for Openness perform better than the null model, and most models for Agreeableness perform better than models for the other traits. FIGURE 6c shows only models whose regressors are correlations between a bodily feature and any combination of ratings for only discrete emotions. In this case the pattern is similar to that when all types are allowed, but the best performing models for Agreeableness are not as strong as for models having only dimensional affect or for all models. This is consistent with Vuoskoski and Eerola’s (2011b) finding that, regarding music-induced emotions, discrete emotions have stronger relationships to individual differences than dimensional affect. When any combination of regressors for discrete emotions and dimensional affect is allowed, then the maximum Δ RMSE for Extraversion is significantly higher than when either only dimensional affect or discrete emotions are considered.

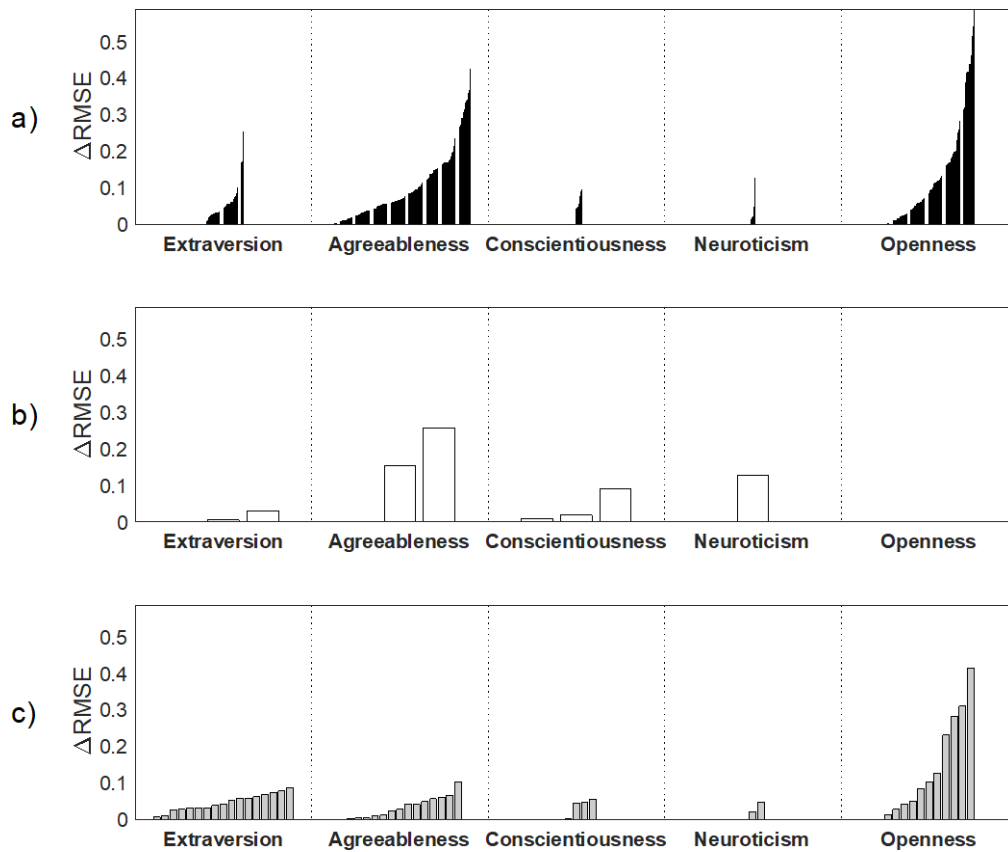


FIGURE 6: Relevant models of personality traits.

The best fitting models are shown in TABLE 1. When looking at the regressors in more detail, it is worthwhile to note that the embodiments of Valence and Tenderness by the bodily feature representing vertical acceleration of the centre of the body, are regressors for a model that is the most relevant for Extraversion and one of the most relevant for Openness. A closer inspection of this model reveals that the coefficients for the regression fit are very similar for both personality traits. However, the fit—and prediction power—of this model, is greater for Extraversion.

TABLE 1 : Best fitting models for each personality trait.

Personality Trait	Bodily Feature	Perceived Emotions	R^2	p^*	$\Delta RMSE$
Extraversion	Body center vertical acceleration	Valence and Tenderness	.173	.004	.254
Agreeableness	Body center speed ²	Arousal, Happiness and Tenderness	.281	<.001	.425
Conscientiousness	Head speed	Arousal and Anger	.122	.025	.096
Neuroticism	Hands distance	Valence	.089	.021	.127
Openness	Area	Arousal, Valence, Happiness and Sadness	.308	<.001	.587

* p -values are not adjusted. See Rothman (1990) and Althouse (2016).

Looking at both analyses, no single bodily feature embodying a musical emotion was a high rank correlate of all personality traits. Likewise, no single combination of bodily features embodying any combination of perceived emotions predicted all personality traits.

The model selection method leading to the results presented above is focused on the prediction performance of models, allowing the best combinations of regressors for each model, with the sole constraint of having regressors for only one bodily feature for each model. However, this means that regressors are removed from a model only to improve its prediction power. Even when the models have been cross-validated, it is possible that regressors remain in the model because of their noise instead of their true explanatory power. Therefore, it is convenient to also examine only models that have all regressors for each type of emotional rating and also the models that have all emotional ratings. TABLE 2 shows all relevant models that have regressors considering all emotional ratings, all dimensional affect ratings, or all discrete emotions ratings. In these conditions, no relevant models are found for Extraversion or Neuroticism. Additionally, all except the following bodily features appear in regressors for at least one relevant model: Speed of all body parts, jerk of all body parts, shoulders' acceleration, and squared speed of the head. Hence, these features may be irrelevant.

TABLE 2 : Relevant models with all regressors of each subset.

Personality Trait	Bodily Feature	Perceived Emotions	R^2	p^*	$\Delta RMSE$
Agreeableness	Hands distance	dimensional affect	.191	.002	.257
Agreeableness	Body center speed ²	all	.307	.003	.169
Conscientiousness	head speed	dimensional affect	.118	.028	.091
Openness	Area	discrete	.245	.003	.310
Openness	Area	all	.316	.002	.315

* p -values are not adjusted.

4.3 Application

4.3.1 Article VI

This article describes a system that was devised as a proof of concept for the feasibility of unsupervised learning of patterns in a continuous input signal, for gestural control in a musical application that doesn't require quick responsiveness. The system is conceptually a musical instrument in a broad sense, for it essentially allows a user to control sound. One key quality of this instrument and the innovation presented in the article, is that it segments gestures without the need of explicitly indicating their starting and ending points. These gestures can be used to trigger actions that control sound. However, the detection of gestures needs some time to occur. It might be just a fraction of a second, but it is enough to be perceived as not instantaneous. This would normally be considered a disadvantage, but in this context it is not.

A polystyrene ball having 12 cm. of diameter was cut in half and the interior was carved to fit a Myo armband controller (FIGURE 7). The Myo was originally designed by Thalmic Labs to be worn on the forearm. It has several sensors, but the system described here only utilised the triaxial accelerometer. The two halves of the ball are put together restoring the spherical shape, but it can be easily disassembled to recharge the battery of the Myo. The data from the sensors is broadcast in real time using the Bluetooth Low-Energy (BLE) specification. The BLE signal is captured by a personal computer nearby, and a piece of software (Visi, 2017) outputs the data in real time using the Open Sound Control (OSC) format. This data is sent to a User Datagram Protocol (UDP) port, where it can be accessed by other software as described below. The Myo was used for its convenience, as it was available to the researcher along with the software to access the data in real time.

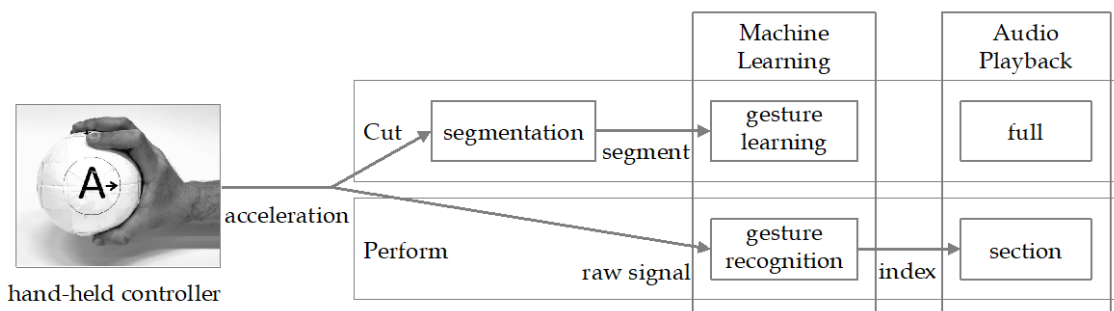


FIGURE 7: Proof of concept for musical gesture segmentation.

The online temporal segmentation method described in Article III can detect boundaries between gestures performed with the hand-held controller continuously, without the need to indicate their start or end. Segmentation occurs in real time, and the result (a gesture boundary) is obtained with a delay time comprising the lag of the procedure as described in Article III, plus computation time. Given these properties, the segmentation procedure was

deemed to be useful as a preprocessing stage to map data from the gestural controller to sound, instead of directly mapping raw sensor data. However, because of the delay in the result, any action connected to the detection of a boundary cannot be performed immediately. Also, a further meta-parameter was incorporated to prevent segments having less than a given duration. This parameter was incorporated because the transitions between gestures may be detected by the system as short segments. The segmentation procedure, the musical system and its graphical user interface were implemented in the Pure Data programming environment, which receives the accelerometry data using OSC as described above.

The detected segments are stored in memory and fed to a machine-learning process, so that later they can be recognised when performed continuously. The DTW algorithm was used for gesture learning and recognition. This algorithm was chosen as it is available in the easy-to-use software Wekinator (Fiebrink et al., 2009), which communicates with Pure Data using OSC over a UDP port. However, another online learning and recognition algorithm could be used (e.g., HMM). As with segmentation, the result of the recognition has a lag due to buffering and computation.

The segmentation and machine-learning processes are incorporated into a system that allows a user to reorder sections of an audio file. The sections are indicated by the segments detected in the hand-held controller's signal. The use of the system comprises two main stages: Cut and Perform (FIGURE 7). In the Cut stage the full audio file is played while the user performs distinct gestures. The boundaries between these gestures are detected in real time by the segmentation process and their time location is stored and labelled with a sequential index (i.e., 1, 2, 3...). As this happens, the segments of the triaxial accelerometry signal are fed as individual examples to the gesture learning process (i.e., "one-shot" learning) and stored for later recognition. Also, a green vertical line is placed in the graphical user interface over a plot of the signal, to indicate that the gesture has been successfully segmented.

After the Cut stage is executed, in the Perform stage the gesture recognition process is continuously comparing the incoming triaxial accelerometry signal, to all the segments that were stored in the Cut stage. The segment that is closest to a stored segment will be deemed a match and its corresponding audio section will be played in a loop. The learning process will keep assessing similarity between the incoming signal and the stored segments. If a gesture corresponding to a different section than the current one is recognised, then the corresponding audio section will be played once the current section reaches its end.

During the development of the proof-of-concept, the system's components were tested separately and progressively combined. This was done to inform the system's design and implementation. One outcome of testing that is particular to the interaction paradigm presented here, is the discovery of gestures that work well with the system. This means static and continuous gestures (see BACKGROUND) and combinations of gestures that can be

segmented, learned, and recognised by the system. This involved the adjustment and fine-tuning of the system's meta-parameters: timescale (i.e., granularity, given by the size of the algorithm's checkerboard kernel), minimum duration of segments, novelty smoothing, and threshold for novelty peaks. Also, parameters of the DTW process had to be adjusted but those are not discussed as the DTW algorithm is well known and documented (Gillian et al., 2011). Testing was done in three phases: the researcher alone; other researchers and students of music; the general public at an outreach event of the University of Jyväskylä. In the first testing phase a sequence of seven gestures (static and continuous) was formulated to work well with a given setting of meta-parameters. The combination of gestures required rotating the ball. To aid in this, a printed letter was placed at each of the six orthogonal orientations, with an arrow indicating the direction of the next gesture in the sequence (see FIGURE 7). This sequence was used in the two following testing phases, in which participants were also allowed to freely experiment and discover other gestures that might work well. Several pieces of music were used, but to standardise the test, observations were focused on the usage with an electronic dance music piece. The findings of the testing were as follows:

Static gestures: Any set of orientations that are different enough among them will work, but it was observed that the six orthogonal orientations along the three axes work flawlessly.

Continuous gestures: Any combination and variation of repeated movements along the three orthogonal axes works well. These axes need not to be aligned with the horizon (i.e., they may be diagonal). Movements that are sudden and energetic work best, as these have high acceleration and therefore are better measured by an accelerometer. Also, circular and semi-circular motion in different orientations, and "8 figure" trajectories could be segmented well as long as the speed, and therefore radial acceleration in the case of circular motion, was powerful enough to produce a novelty score above the set threshold. Conversely, smooth movement will yield little or no acceleration, and might not be detected by the system.

Transitions: Some transitions between gestures may be longer than the set minimum duration. In the Perform stage the system may get stuck looping such short segments, resulting in what one participant called "a DJ effect". Another participant experienced the same result and both expressed that they liked it.

Form factor: One participant of the second testing phase tried to use the device with closed eyes to explore the possibility of not having to look at the ball when manipulating it. This happened after the participant realised that it was necessary to look at the ball when changing its orientation and to look at the computer screen to check if the gesture was successfully segmented. A discussion ensued, which led to conclude that, since the ball is fully symmetric, it is not possible to be aware of its orientation without looking at it.

User experience: The task proved to be challenging to different extents. Some participants expressed that they wanted to try again to improve the number of correctly segmented gestures. All participants showed engagement

and enjoyment. However, it is important to consider that participation was voluntary. It is to expect that researchers and students have interest as the experience is related to their profession and studies. Likewise, it may be safely assumed that visitors at the outreach event attended because they had curiosity about what they may be presented with.

4.3.2 Article VII

This article reports on the development of a system to produce musical sonification of daily activity data recorded by wearable devices equipped with accelerometers. The method employs a novel approach to multigranular temporal segmentation, that results in a clear correspondence between daily events and sound. Additionally, the system does not require the final user to do any fine-tuning of parameters. This system could be used in healthcare by helping people to be aware of their own daily physical activity in a novel and engaging way.

To develop the system, two accelerometry recordings from 75-year-old adults were used. These were chosen from the AGNES database (Portegijs et al., 2019; Rantanen et al., 2018) so that one corresponds to a low-activity sedentary subject while the other corresponds to a high-activity non-sedentary subject. The data was recorded by two tri-axial accelerometers, one chest-worn and the other thigh-worn. These data were pre-processed to obtain the Mean Absolute Deviation (MAD) of the square norm (Vähä Ypyä, 2015), from the thigh-worn accelerometer (FIGURE 8a). Also, activities were identified from the orientation of the accelerometers: lying, sitting, upright posture and walking (Rantalainen et al., 2022) (FIGURE 8b). Then, MAD and activities were integrated, resulting in continuous and smooth curves. For MAD the integration was logarithmically scaled to preserve distribution, as the relation between time of inactivity and activity follows a power-law distribution (Chastin & Granat, 2010).

The integrated data was segmented using the algorithm described in Articles II and III. However, in this study several segmentation boundary sequences were obtained by correlating checkerboard kernels of different sizes upon the diagonal of the distance matrix for the whole data (FIGURE 8c). The resulting sequences at different granularities represent different timescales. The boundaries at different granularities are not perfectly aligned in time (FIGURE 8d) because, as the kernel gets larger, it incorporates more information causing the novelty peaks to move slightly in either direction. Since the sizes of kernels were set to be minimally different, it is safe to assume that they correspond to the same segment. Thus, every coarser granularity boundary has an origin in a finer granularity boundary, except at the borders. The structure is hierarchical, where segments are embedded in larger segments. This reflects the structure of human daily activity. For example, a large portion of the day such as the morning, may contain activities like waking-up and getting ready, breakfast, commuting, and so forth. This hierarchical structure is also analogous to musical structure. For example, a song has sections like introduction, verse, and chorus, each of which have sub-sections, such as melodic lines. However, in

music the boundaries of each section exactly match in time, unlike the structure resulting from the procedure described above. If that multigranular structure was used for musical sonification, it would result in a seemingly unnatural performance. For example, each granularity level may be assigned to a different musical instrument. If so, then the instruments would begin and change sections of the song at different times.

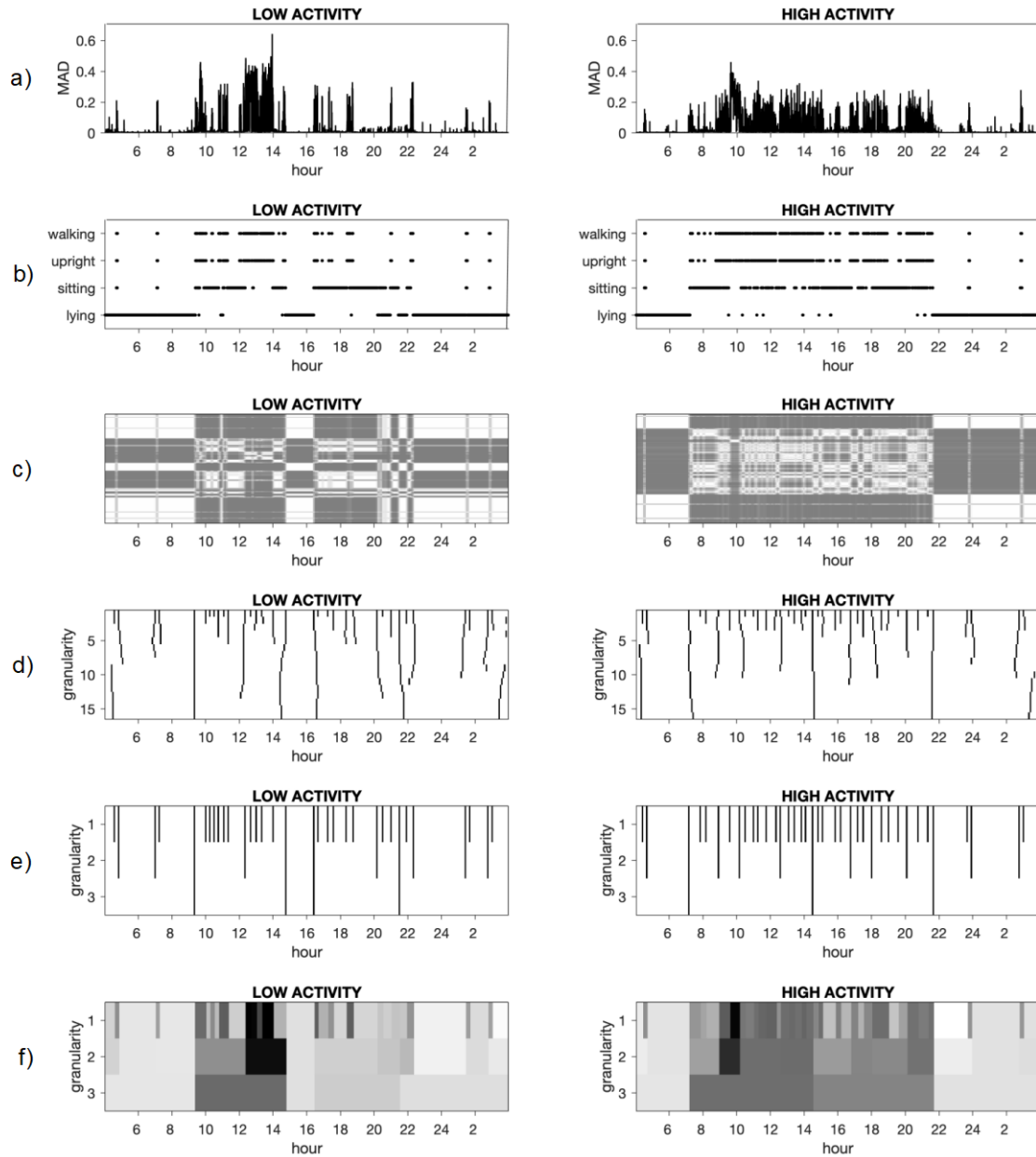


FIGURE 8: Multigranular segmentation of daily activity accelerometry.

Therefore, the segmentation boundaries were aligned to the finest-granularity boundary. Also, the boundaries at the borders were removed. This resulted in sequences at different granularities being identical or slightly different. Thus, the finest and coarsest granularity sequences were kept, as well as the sequences that provide greatest variety in number of boundaries. For the

examples given here, the reduction resulted in sequences at three levels of granularity: fine, medium, and coarse (FIGURE 8e). Finally, the median activity (integrated MAD) was computed for each segment at each granularity level (FIGURE 8f). The result of the process is further referred to as “Segmented Activity” and consists of the length and the amount of activity for each segment, at each granularity level.

The Segmented Activity was read as a musical score by a sonification system programmed in the Pure Data environment, such that each segment is a note. Notes are generated by pseudo-random generators, such that they are part of a user-defined scale (e.g., a pentatonic) if over a defined threshold, or chromatic (i.e., any note) if below the threshold. This results in dissonant melodies when activity is low and vice versa. Each granularity level of Segmented Activity is mapped to a different octave of synthesisers producing bell-like sounds, with notes corresponding to finer granularities being of higher pitch and vice versa. Also, a drum machine plays a user-defined rhythmical pattern that can comprise bass drum, snare, and cymbals. The drum machine will play only bass drum if the activity is low and will incorporate first the cymbals and then the snare drum as the activity increases.

The user specifies how long the performance will last and the program computes the duration of each segment accordingly. The sum of activity for all segments is used as the seed for all pseudo-random generators, to obtain a deterministic performance. This means that the sonification of a given Segmented Activity will always be the same. The mean of activity for all segments sets the tempo. The mean between the values of both subjects was mapped to 120 BPM (beats per minute) for crotchet notes (60 BPM for minim notes), as the typical healthy average heartbeat at rest is just over 60 BPM (Nanchen et al., 2013) and both preferred musical tempo and average walking steps have a period of about 120 BPM (Burger et al., 2014). Hence, the sonification for the high-activity subject will have a slightly higher tempo than the sonification for the low-activity subject. Also, the mean activity is computed for segments of different granularities occurring at the same time. This is mapped to a low pass-filter whose cut-off frequency is increased as activity increases. For example (see FIGURE 8f), the full drum set and only notes within the defined scale play between about 12:30 and 15:00 for the low-activity subject and from about 9:00 to 10:00 for the high-activity subject. As these sections are of high activity, the sound is also spectrally rich and bright. Conversely, the sections with lower activity sound mellow as high frequencies are reduced.

Two audio files were produced with the method described above, using excerpts from 6:00 to 23:00 of the data. These audio files were used as stimuli for a perceptual assessment. Data for this assessment were collected during 31 days by means of a short survey on internet social media, targeting users in Finland and major English-speaking countries. In the survey, participants were asked to listen to each audio file, whose order of presentation was random, and to indicate which of them represented the more active person. A total of 1847 responses were collected by the survey, of which 66.3% correctly identified the

sonification corresponding to high activity. A one-proportion z-test was performed to evaluate the statistical significance of the results, yielding $z = 14.03$, with a p -value $< 1 \times 10^{-5}$. This may be sufficient to reject the null hypothesis, suggesting that the proportion of correct responses was significant.

5 DISCUSSION

This section discusses the main outcomes of the research corresponding to this dissertation. While the presentation of background, research aims and articles was organised according to the initially aimed research areas Theory, Measurement and Application, this section is organised according to the resulting logical linkages among the topics investigated. The discussion starts with a succinct account of the theoretical framework proposed in Article I and how it relates to the following studies. Then, the studies concerned with temporal segmentation (Articles II, III, VI and VII) are discussed. The discussion follows on the studies concerning embodied musical emotions (Articles IV and V). Finally, the theoretical framework presented in Article I is discussed in light of the findings presented in Articles II to VII, and of recent literature.

5.1 Gestures and agency in musical interaction

Article I is an essay that describes the formulation of a theoretical model of embodied musical interaction proposing that both a musical machine (e.g., a musical instrument) and its user (i.e., the musician) can be considered as cognitive agents that communicate by means of musical gestures. In this context, musical gestures are defined as multimodal information that happens in time. The information that gestures carry shape the musical interaction, a concept referred to as “Gestural Agency”. The model is suitable to analyse musical interaction involving any type of musical instrument, ranging from traditional instruments like the violin or the piano, to state-of-the art digital musical instruments that incorporate artificial intelligence.

Because of its broadness, the model is proposed as a general framework to analyse musical interactions integrating humans, traditional musical instruments, and newer electronic musical instruments. Thus, this framework was used as a starting point towards the application of the concept “Gestural Agency” in the design of novel musical instruments that may take advantages

of novel technologies and techniques such as gestural control by means of motion sensors and machine learning. To that end, two specific yet dissimilar aspects of the model were investigated with experimental and quantitative methods. The first one was the parsing of the gestures involved in musical interaction, namely segmentation. The second one was the emotional information carried by such gestures.

5.2 Temporal segmentation of bodily motion

5.2.1 Online temporal segmentation

The method for online temporal segmentation tested in the study corresponding to Articles II and III, yields optimised sequences of computed boundaries that are substantially similar to the human-annotated sequences. However, given that median computation lag is 0.35s. and maximum tested lag was 0.5s., the system is not suitable for any practical application that requires immediate perceptual real-time response (i.e., up to about 10 to 50 milliseconds, according to the responsiveness standards outlined in the INTRODUCTION, subsection 2.4.1). However, this lag time is suitable for applications in which the occurrence of a segmentation boundary is not to be acted upon immediately. For example, the delayed response may be mapped to a procedure that changes the music to which a person is dancing, in such a way that it prompts the person to change the dance, thus creating a feedback loop. Another possible use of this delayed response might be to record the segments, then compute statistics (e.g., mean, standard deviation), and use those for a larger timescale control of music, lights, or other actionable medium. Furthermore, the segmentation result may be used to produce a near-real-time visual or sonic display that may be useful in clinical applications and research in biomechanics, for example.

The sequences of assessed annotated and computed boundaries are visualised in FIGURE 4. While a substantial number of boundaries were confirmed, it is prudent to inspect the results sceptically. For instance, the fifth and sixth boundaries of Annotation 2 seem to be too far for any of them to correspond to the fifth computed boundary. However, this boundary was confirmed in the perceptual assessment. It is not possible to conclude whether this boundary corresponds to any of the annotated boundaries, or if it is a new boundary that was unseen at the annotation task (i.e., serendipity effect) or if it was a mistake made by the annotator in the assessment task.

Another problem is that most annotators rejected boundaries that they had previously annotated. While the values of Precision for annotated boundaries are fairly high, some assessment responses look counter-intuitive. For example, the third boundary of Annotation 4 is evidently close enough to its computed counterpart to be considered an exact match. However, the computed boundary was rejected as shown by the dotted line. Another example

that may cast doubt on the perceptual task is the second and fourth boundaries of Annotation 3. These were rejected but their computed counterparts, even being noticeably very near, were confirmed. These odd assessment responses are not the norm, but they raise questions about the reliability of the perceptual tasks. Additionally, it is worthwhile to bear in mind that the annotation and assessment tasks were done at different times. This might explain the odd responses mentioned above.

The aforementioned problems may be solved by integrating annotation, automatic segmentation, optimisation, and assessment, into one procedure. The assessment questionnaire may be improved by including a task that shows both annotated and computed boundaries in the same timeline, thus making evident to the annotator the difference between them. In addition, the task would require the annotator to explicitly indicate the corresponding annotated boundary for each computed boundary and vice-versa, if such correspondence exists. Despite the drawbacks of the segmentation and assessment methods, the best-case scenario reveals very high Precision and Recall values. This is relevant as the best-case scenario is akin to the best possible re-tuning of parameters that a user could make in a practical application scenario.

This study contributes to the understanding of the behaviour of the segmentation algorithm when specifically applied to perceptually relevant segmentation of music-induced motion data from a hand-held accelerometry device. The knowledge produced by this study was significant for the design and implementation of practical applications of the segmentation method, as discussed in the two following subsections.

5.2.2 Delayed gestural control of musical sound

The online segmentation procedure discussed above was implemented to run in real time to detect pattern changes in a triaxial accelerometer signal. The segmentation process ineluctably produces a lagged response. In other words, noticeable time passes between the occurrence of the change from one pattern to the next, and the reporting that it had occurred. Nonetheless, the musical application presented in Article VI conforms to this constraint. The application allows a user to rearrange the playback of an audio file in real-time, by performing distinct continuous gestures with a hand-held device.

The testing of the system used audio files of recorded music and the meta-parameters of the system were adjusted for the experience with the chosen music. However, any audio recording may be used, and the parameters may be tweaked for further exploration that may lead to unexpected yet interesting results. The testing also revealed that while the system is not able to segment all possible gestures, it can still segment a substantial variety of possible gestures, comprising static orientations and variations of dynamic gestures such as straight and curved trajectories in different orientations. This result was obtained using a single setting of meta-parameters, showing a substantial degree of generalisation. This was unexpected, as the perceptual evaluation reported in Article III suggests that fine-tuning might have been needed for each different

user. Furthermore, participants of the assessment tended to take the task as a challenge, which in combination with the discovery of new meaningful gestures, and the sense-making of the constraints, turned the experience into a ludic one.

While the system appeared to be promising, it also prompted reflection on opportunities for improvement. One immediate improvement that could be made is the form factor. The hand-held device would improve by having a form that allows the user to manipulate it without needing to look at it. Further variations of the system may include several sensors and the detection of patterns using features extracted from the raw triaxial acceleration. For example, a second hand-held device may be incorporated, or sensors that are not hand-held but wearable. Other sensing technologies may be used as well, such as optical motion capture, video tracking, skeleton recognition from video, or clothing that measures posture changes. In addition, the use of several sensors by more than one person at the same time would allow shared control, turning the experience into a group activity (e.g., Staudt et al., 2022; Tahiroglu et al., 2013). Yet a further idea for future research is the implementation of real-time multigranular segmentation, meaning the detection of gestural boundaries at different timescales. This implies running parallel instances of the segmentation algorithm, each with a checkerboard kernel of different size. Notwithstanding, an offline implementation of multigranular segmentation was devised for the system described by Article VII, discussed in the following subsection.

Current limitations to add more sensors, features, and the capability of real-time multigranular segmentation, are algorithmic complexity, processing power and software efficiency. The two latter are due to the high-level programming and interconnected software used in the proof of concept. The solution beyond using faster hardware is to implement the system using low-level programming. Possibly the best solution would involve the use of embedded software and hardware capable of parallel computing of features and granularities. Moreover, it is important to consider that while the setting of meta-parameters generalised well given the specific configuration being tested in this study, a different setting might be needed when using other configurations of hardware and software, when using different music, when the user is different, or when the user intention changes (for example, to explore different outcomes as suggested above). Thus, future research should consider scrutiny on the effects of the meta-parameters in the segmentation process and the user experience.

The system described in Article VI has potential beyond the musical application described as proof of concept. Consider that in the system described here, the online segmentation procedure only contributes to display on the screen an indication when a gesture has been segmented in the Cut stage. The display of a successfully detected gesture change-point occurs shortly after the actual change. This allows the user, for example, to stop the Cut and restart if a gesture change was not detected. Arguably this is an advantage to the performer, but presumably much more can be done to exploit the online segmentation capability. This capability in conjunction with the notion of

delayed control, deserve more research to explore further possibilities for near-real-time interaction. For example, a system may learn gestures as they occur. This may be incorporated to interactive musical systems where both the user and the system discover and learn gestures at the same time, leading to a process of gestural agency as proposed in Article I.

5.2.3 Musical sonification of daily activity

The contributions of the study described in Article VII are firstly the extension of the segmentation algorithm used in the studies discussed above to produce hierarchical, multigranular, rectified, and reduced segmentation, and its application to the segmentation of daily activity recorded by accelerometry. A second contribution is the application of the resulting segmented daily activity to a deterministic musical sonification paradigm.

Specifically, this study has shown how the occurrence of hierarchical daily events, as well as the amount of average energy recorded within those segments, may be mapped to musical sound. Furthermore, a perceptual assessment of sonifications produced with the described system, resulted in correct identification by a significant majority of the surveyed population. It is worthwhile to note that there are potentially infinite ways of mapping data to sound and their appropriateness may be highly contextual. Hence, the procedure described in Article VII should be taken as a proof of concept and not as the only solution to sonification of daily activity.

The described musical sonification system may be useful in public health interventions towards increasing healthy physical activity or reducing sedentary behaviour, by making a person using the system aware of their intraday activity. The system is proposed as an alternative to visual display of information. Producing music with daily bodily movement might be appealing and thus more engaging than presenting the information by other means. Future research following this study should produce a working prototype to be tested with an ecologically valid population such as people diagnosed with sedentary behaviour.

In practice, the musical sonification system would be part of a portable system comprising hardware and software. Such a system would record daily activity, produce the musical sonification and possibly recommend actions to the user. The hardware may be composed of already existing technologies such as miniature accelerometers and mobile computing devices like a smartphone or smartwatch. Preliminary testing shall be carried out to explore the extent to which the musical sonification may work as an engagement strategy, and to identify the conditions in which it may be effective. These conditions may include personal characteristics of target clients such as age, personality, or income, as well as environmental factors.

Along with producing working prototypes for preliminary testing, it might be convenient to do more basic research. In the first place, it would be useful to explore implementing the sonification in different musical genres. This might contribute to engagement if the musical genre can be tailored to the

user's preferences. Also, the possibility of having different musical genres would provide the user with the option to explore different music genres over time. For example, during the first weeks a user might want to produce musical sonifications of their daily activity in their own preferred musical genres. After a few weeks they might be interested in exploring other genres previously not of their interest. In principle the implementation of different musical genres is feasible as it implies the translation of musical composition techniques into algorithms. Nonetheless, the specific implementation of a system that generates music after providing it with the temporal hierarchical structure and the genre is a challenge that deserves further research.

The incorporation of self-similarity into automatic music composition would be another interesting avenue for future research and may be necessary to improve the musical sonification of daily activity. Self-similarity is a fundamental property of most music and can be measured with the distance matrix of the segmentation algorithm. Self-similarity may be implemented, for example, as sequences of pitches (i.e., melodies) that repeat identically or with variations. The same principle may be applied to chords, rhythm, timbre and spatialisation (i.e., location of sound in the stereo or otherwise multichannel auditory image). Concretely, the self-similarity matrix and the multigranular segmentation boundaries may be used to assess the similarity of each segment with the others, within each granularity level. Those relations may be used to identify segments similar enough to be mapped to similar musical content (e.g., a melody, chord, or rhythmic pattern), and to produce musical variation (e.g., one random note of the melody or chord changes or one subdivision of a note in the rhythmic pattern is generated).

While Article VII describes a method for multigranular segmentation and musical sonification of intraday activity of one subject, it is trivial to expand the method to work with different data. First, instead of using classified data for the segmentation, the activity itself may be used. The raw accelerometry signal might be enough, as demonstrated in the studies reported by Articles II, III and VI. Also, instead of using activity corresponding to one day, the average of several days may be used, resulting in a representation of a typical day. Likewise, sonifications may be produced for periods longer than one day, for example a month, or several months. Furthermore, instead of using data for a single subject, a group of subjects may be used. A population may be pre-clustered in groups with homogeneous characteristics, such as age, gender, personality traits and so on. The resulting multigranular temporal segmentation may be useful to examine the typical intraday behaviour of the group. Its musical sonification will represent the group and this may open new and interesting doors for community music-making. For example, daily data of users may be uploaded to a server, in which their data is combined with data of other subjects in their social circle. This sort of collaborative music making may be a relevant avenue for exploration in further research, as it has been observed that social support through collaboration was the primary motivator for adults to maintain the use of wearable activity trackers (Kononova et al. 2019).

Finally, the described system for musical sonification may be inscribed into the framework proposed in Article I. The system is a musical machine that communicates through multimodal signals. Of these signals the most salient are the input (accelerometry) and output (musical sound). The system may act as an agent as its output (musical sonification) may prompt the user to change its daily behaviour to a healthier one. This will be reflected in the input signal to the system, generating a feedback loop of musical interaction.

5.3 Embodied musical emotions

5.3.1 Contribution of sensory modality

In line with the ideas outlined in Article I pertaining to multimodality in musical interaction, the sensory modality of the communication of emotions might affect the contribution to agency. This surmise arises from the observation that emotions shape meaning and purpose. Although emotional content was not discussed in Article I and therefore it is not explicit in the theoretical framework, it has often been considered to be strongly conveyed by musical gestures (see BACKGROUND, subsection 2.1). Therefore, emotional content may be considered to be an important element of musical interaction. In this regard, the results of the study reported in Article IV on the contribution of sensory modality to the perception of musical emotions, might foster the expansion of said framework.

The results reported in Article IV show that perceived emotion ratings were more consistent among responders when audio was present (audiovisual and only audio conditions). This suggests that music provided cues that most responders interpreted in more unified ways, as opposed to a possibly wider variety of interpretations that may have been made when rating only video.

The post-hoc tests of the analyses of variance (ANOVA) revealed that participants were mostly successfully able to decode the performers' expressive intentions based on both visual information alone, and auditory information alone. In the rating conditions where audio was present, composed emotion had a stronger effect on participant ratings than performers' expressive emotional intentions. The greater effect of composition compared with the effect of performance expression is also verified in the analyses of performance features (auditory and kinematic), where musical mode (whether the piece was major and minor) was at the same time the least correlated feature with other performance features and the dominating predictor of perceived emotions. Also, all multiple regression models for ratings where audio was present had better fit after including mode in the model. These observations reveal the prevalence of mode as a predictor of perceived emotions in music.

The direct relationships observed—except for violin performances when only audio is present—between mode and emotional valence, and between audio energy and Anger, are consistent with previous studies (see

BACKGROUND, subsection 2.3). Other relations found in this study have partial agreement with previous research, such as the relation between amount of movement and ratings for Happiness being direct, and inverse for Sadness, as observed by Dahl and Friberg (2007). In this study, that relation was strongly verified for all the motion-capture marker groups on the violinist when video was presented and weak or inverse when only audio was presented. For ratings of piano performances, that relation was only verified for the movement of the right hand when audio was present.

The cause of the inconsistency of relations between performance features and perceived emotions for violin performances presented as only audio, compared to the other presentation conditions, remains unclear. A case can be made that this effect corresponds to the melodic nature of the violin performances, which lacks the additional information provided by chords in the piano performances. Also, it should be noted that the most correlated motion features for ratings of violin performances when video is presented, are performance speed, variability of performance speed, and energy. These are moderately to strongly correlated to motion time-derivatives. This suggests that responders made their assessment of violin performances with more emphasis on movement, and the assessments of piano performances with more emphasis on sound, the presence of chords in piano performances likely being a substantial factor in this difference. In addition, the amount of movement of the right hand of the piano had an important contribution to the perception of valence. These observations reveal important differences in the way that the different musical instruments and possibly also the musicians, conveyed musical emotions.

While this study led to interesting observations, its limitations should be addressed in future research. First, it was observed that there were substantial differences among the musical instruments. Since each instrument was played by only one musician, the observed differences might also be attributed to the playing of each musician. Therefore, it would be convenient to repeat the experiment using more or other instruments than only a piano and a violin, and more than one musician playing each instrument. Following this pathway might reveal interesting opportunities for the exploration of the idiosyncrasies of musical instruments.

A second limitation of this study is that its findings might not be transferrable to music cultures outside the culture following the line of European classical music. This postulation is based on the participants of the experiment, the music used, and the strong effect of mode on perceived emotions. Regarding the participants, it was taken for granted that they had been raised in circumstances that would have exposed them to European classical music. This assumption was made solely on the fact that all participants were of Finnish nationality. Regarding the music used in the experiment—in the style of European classical music—it is strongly reliant on the tonal system. In such system major and minor mode have strong significations of positive and negative valence, respectively. These significations

might not be found in other musical cultures (Smit et al., 2022). Being that as it may, this limitation can be taken as motivation to repeat the experiment using different kinds of music and participants.

A third limitation of this study is the use of mean responses of perceived emotions. While this is justified by the variation across responses being consistent, the high variability still may cast doubt on the reliability of the results. This limitation may also serve as motivation to redesign the experiment towards obtaining responses with less variability. Moreover, it may serve as a rationale for the inquiry on the causes of variability. Such rationale warranted the study reported in Article V, whose results are discussed in the next subsection.

5.3.2 Effect of personality

Here the results of the study reported in Article V are discussed, following the line of thought originated by the framework proposed in Article I. The formulation of the model at the core of the framework started by looking at musical instruments and how their common aspects may be summarised. Such model highly resembles human embodied cognition, leading to an encompassing model in which humans and machines are agents of musical interaction. It is necessary to bear in mind that models—any model—might not account for all the individual characteristics of the exemplars being modelled. The narrative of Article I illustrates differences between musical instruments but takes advantage of what is common in them to accommodate them for the formulation of a model of machines as musical agents. Likewise, Article IV observes variability in responses but finds enough cohesion to use their means to model the communication of musical emotions. The study reported in Article V provides information to understand one source of variability—personality traits—in the model where humans are musical agents.

The study explored relationships between the Big Five personality traits and embodied emotions in spontaneous movement to music. “Embodied emotions” is the name given to the correlation between emotions and bodily features representing kinematic and non-kinematic characteristics of motion and posture. In the analyses conducted, distinct bodily features were found to embody musical emotions, which then were found to distinctly relate to personality traits. These relationships may be summarised in two clusters of personality traits. The first cluster is composed by Openness, Agreeableness and Extraversion, while the second cluster consists of Conscientiousness and Neuroticism. Embodied emotions are moderately related to traits in the first cluster and weakly related to traits in the second cluster. Special cases are Openness and Neuroticism, having the strongest and weakest relations to embodied emotions, respectively.

The two-cluster pattern with the special case for Neuroticism is remarkably consistent with the results obtained by the meta-analysis conducted by Barańczuk (2018). That study observed the relationship between suppression of expression of emotions to be non-significant and weak for Neuroticism and

Conscientiousness, respectively, while inverse for all other traits. This suggests that the embodiment of musical emotions is related to the suppression of expression of emotions.

The special cases for Openness and Neuroticism in the two-cluster pattern, are consistent with previous studies that investigated personality traits and music preference (Brown, 2012; Delsing et al., 2008; Dobrota & Reić Ercegovac, 2015; Dunn et al., 2011; Fricke & Herzberg, 2017; Nave et al., 2018; Reić Ercegovac et al., 2015; Rentfrow & Gosling, 2003; Schäfer & Mehlhorn, 2017; Upadhyay, Shukla, & Chakraborty, 2017; Vuoskoski & Eerola, 2011a; Zweigenhaft, 2008). These studies found Openness to have the strongest correlations with music preference, followed by Agreeableness and Extraversion. Conscientiousness and Neuroticism had the weakest correlations. Carlson et al. (2017) reported similar results, albeit correlation between music preference and Extraversion was much lower. Other studies have found distinct stronger correlations for Openness, while weaker for all other traits (Cleridou & Furnham, 2014; Langmeyer et al., 2012; Upadhyay, Shukla, Tripathi, & Agrawal, 2017). Additionally, these observations are consistent with research that has found that the preference for music is related to the emotional content of music (Hunter et al., 2011; Ladinig & Schellenberg, 2012; Naser & Saha, 2021; Schäfer & Sedlmeier, 2011) or that has hypothesised it based on the relation between preference and bodily features of spontaneous dance (Luck et al., 2014). Likewise, Openness, Agreeableness, and Extraversion, have been found to be associated with positive correlations between music preference and the strength of emotional response to music, Openness having the strongest association (Liljeström et al., 2012; Nusbaum & Silvia, 2011).

The highest association between music preference and musical emotions, being for trait Openness, is consistent with a variety of related phenomena. Openness has been found to correlate positively with chills when listening to music (McCrae, 2007), awe for music (Silvia et al., 2015), and with the direct relation between liking for sad music and emotions elicited by sad music (Vuoskoski et al., 2012). In addition, Openness has consistently been thought of as being related to the experience of complex and strong emotions as a result of sensitivity to aesthetic experiences (Reisenzein & Weber, 2009; Terracciano et al., 2003). These observations may explain the results presented in the study here discussed. Other similar patterns may be found in previous studies on the relations between Big Five personality traits and trait Empathy. The special case of trait Neuroticism may be related to Empathy as Melchers et al. (2016) and Bamford and Davidson (2019) have observed direct correspondence between the Empathy Quotient and all Big Five traits, except Neuroticism. Those studies and the work by DeYoung et al. (2010) have found Empathy to be strongly and directly related with Agreeableness, suggesting that Empathy is a substantial contributor to the embodiment of emotions. Also, the weak relation between Conscientiousness and embodied emotions, may be explained by this trait being the only Big Five trait not related to emotional dispositions (Reisenzein & Weber, 2009).

One limitation of this study is the sample size and composition. Although it is not easy or cheap to run a study as the one described, the results discussed here should encourage the repetition of the experiment with a different demographic. Another limitation is ratings of perceived emotions made by a group of participants different than the ones that moved to music. Future research could instead use ratings of felt emotions by the same people that moved to the music, and the use of self-chosen music. These recommendations are made because it has been observed that induced emotions affect dancing characteristics more than portrayed emotions (VanDyck et al., 2013), and self-chosen stimuli elicits more intense emotional responses to music (Liljeström et al., 2012). In summary, this study found evidence supporting that musical emotions are embodied differently according to personality traits. The comparison with previous studies suggests that the causes for the embodiment of musical emotions are preference for the music, empathy, and emotional disposition.

5.3.3 Prospect of practical application

Following the overall vision of this dissertation, it is convenient to draft possible practical implications of the findings reported in Articles IV and V, especially those concerned with the development of technologies for making music. The results of Article IV demonstrate different perceptual outcomes for different sensory modalities and for musical instruments, even though there might be commonalities. Correspondingly, the results of Article V demonstrate different relationships between personality traits and embodiment of musical emotions. These observations imply opportunities to challenge existing ways of understanding and exercising musical interaction, taking advantage of the duality of commonalities and individualities of information signals (i.e., sensory modalities) and agents (i.e., people, instruments, machines).

Regarding the similarities and differences of sensory modalities, the effects of juxtapositions have been exploited since the beginning of times. The continuum ranging from coherence to contrast has been the very essence of the musical arts and dance. Thus, awareness of the possibilities of each sensory modality may have a substantial effect in the outcomes of the design of new technologies for making music. For example, the results of Article IV indicate that there is less agreement among people in perceived emotions when only the visual component is present, compared to the condition when the auditory component is present. This knowledge may be used as a guideline in the design of digital musical instruments that expand the concepts presented in Articles VI and VII, from individual use to group music-making.

The relationships between personality traits and embodiment of musical emotions also offer opportunities for innovation. Musical instruments are often not designed considering individual traits, except for those designed for people with disabilities. A similar case is “Accessible Digital Musical Instruments” (Frid, 2019), a concept focusing on inclusion beyond overcoming limitations. The notion of inclusion not only involves people with disabilities. It is

intimately related to easiness of use, of which there are copious and trivial examples. For instance, musical instruments having buttons (e.g., keys) solve the problem of tuning a note with fingers (e.g., the clarinet, then the saxophone having simplified keywork) and may also solve the problem of starting and ending the sound (e.g., the piano, the organ, the multi-purpose digital controller with assignable buttons). Following and extrapolating from this line of thought, the results of Article V show that individual differences might be an important factor in the emotional experience of novel digital musical instruments using broad bodily motion. In light of this, future research on novel musical instruments may consider measurement of personality traits when assessing the users' experience. That might explain the variation of responses indicating emotional engagement. Furthermore, the conjecture posed in Article V, that embodiment of musical emotions might be related to preference for the music, can be linked to the digital instrument described in Article VI. The instrument allows the user to use any recorded audio. Therefore, following said conjecture, it may be hypothesised that the engagement of users is directly related to the preference for the audio recording being used (e.g., favourite song).

The possibilities of digital musical instruments extend beyond the mechanics of control, allowing designers to propose new and sometimes revolutionary musical interaction paradigms. Beyond the question of how much challenge is desirable for a rewarding musical experience, it is possible to see that the design of musical instruments may – and sometimes should – take into account the individual differences of users. The concomitant research questions appear as obvious.

How the engagement with musical devices is affected by personality traits? Why not design musical devices considering the individuality of their users? Can we make musical devices capable of learning characteristics of people? Could musical devices understand us?

5.4 A holistic model of mimetic musical interaction

The essay of Article I built upon theoretical constructs and empirical research in the literature at the time it was written, concerning musical interaction, musical instruments, embodied cognition, musical gestures, and agency. That essay formulated a model of embodied musical interaction, which served as a framework for the research carried out in the subsequent studies described in Articles II to VII. That research focused on specific and previously unexplored facets of the model and devised novel practical applications following the proposed framework. The new knowledge produced by the research, added to related literature published after the publication of Article I, prompts to revisit the model of embodied musical interaction. The following paragraphs outline how the model is thus substantiated and how it could be extended.

Articles II and III describe a method that can be used as a model for human segmentation of bodily motion. The method exploits self-similarity

patterns in time-series signals. While it had been previously used for segmentation of audio, in the study corresponding to Articles II and III it was applied for the segmentation of bodily motion. The fact that the same mechanism can be used for segmentation of signals carrying information corresponding to different perceptual modalities, supports the idea that gestures are multimodal chunks of information that flow within and amongst participants (humans and machines) of musical interaction. This inference is further supported with the observations made in the testing of the system described in Article VI, in which participants successfully engaged in the interaction paradigm involving unsupervised segmentation. The notion that human perception involves division of big chunks of information and clustering of small chunks of information is not new. However, the study corresponding to Article VII, shows a fresh example. This is observed in the leveraging of the hierarchical structuring of human cognition and the power-law dynamics of music and bodily motion, which facilitated the application of the segmentation method based on self-similarity, to the transformation of daily activity into music.

The apparent ubiquity of hierarchical structuring of information based in self-similarity reinforces the notion of mimesis, in the sense of imitation. It is also related to the idea of mimesis as expression of emotions through bodily gesture (see BACKGROUND, subsection 2.1). The latter may be seen as a subset of the former, broader emergent postulation that gestures, in whichever form they are realised (e.g., bodily motion or posture, sound, shape, abstract), follow a hierarchical organisation that permeates all human cognition. These gestures, in turn, may communicate emotions, among other things (discussed below). These conceptualisations also let musical instruments (in the broadest sense, as discussed throughout this dissertation) be extensions of the human body in such a way that the human body itself is embodied in the larger network of interactions with other humans and other musical instruments, or more generally, musical machines.

The discussed model of embodied musical interaction does not explicitly incorporate emotions. However, it seems convenient to include musical emotions in the understanding of the model because of their fundamental role in the purposes of music, chiefly related to communication and engagement. As mentioned above, gestures may carry meaning through different modalities (e.g., sound, vision, touch) but the model of embodied musical interaction is broader as it considers that gestures are embedded in signals. As such, signals may carry information corresponding to sensory modalities, but the signals may be electrical or mechanical (including acoustical). For example, electrical signals might be brain signals or electronic circuitry, while mechanical signals might be movement of the human body or musical instruments. It is important to distinguish meaning from the gestures and from the signals. The hierarchical order in which these are organised is signals, that carry gestures, that carry meaning. Then, meaning can be composed by direct signification and by emotional signification (see BACKGROUND, subsections 2.1 and 2.3).

The studies reported in Articles IV and V contribute to the understanding of musical emotions, which are part of meaning and therefore agency, in the context of the model for embodied musical interaction. The results reported by Article IV provide evidence that the distinct information carried by different sensory modalities has an effect in the perception of musical emotions. This complements the model of embodied musical interaction by remarking that gestural agency may be affected by the sensory modality of the interaction. We may simplistically think that along the pathways of the musical interaction network the information is only transformed. However, the experimental evidence shows that some information may be preserved or lost, sensory modality being a major factor. This shall be regarded as a fundamental property of the model.

The results reported by Article V provide evidence that personality traits are related to the embodiment of musical emotions. Possible causes for the embodiment of musical emotions are musical preference, empathy, and emotional disposition. These possible causes evince that musical emotions are not only shown (voluntarily or not) through bodily movement, but that they are related to significant and immanent characteristics of people. In turn, these individual characteristics are arguably essential in musical communication. Preference for music is intimately related to the memories that preferred music evokes. It relates to identity and of belonging when those preferences are shared with others. Empathy and emotional disposition facilitate understanding and therefore connection.

The mimetic aspect of this phenomenon is evident: musical emotions are exhibited in bodily motion. In other words, the emotion content of one medium (sound) is observable in other medium (the body). The evidence might not be always available to the naked eye, but the described research shows that it is possible to measure it. The evidence for differential embodiment of emotion adds to the body of knowledge on the effect of personality traits on the feeling of musical emotions. In simple terms, it remarks the differences among people. This is a trivial observation, but it is worthwhile to note that in the context of the embodied musical interaction model, people are agents as much as musical instruments. The variety in the composition of these agents might play an essential role in the occurrence of musical interaction and music in general. Thus, this can be deemed as another fundamental property of the model. If for a moment we allow ourselves to indulge in reveries, we may find an analogy with the second law of thermodynamics (see Schrödinger, 1944). The analogy is observed in that the inequality of agents of musical interaction may be a prerequisite for the exchange of information, in the same way heat is transferred from a warm object to a connected cold object. Further into the musing, we may also realise that a system, in its process of attaining maximum entropy, may arrive at local equilibria. Examples of this are objects displaying patterns such as fractals, crystals, solar systems, galaxies, life in its various embodiments, music, and so on. This principle may be the fundamental cause of what has been called enactive behaviour, and thus the driving force of

musical interaction as much as it may be the driving force of life and the universe at large.

Returning to more concrete considerations, Article I was published as a chapter of a book on “embodied music interaction” (Lesaffre et al., 2017), where interaction is noted to be a crucial component of the Embodied Music Cognition theoretical corpus. In fact, in the first chapter of that book (p. 13), Leman proposes a model of interaction with music in which assessments of music occur in real time in order to predict actions or states that could have caused such sounds in the way they are presented, or the “expressive character of the music”. It is argued that this assessment contributes to the formation of meaning and that the body plays an essential role in this process. A further article attempted to define “embodied music cognition” (Leman et al., 2018), again having the notion of interaction at its core, based on the idea that the role of cognition is to build predictive models and apply such models in interactions. The building of the models is said to be dependent on corporeal mediators and internal states, which is a similar proposition to the ones made in Article I. The difference is the treatment of the elements of the model. Leman describes them as states, whereas Article I discusses internal states as processes propelled by enactive and feedback multimodal signals. The ideas proposed by Leman and colleagues are connected, albeit somehow loosely, to the properties of multimodal musical communication as discussed above. For example, some information about the human body (including emotions) might be transferred to the listener through musical sound.

The model of embodied musical interaction proposed in Article I puts an emphasis in human-machine musical interaction. Two years after the article was published, Tanaka (2019) used the term “Embodied Musical Interaction” as a derivative of the concept “embodied interaction”. This term was used for a discussion on how musical human-computer interaction may benefit from three generally accepted paradigms of Human-Computer Interaction (Fallman, 2003). These paradigms are approaches to design: conservative, pragmatic, and romantic. The proposition is that these paradigms characterise designer, problem, product, process, knowledge, and role model involved in the design. For example, in a conservative approach a problem is ill-defined (to be defined) and a product is the result of the process. In a pragmatic approach the problem is unique to the situation and the product is integrated in the world. In a romantic approach the problem is subordinate to the final product and the product is artwork. The ideas expounded by Tanaka are pertinent to the application of the model for gestural agency in human-machine musical interaction. For example, it could be said that the study corresponding to Article VI is mostly conservative, while that of Article VII is mostly pragmatic.

As it has been proposed throughout this dissertation, a long-term vision of the described research is the design and implementation of musical systems as agents. This postulation resonates with the concept “Emergent Interaction” (Murray-Browne & Tigas, 2021). This concept is an approach to the design of digital musical instruments, which seems to follow the romantic paradigm as

described above. The concept is rooted in the notion that “unsupervised machine learning allows representations to emerge directly from the situation in which interaction happens”. Those “emergent representations allow us to create expressive gestural interactions without explicitly declaring input or output”. The cited article further describes a system for the sonification of dance. That system applies the proposed concept although the machine learning of the system is not fully unsupervised. Nonetheless, the independence of the user from the assumptions of the system’s designer is stressed as a motivation for further research. Further connections of the embodied musical interaction model with recent literature may be seen in works that had cited Article I at the time of writing of this dissertation. These works have taken the concept of agency amongst humans and machines, as a theoretical background for the exploration and design of digital musical systems involving control by bodily movement (Christoffersson, 2018; Erdem et al., 2020; Morand, 2019; Staudt et al., 2022; Oriolo, 2019).

To summarise, the resulting updated model of embodied musical interaction is holistic, integrating multimodal signals as well as human and non-human agents. One key property of the model is that signals are organised in a hierarchical temporal structure. Another key property is that agents and signals are diverse in kind and in time (e.g., signals might acquire or lose information, agents might change), and this diversity affects the functioning of interaction. Arguably these properties lead to the richness in form and substance that we may experience when engaging in musical activity.

5.5 Concluding remarks

The research project of this dissertation started with the goal of investigating the mimetic relationships between music and the movement of the human body. The resulting research produced firstly a holistic theoretical model of embodied musical interaction. This model was used as a framework for the subsequent research, aimed to measure aspects of the theoretical framework. The first aspect of measurement was the temporal segmentation of bodily motion, for which an automatic method was formulated and tested. The second aspect of measurement was embodiment of musical emotions, for which the contribution of visual and auditory sensory modalities, as well as personality traits, were quantified and modelled. Finally, the method for automatic temporal segmentation was applied to a gesturally controlled digital musical instrument, and to a system for musical sonification of daily activity.

Directions for future research can be summarised in two broad and possibly overlapping tracks. The first track comprises the understanding of music as a phenomenon. As it was discussed, it is convenient to look at music as an interaction process. Thus, a holistic model of musical interaction should consider the underlying nature of interactions, for which I suggested that the second law of thermodynamics might be, if not a precise explanation, a source

of inspiration. Beyond the theoretical work, as a matter of course, repetition of the experiments and exploration of different methodologies is necessary. The most urgent practical aspect for the advancement of the experimental research, is the collection of more and more diverse data. Specifically, optical and accelerometry motion-capture, audio, video, annotation of segmentation and daily activity, perceived and felt emotions, personality traits and other personal information, of experts and non-experts, playing instruments, dancing, and living their daily lives.

The second track for future research is the application of knowledge of musical phenomena, especially the newest knowledge, to solve real-world problems. Given that the context is music, this encompasses the exploration of novel ways of making music, including musical instruments. In this regard, the multigranular online temporal segmentation method described in this dissertation is a key piece for the implementation of fully unsupervised gesture-learning systems. This would ultimately result in intelligent automatic musical agents, perhaps the next evolutionary stage of musical instruments.

To conclude, I shall add that this dissertation reflects in an orderly manner the main lessons learned along a journey that started with a desire to learn more about the relations between music and the human body. The journey was not always orderly, sometimes diverting to further explore questions that arose along the way. This led to shape the dissertation in its final form, resulting in, on one hand, the considerably broad holistic theory for musical interaction. On the other hand, it resulted in the examination of the narrow and disparate problems of unsupervised temporal segmentation and the measurement of different facets of embodied musical emotions. In the course of such exploration other topics were investigated as well, but much of the work done failed to produce anything worthy of being reported. Likewise, I have failed to fully answer the big question “How and why does music move us?”. Even so, I have offered some ideas that might be useful to understand how music brings people to move together in rejoice at a concert or in grief at a funeral, or why some people at a dance show might move to the music as if they were part of the performance while some others might prefer to stay still, or why a baby stops crying when mommy sings and sways. These ideas may kindle dreaming of and coming up with new technologies to help athletes to lift heavier, non-athletes to move more, rockstars to arouse utter rapture and non-rockstars to explore different ways of experiencing the joy of making music.

YHTEENVETO (SUMMARY IN FINNISH)

Mimeettisten kehon liikkeiden ja musiikillisten rakenteiden suhteet: teoria, mittaus ja soveltaminen

Tämä väitöstutkimus lähti liikkeelle tavoitteesta tutkia ihmiskehon liikkeen ja musiikillisen rakenteen välisiä suhteita. Tutkimus aloitettiin teoreettisen viitekehyksen kehittämällä, jonka jälkeen suoritettiin kvantitatiivisia mittauksia tämän viitekehyksen eri näkökulmista. Lopulta luotiin teorian ja mittausten sovelluksia. Seuraavissa kappaleissa esitetään yhteenveto tuloksista ja ehdotetaan suuntaa tulevalle tutkimukselle.

Teoreettisena kontribuutiona on malli keholliselle musiikilliselle vuorovaikutukselle, jonka keskiössä on ihmisten ja koneiden (eli soittimien) välinen toimijuus. Tässä mallissa toimijat kommunikoivat multimodaalisiin (esim. auditiivisiin ja visuaalisiin) signaaleihin upotetuilla eleillä. Tällaisten eleiden ajallinen rakenne on hierarkkinen, jossa lyhyemmät eleet on upotettu pidempi kestoisiin eleisiin. Tutkimuksen mittaus- ja sovellusosien tulosten perusteella voitiin ehdottaa malliin parannuksia, jotka koskivat signaalien ja toimijoiden moninaisuutta vuorovaikutuksen keskeisinä ominaisuuksina. On selvää, että tulevissa tutkimuksissa mallia olisi testattava eri tavoin, jotta nähdään voidaanko se kumota.

Mittaukseen liittyviä kontribuutioita on kaksi. Ensimmäinen käsittelee kehon liikkeiden segmentointia ja toinen kehollisia musiikillisia tunteita. Ensimmäinen kontribuutio liittyy antureilla mitatun kehon liikkeen automaattista ajallista segmentointia koskevan menetelmän kehittämiseen. Menetelmän tärkeitä ominaisuuksia ovat kyky havaita muutos yhdestä eleestä toiseen, malli on ohjaamaton eikä tarvitse aiempaa tietoa eleistä, sen tulokset voivat vastata hyvin ihmisen havaitsemaa segmentointia ja se voidaan toteuttaa toimimaan reaaliajassa. Tulevassa tutkimuksessa olisi hyvä harkita havaintoarvioinnin protokollan parantamista: segmentointirajojen merkitseminen, laskeminen ja laskettujen tulosten arviointi olisi tehtävä yhdellä kertaa. Parannettu protokolla on hyödyllinen myös minkä tahansa ajallisen segmentointimenetelmän arvioinnissa.

Tutkimuksella oli kaksi tunteiden mittaamiseen liittyvää kontribuutiota. Ensimmäinen on sen kvantifiointi miten musiikkiesityksen visuaaliset ja auditiiviset komponentit vaikuttavat tunteiden havaitsemiseen. Suoritettiin koe, jonka tuloksena osallistujien havainnot tunteistaan olivat samansuuntaisia, kun kuulokomponentti oli läsnä (verrattuna vain visuaaliseen komponenttiin), ja tonaliteetti (eli duuri tai molli) oli merkittävin tekijä. Vastaavasti kun esitettiin vain visuaalista informaatiota, osallistujien vastausten välinen samanlaisuus oli heikko, vaikkakin jotkin visuaaliset ominaisuudet olivat edelleen merkittävästi yhteydessä havaittuihin tunteisiin. Tutkimuksen tärkeimpiä rajoituksia ovat osallistujien vähäinen määrä ja erityinen kulttuuritausta, vain kahden soittimen ja esittäjän (piano ja viulu) käyttö sekä testattavien elementtien kompleksisuus

(sävelletyt, ilmaistut ja havaitut tunteet). Tulevassa tutkimuksessa olisi pyrittävä toteuttamaan kokeen muunnelmia, joissa nämä rajoitukset otetaan huomioon. Toinen musiikillisten tunteiden mittaamiseen liittyvä kontribuutio kvantifioi persoonallisuuspiirteiden vaikutusta siihen, missä määrin musiikissa koetut tunteet heijastuvat spontaaniin tanssiin. Tulokset osoittavat, että tämä yhteys pätee pääasiassa avoimuuden piirteeseen ja hyvin heikosti tunnollisuuden ja neuroottisuuden piirteisiin, kun taas yhteys sovinnollisuuden ja ekstroversion piirteisiin on kohtalainen. Aiemmat tutkimukset huomioon ottaen nämä tulokset viittaavat siihen, että musiikillisten tunteiden kehollistumisen syitä ovat mieltymys musiikkiin, empatia sekä emotionaalinen taipumus. Tämän tutkimuksen merkittävä rajoitus on osallistujien vähäinen määrä. Tulevassa tutkimuksessa koe tulisi toistaa eri kohortilla, jotta tulosten tilastollinen voima parani.

Soveltamiseen liittyvät kontribuutiot sisältävät kaksi järjestelmää, jotka osoittavat ajallisen segmentointimenetelmän toteutettavuuden sen käyttämiseksi musiikin tekemisessä. Konkreettisina tuloksina on kaksi ohjelmistoa ja järjestelmien alustavaa testausta. Ensimmäisessä järjestelmässä hyödynnetään menetelmän kykyä toimia reaaliajassa. Järjestelmää kehitettäessä oli otettava huomioon segmentointimenetelmän väistämätön viivästynyt vaste. Tämä johti kokeiluun hyödyntää ihmiskehon laajoja liikkeitä ja musiikkiäänien viivästynyttä ohjausta. Lisäksi järjestelmän havaittiin tukevan osallistavaa käyttäjäkokemusta. Tulevaisuuden tutkimuskohteet voidaan jakaa neljään polkuun, jotka voivat olla risteäviä. Ensimmäinen on tekninen kehitys, ja siihen kuuluvat laskennallisen tehokkuuden parantaminen, monikranulaarisen segmentoinnin toteuttaminen ja erilaisten muototekijöiden kokeileminen. Toinen on ohjaamattoman segmentointiparadigman sisällyttäminen koneoppimisjärjestelmään, joka oppii sujuvasti ilman opetusaineistoa. Vaikka toteutukseen liittyy useita haasteita, idea on pohjimmiltaan yksinkertainen, sillä siinä käytetään ohjaamatonta segmentointimenetelmää minkä tahansa sopivan koneoppimisjärjestelmän (esim. neuroverkon) esikäsittelevä vaiheena. Kolmas suunta tulevaisuuden tutkimukselle on suoraviivaisempi ja se voidaan helposti toteuttaa tuotetusta ohjelmistosta muuttamalla sen käytettävissä olevia parametreja (esim. käyttämällä erilaisia äänitiedostoja, säätämällä segmentointiparametreja), muokkaamalla itse ohjelmistoa (esim. korvaamalla äänentoisto synteeseillä) tai käyttämällä erilaisia ohjauslaitteita (esim. käyttämällä jotakin muuta anturia kuin kiihtyvyysanturia). Neljäs tulevaisuuden tutkimussuunta on musiikillisen vuorovaikutuksen järjestelmällinen tutkiminen kehon laajojen liikkeiden ja viivästetyn ohjauksen avulla.

Toisessa sovelluksessa hyödynnetään segmentointimenetelmän kykyä toimia eri aikaskaaloilla. Tätä kykyä hyödynnettiin päivittäisen toiminnan segmenttien tunnistamiseen kehoon kiinnitetyistä liikeantureista saatujen tietojen avulla. Nämä segmentit edustavat päivittäisen toiminnan tapahtumia (esim. nukkuminen, aamurutiinit, kävely ruokakauppaan jne.), ja niitä käytettiin tuottamaan automaattisesti musifikaatioita (lyhyitä musiikkikappaleita), jotka edustavat kyseisiä tapahtumia ja niitä vastaavaa toiminnan määrää. Laajamittainen tutkimus osoitti, että musiikkikappaleet voivat tehokkaasti edustaa sitä, kuinka

aktiivinen henkilö on. Näin ollen järjestelmää ehdotetaan apuvälineeksi terveydenhuoltoon vähentämään sedentarismia. Jatkotutkimus voitaisiin toteuttaa suunnilleen kolmessa peräkkäisessä vaiheessa. Ensimmäisessä vaiheessa tuotetaan toimiva prototyyppi, joka koostuu helposti saatavilla olevasta puettavasta teknologiasta ja mobiililaitteesta sekä ohjelmistosta, joka voi tuottaa musifikaatioita ja ladata segmentointitietoja palvelimelle. Toisessa vaiheessa järjestelmää testataan todellisella väestöllä, kuten potilailla, joilla on diagnosoitu sedentarismi. Tässä vaiheessa olisi myös kerättävä tietoja. Anturitietojen lisäksi järjestelmän pitäisi antaa käyttäjille mahdollisuus kirjata päivittäiset tapahtumansa. Kolmannessa vaiheessa analysoidaan tietoja käyttäjäkokemuksesta, kliinisiä tuloksia (jos niitä käytetään kliinisen intervention apuvälineenä) ja sitä, millä asteella segmentoidut anturitiedot ja kirjatut päivittäiset tapahtumat vastaavat toisiaan.

Yhteenvedona voidaan todeta, että tämä tutkimushanke on tuottanut uutta tietoa musiikin ja ihmiskehon välisistä suhteista. Tutkimuksen jatkaminen voi tapahtua jommalla kummalla kahdesta mahdollisesti päällekkäisestä suunnasta. Ensimmäinen suunta käsittää tutkimusta, jonka tavoitteena on ymmärtää paremmin kehon roolia musiikillisessa aktiivisuudessa. Toinen suunta käsittää tiedon soveltamisen musiikin tekemiseen liittyvien uusien teknologioiden kehittämiseen, jotka voivat olla esimerkiksi terveydenhuollon apuvälineitä tai älykkäitä soittimia ja jotka voivat auttaa ratkaisemaan reaali maailman ongelmia.

REFERENCES

- Abeles, H. F. & Chung, J. W. 1996. Responses to Music. In D. Hodges (Ed.), *Handbook of Music Psychology* (pp. 285-342). IMR Press.
- Agrawal, Y., Jain, S., Carlson, E., Toiviainen, P., & Alluri, V. (2020). Towards Multimodal MIR: Predicting individual differences from music-induced movement. In *Proceedings of the 21st International Society for Music Information Retrieval Conference*, Montréal, Canada.
<https://doi.org/10.5281/zenodo.4245368>
- Ahad, M. A. R., Tan, J. K., Kim, H. S., & Ishikawa, S. (2008). Human activity recognition: Various paradigms. In *2008 international Conference on Control, Automation and Systems* (pp. 1896-1901). IEEE.
<https://doi.org/10.1109/ICCAS.2008.4694407>
- Ahmed, F., Bari, A. H., & Gavrilova, M. L. (2019). Emotion recognition from body movement. *IEEE Access*, 8, 11761-11781.
<https://doi.org/10.1109/ACCESS.2019.2963113>
- Althouse, A. D. (2016). Adjust for multiple comparisons? It's not that simple. *The Annals of thoracic surgery*, 101(5), 1644-1645. [https://www.annalsthoracicsurgery.org/article/S0003-4975\(15\)01873-1/fulltext](https://www.annalsthoracicsurgery.org/article/S0003-4975(15)01873-1/fulltext)
- Aminikhanghahi, S., & Cook, D. J. (2017). A survey of methods for time series change point detection. *Knowledge and information systems*, 51(2), 339-367.
<https://doi.org/10.1007/s10115-016-0987-z>
- Anderson, M. L. (2003). Embodied cognition: A field guide. *Artificial Intelligence*, 149(1), 91-130. [https://doi.org/10.1016/S0004-3702\(03\)00054-7](https://doi.org/10.1016/S0004-3702(03)00054-7)
- Armstrong, N. (2006). *An enactive approach to digital musical instrument design* [Doctoral dissertation, Princeton University].
<https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=71be05be9c05b41bb21a1ab5266318aedee16146>
- Bamford, J. M. S., & Davidson, J. W. (2019). Trait Empathy associated with Agreeableness and rhythmic entrainment in a spontaneous movement to music task: Preliminary exploratory investigations. *Musicae Scientiae*, 23(1), 5-24. <http://dx.doi.org/10.1177/1029864917701536>
- Barańczuk, U. (2019). The five factor model of personality and emotion regulation: A meta-analysis. *Personality and Individual Differences*, 139, 217-227. <https://doi.org/10.1016/j.paid.2018.11.025>
- Barbič, J., Safonova, A., Pan, J. Y., Faloutsos, C., Hodgins, J. K., & Pollard, N. S. (2004). Segmenting motion capture data into distinct behaviors. In *Proceedings of Graphics Interface 2004* (pp. 185-194).
<https://dl.acm.org/doi/10.5555/1006058.1006081>
- Battcock, A., & Schutz, M. (2019). Acoustically expressing affect. *Music Perception*, 37(1), 66-91. <https://doi.org/10.1525/mp.2019.37.1.66>

- Bernard, J., Dobermann, E., Vögele, A., Krüger, B., Kohlhammer, J., & Fellner, D. W. (2017). Visual-interactive semi-supervised labeling of human motion capture data. In *Proceedings of IS&T International Symposium on Electronic Imaging: Visualization and Data Analysis* (pp. 34-45).
<https://doi.org/10.2352/ISSN.2470-1173.2017.1.VDA-387>
- Bernier, E., Chellali, R., & Thouvenin, I. M. (2013, August). Human gesture segmentation based on change point model for efficient gesture interface. In *2013 IEEE RO-MAN: The 22nd IEEE International Symposium on Robot and Human Interactive Communication* (pp. 258-263). IEEE
<https://doi.org/10.1109/ROMAN.2013.6628456>
- Bevilacqua, F., Zamborlin, B., Sypniewski, A., Schnell, N., Guédy, F., & Rasamimanana, N. (2010). Continuous Realtime Gesture Following and Recognition. In S. Kopp & I. Wachsmuth (Eds.) *Gesture in Embodied Communication and Human-Computer Interaction*. (pp. 73-84). Springer.
https://doi.org/10.1007/978-3-642-12553-9_7
- Bläsing, B.E. (2015). Segmentation of dance movement: effects of expertise, visual familiarity, motor experience and music. *Frontiers in psychology* 5, 1500. <https://doi.org/10.3389/fpsyg.2014.01500>
- Bosi, M., & Jordà, S. (2012). Towards fast multi-point force and hit detection in tabletops using mechanically intercoupled force sensing resistors. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. <http://doi.org/10.5281/zenodo.1178217>
- Bown, O., Eldridge, A., & McCormack, J. (2009). Understanding interaction in contemporary digital music: From instruments to behavioural objects. *Organised Sound*, 14(2), 188–196.
<https://doi.org/10.1017/S1355771809000296>
- Brickwood, K. J., Watson, G., O'Brien, J., & Williams, A. D. (2019). Consumer-based wearable activity trackers increase physical activity participation: systematic review and meta-analysis. *JMIR Mhealth and Uhealth*, 7(4), e11819. <https://doi.org/10.2196/11819>
- Brown, R. A. (2012). Music preferences and personality among Japanese university students. *International Journal of Psychology*, 47(4), 259–268.
<https://doi.org/10.1080/00207594.2011.631544>
- Buckingham, S. A., Williams, A. J., Morrissey, K., Price, L., & Harrison, J. (2019). Mobile health interventions to promote physical activity and reduce sedentary behaviour in the workplace: a systematic review. *Digital health*, 5, 2055207619839883. <https://doi.org/10.1177/2055207619839883>
- Burger, B., Polet, J., Luck, G., Thompson, M. R., Saarikallio, S., & Toiviainen, P. (2013). Investigating relationships between music, emotions, personality, and music-induced movement. In *The 3rd International Conference on Music & Emotion*. University of Jyväskylä, Department of Music.
<http://urn.fi/URN:NBN:fi:ju-201305281794>
- Burger, B., Saarikallio, S., Luck, G., Thompson, M. R., & Toiviainen, P. (2013). Relationships between perceived emotions in music and music-induced movement. *Music Perception*, 30(5), 517-533.
<https://doi.org/10.1525/mp.2013.30.5.517>

- Burger, B., Thompson, M. R., Luck, G., Saarikallio, S., & Toiviainen, P. (2012). Music moves us: Beat-related musical features influence regularity of music-induced movement. In *Proceedings of the 12th International Conference in Music Perception and Cognition and the 8th Triennial Conference of the European Society for the Cognitive Sciences for Music* (pp. 183-187). http://icmpc-escom2012.web.auth.gr/files/papers/183_Proc.pdf
- 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, 183. <https://doi.org/10.3389/fpsyg.2013.00183>
- Burger, B., Thompson, M. R., Luck, G., Saarikallio, S. H., & Toiviainen, P. (2014). Hunting for the beat in the body: on period and phase locking in music-induced movement. *Frontiers in human neuroscience*, 8, 903. <https://doi.org/10.3389/fnhum.2014.00903>
- Cajita, M. I., Kline, C. E., Burke, L. E., Bigini, E. G., & Imes, C. C. (2020). Feasible but not yet efficacious: a scoping review of wearable activity monitors in interventions targeting physical activity, sedentary behavior, and sleep. *Current epidemiology reports*, 7, 25-38. <https://doi.org/10.1007/s40471-020-00229-2>
- Cambridge University Press & Assessment. (n.d.). Gesture. In *Cambridge dictionary*. <https://dictionary.cambridge.org/dictionary/english/gesture>
- Camurri, A., De Poli, G., Leman, M., & Volpe, G. (2001). A multi-layered conceptual framework for expressive gesture applications. In *Proceedings of the MOSART workshop*. <http://mtg.upf.edu/mosart/papers/p45.pdf>
- Camurri, A., Mazzarino, B., Ricchetti, M., Timmers, R., & Volpe, G. (2004). Multimodal analysis of expressive gesture in music and dance performances. In A. Camurri & G. Volpe (Eds.) *Gesture-Based Communication in Human-Computer Interaction* (pp. 20-39). Springer. https://doi.org/10.1007/978-3-540-24598-8_3
- Caramiaux, B., Françoise, J., Schnell, N., & Bevilacqua, F. (2014). Mapping through listening. *Computer Music Journal*, 38(3), 34-48. <https://muse.jhu.edu/article/554086/pdf>
- Caramiaux, B., & Tanaka, A. (2013). Machine learning of musical gestures. In *Proceedings of the International Conference on New Interfaces for Musical Expression* (pp. 513-518). <http://doi.org/10.5281/zenodo.1178490>
- Carlson, E., Burger, B., London, J., Thompson, M. R., & Toiviainen, P. (2016). Conscientiousness and extraversion relate to responsiveness to tempo in dance. *Human Movement Science*, 49, 315-325. <https://doi.org/10.1016/j.humov.2016.08.006>
- Carlson, E., Saari, P., Burger, B., & Toiviainen, P. (2017). Personality and musical preference using social-tagging in excerpt-selection. *Psychomusicology: Music, Mind, and Brain*, 27(3), 203-212. <http://urn.fi/URN:NBN:fi:jyu-201801091122>

- Castellano, G., Mortillaro, M., Camurri, A., Volpe, G., & Scherer, K. (2008). Automated analysis of body movement in emotionally expressive piano performances. *Music Perception*, 26(2), 103-119.
<https://doi.org/10.1525/mp.2008.26.2.103>
- Chastin, S. F. M., & Granat, M. H. (2010). Methods for objective measure, quantification and analysis of sedentary behaviour and inactivity. *Gait & posture*, 31(1), 82-86. <https://doi.org/10.1016/j.gaitpost.2009.09.002>
- Christoffersson, E. (2018). *Jamming with the Plot Twister: Designing virtual worlds for improvised performances* [Master's thesis, Dalarna University].
<http://urn.kb.se/resolve?urn=urn:nbn:se:du-28695>
- Cleridou, K., & Furnham, A. (2014). Personality correlates of aesthetic preferences for art, architecture, and music. *Empirical Studies of the Arts*, 32(2), 231-255. <https://doi.org/10.2190/EM.32.2.f>
- Cloninger, C. R. (1987). A systematic method for clinical description and classification of personality variants: A proposal. *Archives of General Psychiatry*, 44(6), 573-588.
https://www.researchgate.net/publication/19581432_A_Systematic_Method_for_Clinical_Description_and_Classification_of_Personality_Variants_A_Proposal
- Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI): Professional manual*. Psychological Assessment Resources.
- Coutinho, E., & Cangelosi, A. (2011). Musical emotions: predicting second-by-second subjective feelings of emotion from low-level psychoacoustic features and physiological measurements. *Emotion*, 11(4), 921.
https://livrepository.liverpool.ac.uk/3002881/1/CoutinhoCangelosi2011_Accepted4Publication.pdf
- Creaser, A. V., Cledes, S. A., Costa, S., Hall, J., Ridgers, N. D., Barber, S. E., & Bingham, D. D. (2021). The acceptability, feasibility, and effectiveness of wearable activity trackers for increasing physical activity in children and adolescents: a systematic review. *International journal of environmental research and public health*, 18(12), 6211.
<https://doi.org/10.3390/ijerph18126211>
- Dahl, L. (2015). Studying the timing of discrete musical air gestures. *Computer Music Journal*, 39(2), 47-66. <https://muse.jhu.edu/article/583487>
- Dahl, S., & Bresin, R. (2001). Is the player more influenced by the auditory than the tactile feedback from the instrument. In *Proceedings of the COST G-6 Conference on Digital Audio Effects*.
<https://www.speech.kth.se/prod/publications/files/760.pdf>
- Dahl, S., & Friberg, A. (2004). Expressiveness of musician's body movements in performances on marimba. In A. Camurri & G. Volpe (Eds.) *Gesture-Based Communication in Human-Computer Interaction* (pp. 479-486). Springer.
https://doi.org/10.1007/978-3-540-24598-8_44

- Dahl, S., & Friberg, A. (2007). Visual perception of expressiveness in musicians' body movements. *Music Perception*, 24(5), 433-454. <https://doi.org/10.1525/mp.2007.24.5.433>
- Dalla Bella, S., Peretz, I., Rousseau, L., & Gosselin, N. (2001). A developmental study of the affective value of tempo and mode in music. *Cognition*, 80(3), B1-B10. [https://doi.org/10.1016/s0010-0277\(00\)00136-0](https://doi.org/10.1016/s0010-0277(00)00136-0)
- Delsing, M. J., Ter Bogt, T. F., Engels, R. C., & Meeus, W. H. (2008). Adolescents' music preferences and personality characteristics. *European Journal of Personality*, 22(2), 109-130. <https://doi.org/10.1002/per.665>
- Demos, A. P., & Chaffin, R. (2018). How music moves us: Entraining to musicians' movements. *Music Perception*, 35(4), 405-424. <https://doi.org/10.1525/mp.2018.35.4.405>
- Devi, R., Coomaraswamy, A. K., & Tagore, R. (1913). *Thirty Songs from the Panjab and Kashmir*. Old Bourne Press. <http://hdl.handle.net/1802/17483>
- DeYoung, C. G., Hirsh, J. B., Shane, M. S., Papademetris, X., Rajeevan, N., & Gray, J. R. (2010). Testing predictions from personality neuroscience: Brain structure and the big five. *Psychological Science*, 21(6), 820-828. <https://doi.org/10.1177/0956797610370159>
- DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: 10 aspects of the Big Five. *Journal of Personality and Social Psychology*, 93(5), 880-896. <https://www.jordanbpeterson.com/docs/230/2014/15DeYoung.pdf>
- Dixon, T. (2012). "Emotion": The history of a keyword in crisis. *Emotion Review*, 4(4), 338-344. <https://psycnet.apa.org/doi/10.1177/1754073912445814>
- Dobrota, S., & Reić Ercegovac, I. (2015). The relationship between music preferences of different mode and tempo and personality traits—implications for music pedagogy. *Music Education Research*, 17(2), 234-247. <https://doi.org/10.1080/14613808.2014.933790>
- Dreher, C. R., Kulp, N., Mandery, C., Wächter, M., & Asfour, T. (2017). A framework for evaluating motion segmentation algorithms. In *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)* (pp. 83-90). IEEE. <https://doi.org/10.1109/HUMANOIDS.2017.8239541>
- Dunn, P. G., de Ruyter, B., & Bouwhuis, D. G. (2011). Toward a better understanding of the relation between music preference, listening behavior, and personality. *Psychology of Music*, 40(4), 411-428. <https://doi.org/10.1177/0305735610388897>
- Eaves, D. L., Griffiths, N., Burrige, E., McBain, T., & Butcher, N. (2020). Seeing a drummer's performance modulates the subjective experience of groove while listening to popular music drum patterns. *Musicae Scientiae*, 24(4), 475-493. <https://doi.org/10.1177/1029864919825776>
- Eerola, T., Friberg, A., & Bresin, R. (2013). Emotional expression in music: contribution, linearity, and additivity of primary musical cues. *Frontiers in psychology*, 4, 487. <https://doi.org/10.3389/fpsyg.2013.00487>

- Eerola, T., Lartillot, O., & Toiviainen, P. (2009). Prediction of Multidimensional Emotional Ratings in Music from Audio Using Multivariate Regression Models. In *International Society for Music Information Retrieval Conference* (pp. 621-626). <https://archives.ismir.net/ismir2009/paper/000121.pdf>
- Eerola, T., & Vuoskoski, J. K. (2011). A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music*, 39(1), 18-49. <https://doi.org/10.1177/0305735610362821>
- Eerola, T., Vuoskoski, J. K., & Kautiainen, H. (2016). Being moved by unfamiliar sad music is associated with high empathy. *Frontiers in psychology*, 1176. <https://doi.org/10.3389/fpsyg.2016.01176>
- Eerola, T., Vuoskoski, J. K., Kautiainen, H., Peltola, H. R., Putkinen, V., & Schäfer, K. (2021). Being moved by listening to unfamiliar sad music induces reward-related hormonal changes in empathic listeners. *Annals of the New York Academy of Sciences*, 1502(1), 121-131. <https://doi.org/10.1111/nyas.14660>
- E.B. (1949). A Dictionary of Musical Themes. *Music & Letters*, 30(3), 271-273. <https://www.jstor.org/stable/731016>
- Emmerson, S. (1986). The relation of language to materials. In S. Emmerson (Ed.), *The language of electroacoustic music* (pp. 17-39). Macmillan.
- Endres, D., Christensen, A., Omlor, L., & Giese, M.A. (2011). Emulating human observers with bayesian binning: Segmentation of action streams. *ACM Transactions on Applied Perception*, 8(3), Article 16, 1-12. <https://doi.org/10.1145/2010325.2010326>
- Erdem, Ç., Lan, Q., & Jensenius, A. R. (2020). Exploring relationships between effort, motion, and sound in new musical instruments. *Human Technology*, 16(3), 310. <https://ht.csr-pub.eu/index.php/ht/article/view/236>
- Escalera, S., Athitsos, V., & Guyon, I. (2016). Challenges in multimodal gesture recognition. *Journal of Machine Learning Research*, 17(72), 1-54. <https://www.jmlr.org/papers/v17/14-468.html>
- Fallman, D. (2003, April). Design-oriented human-computer interaction. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 225-232). <https://doi.org/10.1145/642611.642652>
- Fathy, Y., Barnaghi, P., & Tafazolli, R. (2019). An Online Adaptive Algorithm for Change Detection in Streaming Sensory Data. *IEEE Systems Journal*, 13(3), 2688-2699. <https://doi.org/10.1109/JSYST.2018.2876461>
- Fatone, G., Clayton, M., Leante, L., & Rahaim, M., (2011). Imagery, melody and gesture in cross-cultural perspective. In A. Gritten & E. King (Eds.), *New perspectives on music and gesture* (pp. 203-220). Ashgate.
- Fay, T. (1974). Context analysis of musical gestures. *Journal of Music Theory*, 18(1), 124-151. <https://www.jstor.org/stable/843138>
- Fiebrink, R., Trueman, D., & Cook, P. R. (2009). A meta-instrument for interactive, on-the-fly machine learning. In *Proceedings of the International Conference on New Interfaces for Musical Expression* (pp. 280-285). <https://doi.org/10.5281/zenodo.1177513>

- Foote, J. (2000). Automatic audio segmentation using a measure of audio novelty. In *International Conference on Multimedia and Expo* (Vol. 1, pp. 452-455). IEEE. <https://doi.org/10.1109/ICME.2000.869637>
- Fraisse, P. (1982). Rhythm and tempo. In D. Deutsch (Ed.), *The psychology of music* (pp. 149-180). Academic Press.
- Friberg, A., Schoonderwaldt, E., Hedblad, A., Fabiani, M., & Elowsson, A. (2014). Using listener-based perceptual features as intermediate representations in music information retrieval. *The Journal of the Acoustical Society of America*, 136(4), 1951-1963. <https://www.speech.kth.se/prod/publications/files/3952.pdf>
- Fricke, K. R., & Herzberg, P. Y. (2017). Personality and self-reported preference for music genres and attributes in a German-speaking sample. *Journal of Research in Personality*, 68, 114-123. <https://doi.org/10.1016/j.jrp.2017.01.001>
- Frid, E. (2019). Accessible digital musical instruments – a review of musical interfaces in inclusive music practice. *Multimodal Technologies and Interaction*, 3(3), 57. <https://doi.org/10.3390/mti3030057>
- Gabrielsson, A. (2002). Emotion perceived and emotion felt: Same or different?. *Musicae scientiae, Spec Issue 2001-2002*, 123-147. <https://doi.org/10.1177/10298649020050S105>
- Gabrielsson, A., & Juslin, P. N. (1996). Emotional expression in music performance: Between the performer's intention and the listener's experience. *Psychology of music*, 24(1), 68-91. <https://www.proquest.com/docview/1338859?pq-origsite=gscholar&fromopenview=true>
- Garrido, S., MacRitchie, J., Breaden, M., & Stevens, K. (2019). Why music moves us. *InPsych*, 41(1). <https://psychology.org.au/for-members/publications/inpsych/2019/february-issue-1/why-music-moves-us>
- Geeves, A., & Sutton, J. (2014). Embodied cognition, perception, and performance in music. *Empirical Musicology Review*, 9(3-4), 247-253. <https://doi.org/10.18061/emr.v9i3-4.4538>
- Gerra, G., Zaimovic, A., Franchini, D., Palladino, M., Giucastro, G., Reali, N., . . . Brambilla, F. (1998). Neuroendocrine responses of healthy volunteers to techno-music: Relationships with personality traits and emotional state. *International Journal of Psychophysiology*, 28(1), 99-111. [https://doi.org/10.1016/S0167-8760\(97\)00071-8](https://doi.org/10.1016/S0167-8760(97)00071-8)
- Gharghabi, S., Yeh, C.C.M., Ding, Y., Ding, W., Hibbing, P., LaMunion, S., Kaplan, A., Crouter, S.E., & Keogh, E. (2019). Domain agnostic online semantic segmentation for multi-dimensional time series. *Data Mining and Knowledge Discovery*, 33(1), 96-130. <https://doi.org/10.1007/s10618-018-0589-3>
- Gibb, B., Gibb, R., & Gibb, M. (1977). Stayin' alive [Song recorded by The Bee Gees]. On *Saturday Night Fever, The Original Motion Picture Soundtrack*. RSO.

- Gillian, N. (2011). *Gesture recognition for musician computer interaction* [Doctoral thesis, Queen's University Belfast].
<https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=dd94938b666b4345b1e94ca9c89de3734cf17822>
- Gillian, N., Knapp, B., & O'modhrain, S. (2011). Recognition Of Multivariate Temporal Musical Gestures Using N-Dimensional Dynamic Time Warping. In *Proceedings of the International Conference on New Interfaces for Musical Expression* (pp. 337-342). <https://doi.org/10.5281/zenodo.1178029>
- Gillian, N., & Paradiso, J. A. (2012). Digito: A Fine-Grain Gesturally Controlled Virtual Musical Instrument. In *Proceedings of the International Conference on New Interfaces for Musical Expression*.
<https://doi.org/10.5281/zenodo.1178263>
- Godøy, R. I. (2011). Coarticulated gestural-sonic objects in music. In *New perspectives on music and gesture* (pp. 67-82). Routledge.
- Godøy, R. I. (2014). Understanding coarticulation in musical experience. In M. Aramaki, O. Derrien, R. Kronland-Martinet, & S. Ystad (Eds.) *Sound, Music, and Motion* (pp. 535-547). Springer.
https://doi.org/10.1007/978-3-319-12976-1_32
- Goebel, W., Bresin, R., & Galembo, A. (2005). Touch and temporal behavior of grand piano actions. *The Journal of the Acoustical Society of America*, 118(2), 1154-1165. <https://www.speech.kth.se/prod/publications/files/1339.pdf>
- Gong, D., Medioni, G., & Zhao, X. (2014). Structured time series analysis for human action segmentation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7), 1414-1427.
<https://doi.org/10.1109/TPAMI.2013.244>
- Gurevich, M., & Treviño, J. (2007). Expression and its discontents: Toward an ecology of musical creation. In *Proceedings of the International Conference on New Interfaces for Musical Expression* (pp. 106-111).
<https://doi.org/10.5281/zenodo.1177107>
- Haga, E. (2008). *Correspondences between music and body movement* [Doctoral thesis, University of Oslo]. <http://urn.nb.no/URN:NBN:no-20848>
- Hard, B. M., Tversky, B., & Lang, D. S. (2006). Making sense of abstract events: Building event schemas. *Memory & cognition*, 34(6), 1221-1235.
<https://doi.org/10.3758/BF03193267>
- Hartmann, M., Lartillot, O., & Toiviainen, P. (2017). Interaction features for prediction of perceptual segmentation: Effects of musicianship and experimental task. *Journal of New Music Research*, 46(2), 156-174.
<http://urn.fi/URN:NBN:fi:ju-201705192420>
- Hodges, D. A. (2010). Psychophysiological measures. In P. N. Juslin & J. A. Sloboda (Eds.), *Handbook of music and emotion: Theory, research, applications* (pp. 279-311). Oxford University Press.
<https://doi.org/10.1093/acprof:oso/9780199230143.003.0011>
- Hodges, D. A., & Wilkins, R. W. (2015). How and why does music move us? Answers from psychology and neuroscience. *Music Educators Journal*, 101(4), 41-47. <https://doi.org/10.1177/0027432115575755>

- Hunter, P. G., Schellenberg, E. G., & Stalinski, S. M. (2011). Liking and identifying emotionally expressive music: Age and gender differences. *Journal of Experimental Child Psychology*, 110(1), 80–93.
<https://doi.org/10.1016/j.jecp.2011.04.001>
- Hurley, S. L. (2002). *Consciousness in action*. University Press.
- Jack, R. H., Mehrabi, A., Stockman, T., & McPherson, A. (2018). Action-sound latency and the perceived quality of digital musical instruments: Comparing professional percussionists and amateur musicians. *Music Perception*, 36(1), 109-128. <https://doi.org/10.1525/mp.2018.36.1.109>
- John, O. P., & Srivastava, S. (1999). The Big Five Trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (pp. 102–138). Guilford Press.
<https://www.ocf.berkeley.edu/~johnlab/pdfs/john&srivastava,1999.pdf>
- Jordà, S. (2002). FMOL: Toward user-friendly, sophisticated new musical instruments. *Computer Music Journal*, 26(3), 23-39.
<http://hdl.handle.net/10230/41827>
- Juslin, P. N. (2000). Cue utilization in communication of emotion in music performance: Relating performance to perception. *Journal of Experimental Psychology: Human Perception and Performance*, 26(6), 1797–1812.
https://psyk.uu.se/digitalAssets/510/510552_1juslin_emotion2000.pdf
- Juslin, P. N. (2008). Emotional Responses to Music. In S. Hallam, I. Cross, and M.H. Thaut (Eds), *Oxford Handbook of Music Psychology*.
<https://doi.org/10.1093/oxfordhb/9780199298457.013.0012>
- Juslin, P. N., & Laukka, P. (2004). Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. *Journal of new music research*, 33(3), 217-238.
<https://doi.org/10.1080/0929821042000317813>
- Juslin, P. N., & Lindström, E. (2010). Musical expression of emotions: Modelling listeners' judgements of composed and performed features. *Music Analysis*, 29(1-3), 334-364. <https://doi.org/10.1111/j.1468-2249.2011.00323.x>
- Kahol, K., Tripathi, P., & Panchanathan, S. (2004). Automated gesture segmentation from dance sequences. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings.* (pp. 883-888). IEEE. <https://doi.org/10.1109/AFGR.2004.1301645>
- Kaipainen, M., Ravaja, N., Tikka, P., Vuori, R., Pugliese, R., Rapino, M., & Takala, T. (2011). Enactive systems and enactive media: Embodied human-machine coupling beyond interfaces. *Leonardo*, 44(5), 433–438.
<https://muse.jhu.edu/article/449967>
- Kleinsmith, A., & Bianchi-Berthouze, N. (2012). Affective body expression perception and recognition: A survey. *IEEE Transactions on Affective Computing*, 4(1), 15-33. <https://doi.org/10.1109/T-AFFC.2012.16>
- Koelsch, S., Siebel, W.A., Fritz, T. (2010). Psychophysiological measures. In P. N. Juslin & J. A. Sloboda (Eds.), *Handbook of music and emotion: Theory, research, applications* (pp. 313–344). Oxford University Press.

- Kononova, A., Li, L., Kamp, K., Bowen, M., Rikard, R. V., Cotten, S., & Peng, W. (2019). The use of wearable activity trackers among older adults: focus group study of tracker perceptions, motivators, and barriers in the maintenance stage of behavior change. *JMIR Mhealth and Uhealth*, 7(4), e9832. <https://doi.org/10.2196/mhealth.9832>
- Krasnoskulov, A. (2019). *Family album: How does your daily activity sound?*. [Conference presentation]. ICAD Conference, Northumbria University. https://www.researchgate.net/profile/Alex-Krasnoskulov/publication/341071964_FAMILY_ALBUM_HOW_DOES_YOUR_DAILY_ACTIVITY_SOUND
- Krippendorff, K. (2011). Computing Krippendorff's alpha-reliability. https://repository.upenn.edu/asc_papers/43
- Krüger, V., Kragic, D., Ude, A., & Geib, C. (2007). The meaning of action: A review on action recognition and mapping. *Advanced robotics*, 21(13), 1473-1501. <https://doi.org/10.1163/156855307782148578>
- Krüger, B., Vögele, A., Willig, T., Yao, A., Klein, R., & Weber, A. (2017). Efficient unsupervised temporal segmentation of motion data. *IEEE Transactions on Multimedia*, 19(4), 797-812. <https://doi.org/10.1109/TMM.2016.2635030>
- Ladinig, O., & Schellenberg, E. G. (2012). Liking unfamiliar music: Effects of felt emotion and individual differences. *Psychology of Aesthetics, Creativity, and the Arts*, 6(2), 146-154. <https://www.utm.utoronto.ca/~w3psygs/LadinigSchellenberg2012.pdf>
- Lago, N. P., & Kon, F. (2004). The quest for low latency. In *International Computer Music Conference Proceedings*. <http://hdl.handle.net/2027/spo.bb2372.2004.142>
- Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. The University of Chicago press.
- Lakoff, G., and Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to western thought*. Basic Books.
- Lan, R., & Sun, H. (2015). Automated human motion segmentation via motion regularities. *The Visual Computer*, 31, 35-53. <https://doi.org/10.1007/s00371-013-0902-5>
- Lange, E. B., & Frieler, K. (2018). Challenges and opportunities of predicting musical emotions with perceptual and automatized features. *Music Perception*, 36(2), 217-242. <https://doi.org/10.1525/mp.2018.36.2.217>
- Langmeyer, A., Guglhör-Rudan, A., & Tarnai, C. (2012). What do music preferences reveal about personality? *Journal of Individual Differences*, 33(2), 119-130. https://www.researchgate.net/profile/Alexandra-Langmeyer/publication/254735373_What_Do_Music_Preferences_Reveal_About_Personality_A_Cross-Cultural_Replication_Using_Self-Ratings_and_Ratings_of_Music_Samples
- Lara, O. D., & Labrador, M. A. (2013). A survey on human activity recognition using wearable sensors. *IEEE communications surveys & tutorials*, 15(3), 1192-1209. <https://doi.org/10.1109/SURV.2012.110112.00192>

- Larsen, R. T., Wagner, V., Korfitsen, C. B., Keller, C., Juhl, C. B., Langberg, H., & Christensen, J. (2022). Effectiveness of physical activity monitors in adults: systematic review and meta-analysis. *BMJ*, 376.
<https://doi.org/10.1136/bmj-2021-068047>
- Last, M., & Usyskin, A. (2015). Listen to the sound of data. In A. K. Baughman, J. Gao, J-Y. Pan, & V.A. Petrushin (Eds.) *Multimedia Data Mining and Analytics: Disruptive Innovation* (pp. 419-446). https://www.researchgate.net/publication/282504359_Listen_to_the_Sound_of_Data
- Lee, V. (1881). Cherubino. *The Cornhill magazine*, 44(260), 218-232.
<https://books.google.com/books?id=O4dHAAAAYAAJ>
- Leichtentritt, H. (1924). German Music of the Last Decade. *The Musical Quarterly*, 10(2), 193-218. <https://www.jstor.org/stable/738268>
- Leman, M. (2008). *Embodied music cognition and mediation technology*. MIT Press.
- Leman, M., & Godøy, R. I. (2010). Why study musical gestures?. In R.I. Godøy & Leman, M. (Eds.), *Musical Gestures* (pp. 3-11). Routledge.
- Leman, M., & Maes, P. J. (2014a). The role of embodiment in the perception of music. *Empirical Musicology Review*, 9(3-4), 236-246.
<https://doi.org/10.18061/emr.v9i3-4.4498>
- Leman, M., & Maes, P.J. (2014b). Music perception and embodied music cognition. In L. Shapiro (Ed.), *The Routledge handbook of embodied cognition* (pp. 81-89). Routledge.
- Leman, M., Maes, P. J., Nijs, L., & Van Dyck, E. (2018). What is embodied music cognition?. *Springer handbook of systematic musicology*, 747-760.
https://www.researchgate.net/profile/Luc-Nijs/publication/319260279_What_Is_Embodied_Music_Cognition
- Lesaffre, M., Maes, P. J., & Leman, M. (Eds.). (2017). *The Routledge companion to embodied music interaction*. Routledge.
- Levitin, D. J., Chordia, P., & Menon, V. (2012). Musical rhythm spectra from Bach to Joplin obey a 1/f power law. *Proceedings of the National Academy of Sciences*, 109(10), 3716-3720. <https://doi.org/10.1073/pnas.1113828109>
- Ley-Flores, J., Turmo Vidal, L., Berthouze, N., Singh, A., Bevilacqua, F., & Tajadura-Jiménez, A. (2021). Soniband: Understanding the effects of metaphorical movement sonifications on body perception and physical activity. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Article 521, pp. 1-16).
<https://doi.org/10.1145/3411764.3445558>
- Liljeström, S., Juslin, P. N., & Västfjäll, D. (2012). Experimental evidence of the roles of music choice, social context, and listener personality in emotional reactions to music. *Psychology of Music*, 41(5), 579-599.
<https://doi.org/10.1177/0305735612440615>
- Lin, J.F.S., Karg, M., & Kulić, D. (2016). Movement primitive segmentation for human motion modeling: A framework for analysis. *IEEE Transactions on Human-Machine Systems* 46(3), 325-339.
<https://doi.org/10.1109/THMS.2015.2493536>

- Liu, J. Y. W., Kor, P. P. K., Chan, C. P. Y., Kwan, R. Y. C., & Cheung, D. S. K. (2020). The effectiveness of a wearable activity tracker (WAT)-based intervention to improve physical activity levels in sedentary older adults: A systematic review and meta-analysis. *Archives of Gerontology and Geriatrics*, 91, 104211. <https://doi.org/10.1016/j.archger.2020.104211>
- Liu, S., Yamada, M., Collier, N., & Sugiyama, M. (2013). Change-point detection in time-series data by relative density-ratio estimation. *Neural Networks*, 43, 72-83. <https://doi.org/10.1016/j.neunet.2013.01.012>
- Luck, G., Saarikallio, S., Burger, B., Thompson, M., & Toiviainen, P. (2010). Effects of the Big Five and musical genre on music-induced movement. *Journal of Research in Personality*, 44(6), 714-720. <https://doi.org/10.1016/j.jrp.2010.10.001>
- Luck, G., Saarikallio, S., Burger, B., Thompson, M., & Toiviainen, P. (2014). Emotion-driven encoding of music preference and personality in dance. *Musicae Scientiae*, 18(3), 307-323. <https://doi.org/10.1177/1029864914537290>
- Lv, F., & Nevatia, R. (2006). Recognition and segmentation of 3-D human action using HMM and Multi-Class Adaboost. In A. Leonardis, H. Bischof, & A. Pinz (Eds.) *Computer Vision-ECCV* (pp. 359-372). Springer. https://doi.org/10.1007/11744085_28
- MacRitchie, J., Buck, B., & Bailey, N. J. (2013). Inferring musical structure through bodily gestures. *Musicae Scientiae*, 17(1), 86-108. <https://doi.org/10.1177/1029864912467632>
- Margulis, E. H. (2007). Moved by nothing: listening to musical silence. *Journal of Music Theory*, 51(2), 245-276. <https://doi.org/10.1215/00222909-2009-003>
- Markou, M., & Singh, S. (2003). Novelty detection: a review – part 1: statistical approaches. *Signal processing*, 83(12), 2481-2497. <https://doi.org/10.1016/j.sigpro.2003.07.018>
- Martin, C. P., Glette, K., Nygaard, T. F., & Torresen, J. (2020). Understanding musical predictions with an embodied interface for musical machine learning. *Frontiers in Artificial Intelligence*, 3, 6. <https://doi.org/10.3389/frai.2020.00006>
- Matyja, J. R. (2016). Embodied music cognition: Trouble ahead, trouble behind. *Frontiers in psychology*, 7, 1891. <https://doi.org/10.3389/fpsyg.2016.01891>
- Matyja, J. R., & Schiavio, A. (2013). Enactive music cognition: Background and research themes. *Constructivist Foundations*, 8(3), 351-357. https://www.researchgate.net/publication/272169394_Enactive_music_cognition
- Mazzola, G. (2012). Singular homology on hypergestures. *Journal of Mathematics and Music*, 6(1), 49-60. <https://doi.org/10.1080/17459737.2012.680311>
- Mazzola, G., Mannone, M., Pang, Y., O'Brien, M., & Torunsky, N. (2016). *All About Music – The Complete Ontology: Realities, Semiotics, Communication, and Embodiment*. Springer. <https://doi.org/10.1007/978-3-319-47334-5>
- McCrae, R. R. (2007). Aesthetic chills as a universal marker of openness to experience. *Motivation and Emotion*, 31(1), 5-11. <https://doi.org/10.1007/s11031-007-9053-1>

- McCrae, R. R., & Costa P. T., Jr. (2004). A contemplated revision of the NEO Five-Factor Inventory. *Personality and Individual Differences*, 36(3), 587–596. [https://doi.org/10.1016/S0191-8869\(03\)00118-1](https://doi.org/10.1016/S0191-8869(03)00118-1)
- McPherson, A. P., Jack, R. H., & Moro, G. (2016). Action-sound latency: Are our tools fast enough?. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. <https://doi.org/10.5281/zenodo.3964611>
- Meier, F., Theodorou, E., Stulp, F., & Schaal, S. (2011). Movement segmentation using a primitive library. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 3407-3412). IEEE. <https://doi.org/10.1109/IROS.2011.6094676>
- Melchers, M. C., Li, M., Haas, B. W., Reuter, M., Bischoff, L., & Montag, C. (2016). Similar personality patterns are associated with empathy in four different countries. *Frontiers in Psychology*, 7, Article 290. <https://doi.org/10.3389/fpsyg.2016.00290>
- Mendoza, J.I. (2014). *Self-report measurement of segmentation, mimesis and perceived emotions in acousmatic electroacoustic music* [Master's thesis, University of Jyväskylä]. <http://urn.fi/URN:NBN:fi:jyu-201406192112>
- Merriam-Webster. (n.d.) Gesture. In *Merriam-Webster.com dictionary*. <https://www.merriam-webster.com/dictionary/gesture>
- Merrill, D. J., & Paradiso, J. A. (2005, April). Personalization, expressivity, and learnability of an implicit mapping strategy for physical interfaces. In *Proceedings of the CHI Conference on Human Factors in Computing Systems, Extended Abstracts* (pp. 2152-2161). https://resenv.media.mit.edu/pubs/papers/2005-04-alt_chi-flexigesture.pdf
- Michie, S., Richardson, M., Johnston, M., Abraham, C., Francis, J., Hardeman, W., ... & Wood, C. E. (2013). The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. *Annals of behavioral medicine*, 46(1), 81-95. <https://doi.org/10.1007/s12160-013-9486-6>
- Mitra, S., & Acharya, T. (2007). Gesture recognition: A survey. In *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 37(3), 311-324. <https://doi.org/10.1109/TSMCC.2007.893280>
- Moore, F. R. (1988). The dysfunctions of MIDI. *Computer music journal*, 12(1), 19-28. <https://doi.org/10.2307/3679834>
- Morand, F. (2019) *In the intersection of emotion, biosensors and sound: emerging corporealities in Emovere's interaction*. <https://archee.uqam.ca/decembre-2019-in-the-intersection-of-emotion-biosensors-and-sound-emerging-corporealities-in-emoveres-interaction/>
- Moro, G., & McPherson, A. (2020). A platform for low-latency continuous keyboard sensing and sound generation. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. <https://doi.org/10.5281/zenodo.4813253>

- Murad, D., Ye, F., Barone, M., & Wang, Y. (2017). Motion initiated music ensemble with sensors for motor rehabilitation. In *2017 international conference on orange technologies* (pp. 87-90). IEEE.
<https://doi.org/10.1109/ICOT.2017.8336095>
- Murray-Browne, T., & Tigas, P. (2021). Emergent interfaces: Vague, complex, bespoke and embodied interaction between humans and computers. *Applied Sciences*, 11(18), 8531. <https://doi.org/10.3390/app11188531>
- Music in Relation to Other Arts. (1910). Music in Relation to Other Arts by H. Walford Davies. *The Musical Times*, 51(806), 236-238.
<https://www.jstor.org/stable/906827>
- Müller, M., Röder, T., & Clausen, M. (2005). Efficient content-based retrieval of motion capture data. In *ACM SIGGRAPH 2005 Papers* (pp. 677-685).
<https://doi.org/10.1145/1186822.1073247>
- Nanchen, D., Leening, M. J., Locatelli, I., Cornuz, J., Kors, J. A., Heeringa, J., ... & Dehghan, A. (2013). Resting heart rate and the risk of heart failure in healthy adults: the Rotterdam Study. *Circulation: Heart Failure*, 6(3), 403-410. <https://doi.org/10.1161/CIRCHEARTFAILURE.112.000171>
- Naser, D. S., & Saha, G. (2021). Influence of music liking on EEG based emotion recognition. *Biomedical Signal Processing and Control*, 64, 102251.
<https://doi.org/10.1016/j.bspc.2020.102251>
- Nave, G., Minxha, J., Greenberg, D. M., Kosinski, M., Stillwell, D., & Rentfrow, J. (2018). Musical preferences predict personality: Evidence from active listening and Facebook likes. *Psychological Science*, 29(7), 1145-1158.
<https://doi.org/10.1177/0956797618761659>
- Newton, D. (1973). Attribution and the unit of perception of ongoing behavior. *Journal of personality and social psychology*, 28(1), 28-38.
<https://doi.org/10.1037/h0035584>
- Ni, Q., Garcia Hernando, A. B., & De la Cruz, I. P. (2015). The elderly's independent living in smart homes: A characterization of activities and sensing infrastructure survey to facilitate services development. *Sensors*, 15(5), 11312-11362. <https://doi.org/10.3390/s150511312>
- NIME Proceedings Archive (n.d.) *Proceedings of the International Conference on New Interfaces for Musical Expression*. <https://www.nime.org/archives/>
- Noor, M. H. M., Salci, Z., Kevin, I., & Wang, K. (2017). Adaptive sliding window segmentation for physical activity recognition using a single tri-axial accelerometer. *Pervasive and Mobile Computing*, 38, 41-59.
<https://doi.org/10.1016/j.pmcj.2016.09.009>
- Noroozi, F., Corneanu, C. A., Kamińska, D., Sapiński, T., Escalera, S., & Anbarjafari, G. (2018). Survey on emotional body gesture recognition. *IEEE transactions on affective computing*, 12(2), 505-523.
<https://doi.org/10.1109/TAFFC.2018.2874986>
- Novello, Ewer and Co. (1894). Just published. *The Musical Times and Singing Class Circular*, 35(612), 128. <https://www.jstor.org/stable/3361140>

- Nusbaum, E. C., & Silvia, P. J. (2011). Shivers and timbres: Personality and the experience of chills from music. *Social Psychological and Personality Science*, 2(2), 199–204. <https://doi.org/10.1177/1948550610386810>
- O’Keeffe, N., Scheid, J. L., & West, S. L. (2020). Sedentary behavior and the use of wearable technology: An editorial. *International Journal of Environmental Research and Public Health*, 17(12), 4181. <https://doi.org/10.3390/ijerph17124181>
- Oriolo, E. (2019). *Il suono dei gesti: tecnologie “incorporate” nella performance vocale della popular music* [Master’s thesis]. https://www.lim.di.unimi.it/download/theses/921770_oriolo/tesi.pdf
- Otondo, F. (2008). Ciguri [Electroacoustic piece]. On *Tutuguri*. Sargasso.
- Park, M., Hennig-Fast, K., Bao, Y., Carl, P., Pöppel, E., Welker, L., . . . Gutyrchik, E. (2013). Personality traits modulate neural responses to emotions expressed in music. *Brain Research*, 1523, 68–76. <https://doi.org/10.1016/j.brainres.2013.05.042>
- Parker, F. W. (1894). *Talks on pedagogics: An outline of the theory of concentration*. E.L. Kellogg & Co. <https://books.google.com/books?id=LbwKAAAIAAJ>
- Patrona, F., Chatzitofis, A., Zarpalas, D., & Daras, P. (2018). Motion analysis: Action detection, recognition and evaluation based on motion capture data. *Pattern Recognition*, 76, 612–622. <https://doi.org/10.1016/j.patcog.2017.12.007>
- Patterson, T., Khan, N., McClean, S., Nugent, C., Zhang, S., Cleland, I., & Ni, Q. (2017). Sensor-based change detection for timely solicitation of user engagement. *IEEE Transactions on Mobile Computing*, 16(10), 2889–2900. <https://doi.org/10.1109/TMC.2016.2640959>
- Peretz, I., Gagnon, L., & Bouchard, B. (1998). Music and emotion: perceptual determinants, immediacy, and isolation after brain damage. *Cognition*, 68(2), 111–141. [https://doi.org/10.1016/S0010-0277\(98\)00043-2](https://doi.org/10.1016/S0010-0277(98)00043-2)
- Petzdold, C. (ca. 1725). Minuet in G major. *The Anna Magdalena Bach Notebook*, Anh. 114.
- Portegijs, E., Karavirta, L., Saajanaho, M., Rantalainen, T., & Rantanen, T. (2019). Assessing physical performance and physical activity in large population-based aging studies: home-based assessments or visits to the research center?. *BMC Public Health*, 19, 1570. <https://doi.org/10.1186/s12889-019-7869-8>
- Rantalainen, T., Koivunen, K., Portegijs, E., Rantanen, T., Palmberg, L., Karavirta, L., & Chastin, S. (2022). Is Complexity of Daily Activity Associated with Physical Function and Life-Space Mobility among Older Adults?. *Medicine and science in sports and exercise*, 54(7), 1210. <https://doi.org/10.1249/mss.0000000000002883>
- Rantanen, T., Saajanaho, M., Karavirta, L., Siltanen, S., Rantakokko, M., Viljanen, A., ... & Portegijs, E. (2018). Active aging–resilience and external support as modifiers of the disablement outcome: AGNES cohort study protocol. *BMC Public Health*, 18, 565. <https://doi.org/10.1186/s12889-018-5487-5>

- Rasch, R. A. (1981). *Aspects of the perception and performance of polyphonic music* [Doctoral dissertation, University of Groningen].
<https://core.ac.uk/download/pdf/232460432.pdf>
- Redmond, S. J., Scalzi, M. E., Narayanan, M. R., Lord, S. R., Cerutti, S., & Lovell, N. H. (2010, August). Automatic segmentation of triaxial accelerometry signals for falls risk estimation. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology* (pp. 2234-2237). IEEE.
<https://doi.org/10.1109/IEMBS.2010.5627384>
- Reić Ercegovic, I. R., Dobrota, S., & Kuščević, D. (2015). Relationship between music and visual art preferences and some personality traits. *Empirical Studies of the Arts*, 33(2), 207-227.
<https://doi.org/10.1177/0276237415597390>
- Reisenzein, R., & Weber, H. (2009). Personality and emotion. In P. J. Corr & G. Matthews (Eds.), *The Cambridge handbook of personality psychology* (pp. 54-71). Cambridge University Press.
- Rentfrow, P. J., & Gosling, S. D. (2003). The do re mi's of everyday life: The structure and personality correlates of music preferences. *Journal of Personality and Social Psychology*, 84(6), 1236-1256.
<https://osf.io/jmbp7/download>
- Rothman, K. J. (1990). No adjustments are needed for multiple comparisons. *Epidemiology*, 1(1), 43-46. <https://www.jstor.org/stable/20065622>
- Rubine, D., & McAvinney, P. (1990). Programmable finger-tracking instrument controllers. *Computer music journal*, 14(1), 26-41.
<https://www.jstor.org/stable/3680114>
- Saganowski, S., Dutkowiak, A., Dziadek, A., Dzieżyc, M., Komoszyńska, J., Michalska, W., ... & Kazienko, P. (2020). Emotion recognition using wearables: A systematic literature review-work-in-progress. In *2020 IEEE International Conference on Pervasive Computing and Communications Workshops* (pp. 1-6). IEEE.
<https://doi.org/10.1109/PerComWorkshops48775.2020.9156096>
- Salamah, S., Zhang, L., & Brunnett, G. (2015). Hierarchical method for segmentation by classification of motion capture data. In G. Brunnett, S. Coquillart, R. van Liere, G. Welch, & L. Váša (Eds.) *Virtual Realities*, 8844 (pp. 169-186). Springer. https://doi.org/10.1007/978-3-319-17043-5_10
- Santos, L., Khoshhal, K., & Dias, J. (2015). Trajectory-based human action segmentation. *Pattern Recognition*, 48(2), 568-579.
<https://doi.org/10.1016/j.patcog.2014.08.015>
- Sapiński, T., Kamińska, D., Pelikant, A., & Anbarjafari, G. (2019). Emotion recognition from skeletal movements. *Entropy*, 21(7), 646.
<https://doi.org/10.3390/e21070646>
- Sessions, R. (2015). *Musical experience of composer, performer, listener*. Princeton University Press.
- Schaffert, N., Janzen, T. B., Mattes, K., & Thaut, M. H. (2019). A review on the relationship between sound and movement in sports and rehabilitation. *Frontiers in psychology*, 10, 244. <https://doi.org/10.3389/fpsyg.2019.00244>

- Schäfer, T., & Mehlhorn, C. (2017). Can personality traits predict musical style preferences? A meta-analysis. *Personality and Individual Differences*, 116, 265–273. <https://doi.org/10.1016/j.paid.2017.04.061>
- Schäfer, T., & Sedlmeier, P. (2011). Does the body move the soul? The impact of arousal on music preference. *Music Perception*, 29(1), 37–50. <https://doi.org/10.1525/mp.2011.29.1.37>
- Schätti, G. (2007). Real-Time Audio Feature Analysis for Decklight3. <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.85.7916&rep=%20rep1&type=pdf>
- Schedl, M., Eghbal-Zadeh, H., Gómez Gutiérrez, E., & Tkalčič, M. (2016). An analysis of agreement in classical music perception and its relationship to listener characteristics. In J. Devaney, M.I. Mandel, D. Turnbull, and G. Tzanetakis (Eds.), *Proceedings of the 17th International Society for Music Information Retrieval Conference*, (pp. 578-583). <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=eb1ca5a6f9ed9ddecdd85f117ccb8019be3cca>
- Schiavio, A., & Menin, D. (2013). Embodied music cognition and mediation technology: A critical review. *Psychology of music*, 41(6), 804-814. <https://doi.org/10.1177/0305735613497169>
- Schneider, A. (2010). Music and gestures: A historical introduction and survey of earlier research. In R.I. Godøy & Leman, M. (Eds.), *Musical Gestures* (pp. 81-112). Routledge.
- Schoen-Nazzaro, M.B. (1978) Plato and Aristotle on the Ends of Music. *Laval théologique et philosophique* 34(3), 261-273. <https://doi.org/10.7202/705684ar>
- Schrock, K. (2009). Why music moves us. *Scientific American Mind*, 20(4), 32-37. <http://www.jstor.org/stable/24940149>
- Schrödinger, E. (1944). *What is Life? The Physical Aspect of the Living Cell*. Cambridge University Press.
- Shapiro, L. (2011). *Embodied cognition*. Routledge.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2), 420. <http://www.rokwa.x-y.net/Shrout-Fleiss-ICC.pdf>
- Schubart, M. A. (1946). Roger Sessions: Portrait of an American Composer. *The Musical Quarterly*, 32(2), 196-214. <https://www.jstor.org/stable/739131>
- Schubert, E. (2010). Continuous self-report methods. In P. N. Juslin & J. A. Sloboda (Eds.), *Handbook of music and emotion: Theory, research, applications* (pp. 223–253). Oxford University Press.
- Sievers, B., Polansky, L., Casey, M., & Wheatley, T. (2013). Music and movement share a dynamic structure that supports universal expressions of emotion. *Proceedings of the National Academy of Sciences*, 110(1), 70-75. <https://doi.org/10.1073/pnas.1209023110>

- Silvia, P. J., Fayn, K., Nusbaum, E. C., & Beaty, R. E. (2015). Openness to experience and awe in response to nature and music: Personality and profound aesthetic experiences. *Psychology of Aesthetics, Creativity, and the Arts*, 9(4), 376-384. https://www.researchgate.net/profile/Roger-Beaty/publication/280112363_Openness_to_Experience_and_Awe_in_Response_to_Nature_and_Music_Personality_and_Profound_Aesthetic_Experiences
- Simpson, W. S., Peake, G., & Hammond, T. J. (1894). The Jubilee of "The Musical Times". *The Musical Times and Singing Class Circular*, 35(616), 416-418. <https://www.jstor.org/stable/3362611>
- Smalley, D. (1986). Spectro-morphology and structuring processes. In S. Emmerson (Ed.), *The language of electroacoustic music* (pp. 61-93). Macmillan.
- Smit, E. A., Milne, A. J., Sarvasy, H. S., & Dean, R. T. (2022). Emotional responses in Papua New Guinea show negligible evidence for a universal effect of major versus minor music. *Plos one*, 17(6), e0269597. <https://doi.org/10.1371/journal.pone.0269597>
- Solberg, R.T. (2015). 'Moved by the music': Affective arousal, body movement and musical features of electronic dance music. In J. Ginsborg, A. Lamont, & S. Bramley (Eds.) *Proceedings of the Ninth Triennial Conference of the European Society for the Cognitive Sciences of Music* (pp. 744-745). https://www.escomsociety.org/_files/ugd/772b99_a17cfacc8bb64348a82cfff460d111a.pdf
- Staudt, P., Sarigöl, E., Lussana, M., Rizzonelli, M., Hyun Kim, J. (2022) Automatic Classification of Interactive Gestures for Inter-Body Proximity Sonification. In S. Pauletto, S. Delle Monache, and R. Selfridge (Eds) *Proceedings of the Conference on Sonification of Health and Environmental Data* (pp. 20-27). <https://doi.org/10.5281/zenodo.7243901>
- Styns, F., van Noorden, L., Moelants, D., & Leman, M. (2007). Walking on music. *Human movement science*, 26(5), 769-785. <https://doi.org/10.1016/j.humov.2007.07.007>
- Swarbrick, D., Bosnyak, D., Livingstone, S. R., Bansal, J., Marsh-Rollo, S., Woolhouse, M. H., & Trainor, L. J. (2019). How live music moves us: head movement differences in audiences to live versus recorded music. *Frontiers in psychology*, 9, 2682. <https://doi.org/10.3389/fpsyg.2018.02682>
- Tahiroglu, K., Correia, N. N., & Espada, M. (2013). PESI Extended System: In Space, On Body, with 3 Musicians. In *Proceedings of the International Conference on New Interfaces for Musical Expression* (pp. 35-40). <https://doi.org/10.5281/zenodo.1178666>
- Tanaka, A. (2019). Embodied musical interaction: body physiology, cross modality, and sonic experience. In S. Holland, T. Mudd, K. Wilkie-McKenna, A. McPherson, and M. Wanderley (Eds.) *New directions in music and human-computer interaction* (pp. 135-154). <https://research.gold.ac.uk/id/eprint/26000/>
- Ter Bogt, T. F., Mulder, J., Raaijmakers, Q. A., & Nic Gabhainn, S. (2011). Moved by music: A typology of music listeners. *Psychology of music*, 39(2), 147-163. <https://psycnet.apa.org/doi/10.1177/0305735610370223>

- Terracciano, A., McCrae, R. R., Hagemann, D., & Costa, P. T., Jr. (2003). Individual difference variables, affective differentiation, and the structures of affect. *Journal of Personality*, 71(5), 669–704. <https://doi.org/10.1111/1467-6494.7105001>
- Terry, P. C., Lane, A. M., & Fogarty, G. J. (2003). Construct validity of the profile of mood states – Adolescents for use with adults. *Psychology of Sport and Exercise*, 4(2), 125–139. [https://doi.org/10.1016/S1469-0292\(01\)00035-8](https://doi.org/10.1016/S1469-0292(01)00035-8)
- Tetarto Hood (2014). Bouzouki Hiphop – Rempetila [Instrumental].
- Toiviainen, P., Luck, G., & Thompson, M. R. (2010). Embodied meter: Hierarchical eigenmodes in music-induced movement. *Music Perception*, 28(1), 59–70. <https://doi.org/10.1525/mp.2010.28.1.59>
- Trolland, S., Ilsar, A., Frame, C., McCormack, J., & Wilson, E. (2022). AirSticks 2.0: Instrument Design for Expressive Gestural Interaction. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. <https://doi.org/10.21428/92fbeb44.c400bdc2>
- Upadhyay, D., Shukla, R., & Chakraborty, A. (2017). Factor structure of music preference scale and its relation to personality. *Journal of Indian Academy of Applied Psychology*, 43(1), 104–113. <https://ssrn.com/abstract=2882418>
- Upadhyay, D. K., Shukla, R., Tripathi, V. N., & Agrawal, M. (2017). Exploring the nature of music engagement and its relation to personality among young adults. *International Journal of Adolescence and Youth*, 22(4), 484–496. <https://doi.org/10.1080/02673843.2016.1245150>
- Van Dyck, E., Maes, P. J., Hargreaves, J., Lesaffre, M., & Leman, M. (2013). Expressing induced emotions through free dance movement. *Journal of Nonverbal Behavior*, 37(3), 175–190. <https://doi.org/10.1007/s10919-013-0153-1>
- Van Nort, D. (2011). Human: Machine: Human: Gesture, sound and embodiment. *Kybernetes*, 40(7/8), 1179–1188. <https://doi.org/10.1108/03684921111160403>
- Van Zijl, A. G., & Luck, G. (2013). Moved through music: The effect of experienced emotions on performers' movement characteristics. *Psychology of Music*, 41(2), 175–197. <https://doi.org/10.1177/0305735612458334>
- Varela, F. J., Thompson, E., and Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. MIT Press.
- Vähä-Ypyä, H., Vasankari, T., Husu, P., Mänttari, A., Vuorimaa, T., Suni, J., & Sievänen, H. (2015). Validation of cut-points for evaluating the intensity of physical activity with accelerometry-based mean amplitude deviation (MAD). *PloS one*, 10(8), e0134813. <https://doi.org/10.1371/journal.pone.0134813>
- Västfjäll, D. (2010). Indirect perceptual, cognitive, and behavioural measures. In P. N. Juslin & J. A. Sloboda (Eds.), *Handbook of music and emotion: Theory, research, applications* (pp. 255–277). Oxford University Press.

- Vickers, P., & Höldrich, R. (2019) Direct Segmented Sonification of Characteristic Features of the Data Domain. In *The 25th International Conference on Auditory Display*. <http://dx.doi.org/10.21785/icad2019.043>
- Vieillard, S., Peretz, I., Gosselin, N., Khalfa, S., Gagnon, L., & Bouchard, B. (2008). Happy, sad, scary and peaceful musical excerpts for research on emotions. *Cognition & Emotion*, 22(4), 720-752. <https://doi.org/10.1080/02699930701503567>
- Visi, F. (2017). *Methods and technologies for the analysis and interactive use of body movements in instrumental music performance* [Doctoral dissertation, University of Plymouth]. <http://hdl.handle.net/10026.1/8805>
- Vuoskoski, J. K., & Eerola, T. (2011a). Measuring music-induced emotion: A comparison of emotion models, personality biases, and intensity of experiences. *Musicae Scientiae*, 15(2), 159-173. <https://doi.org/10.1177/1029864911403367>
- Vuoskoski, J. K., & Eerola, T. (2011b). The role of mood and personality in the perception of emotions represented by music. *Cortex*, 47(9), 1099-1106. <https://doi.org/10.1016/j.cortex.2011.04.011>
- Vuoskoski, J. K., & Eerola, T. (2017). The pleasure evoked by sad music is mediated by feelings of being moved. *Frontiers in psychology*, 439. <https://doi.org/10.3389/fpsyg.2017.00439>
- Vuoskoski, J. K., Gatti, E., Spence, C., & Clarke, E. F. (2016). Do visual cues intensify the emotional responses evoked by musical performance? A psychophysiological investigation. *Psychomusicology: Music, mind, and brain*, 26(2), 179-188. <https://ora.ox.ac.uk/objects/uuid:636d396d-a540-4eee-9106-49976b654549>
- Vuoskoski, J. K., Thompson, M. R., Clarke, E. F., & Spence, C. (2014). Crossmodal interactions in the perception of expressivity in musical performance. *Attention, Perception, & Psychophysics*, 76, 591-604. <https://doi.org/10.3758/s13414-013-0582-2>
- Vuoskoski, J. K., Thompson, W. F., McIlwain, D., & Eerola, T. (2012). Who enjoys listening to sad music and why? *Music Perception*, 29(3), 311-317. <https://doi.org/10.1525/mp.2012.29.3.311>
- Vuoskoski, J. K., Thompson, M. R., Spence, C., & Clarke, E. F. (2016). Interaction of sight and sound in the perception and experience of musical performance. *Music Perception*, 33(4), 457-471. <https://doi.org/10.1525/mp.2016.33.4.457>
- Vuoskoski, J. K., Zickfeld, J. H., Alluri, V., Moorthigari, V., & Seibt, B. (2022). Feeling moved by music: Investigating continuous ratings and acoustic correlates. *Plos one*, 17(1), e0261151. <https://doi.org/10.1371/journal.pone.0261151>
- Wakabayashi, A., Baron-Cohen, S., Wheelwright, S., Goldenfeld, N., Delaney, J., Fine, D., . . . Weil, L. (2006). Development of short forms of the Empathy Quotient (EQ-Short) and the Systemizing Quotient (SQ-Short). *Personality and Individual Differences*, 41(5), 929-940. <https://doi.org/10.1016/j.paid.2006.03.017>

- Waldbauer, I. (1960). Bartók's 'Four Pieces' for Two Pianos. *Tempo*, (53-54), 17-22. <https://www.jstor.org/stable/942477>
- Wanderley, M. M., Vines, B. W., Middleton, N., McKay, C., & Hatch, W. (2005). The musical significance of clarinetists' ancillary gestures: An exploration of the field. *Journal of New Music Research*, 34(1), 97-113. <https://doi.org/10.1080/09298210500124208>
- Wang, J. Y. (2021). *Analysis of Wireless Interface Latency and Usability for Digital Musical Instruments*. [Doctoral thesis, McGill University]. <https://escholarship.mcgill.ca/concern/theses/0r967843x>
- Wang, W., Cheng, J., Song, W., & Shen, Y. (2022). The effectiveness of wearable devices as physical activity interventions for preventing and treating obesity in children and adolescents: Systematic review and meta-analysis. *JMIR Mhealth and Uhealth*, 10(4), e32435. <https://doi.org/10.2196/32435>
- Weber, F. J. (1891). *Popular History of Music from the Earliest Times*. Simpkin, Marshall, Hamilton, Kent & Co., and Aug. Siegle. <https://books.google.fi/books?hl=en&lr=&id=WOUqAAAYAAJ&oi=fnd&pg=PA3&dq=Popular+History+of+Music+from+the+Earliest+Times>
- Wessel, D., & Wright, M. (2002). Problems and prospects for intimate musical control of computers. *Computer music journal*, 26(3), 11-22. <https://www.jstor.org/stable/3681975>
- Wilkinson, W. C. (1869). *The Dance of Modern Society*. Oakley, Mason & Co. <https://books.google.com/books?id=Zk9KAAAYAAJ>
- Wishart, T. (1996). *On sonic art* (Rev. ed.). Routledge.
- Wöllner, C., & Deconinck, F. (2012). Movement expertise influences gender recognition in point-light displays of musical gestures. In *12th International conference on Music Perception and Cognition; 8th Triennial conference of the European Society for the Cognitive Sciences of Music*. http://icmpec-escom2012.web.auth.gr/files/papers/1136_Proc.pdf
- Wurm, M. (1896). About Teaching Children the Pianoforte. *The Magazine of music*, 13(4), 260-261. https://archive.org/details/sim_magazine-of-music_1896-04_13_4
- Zacks, J. M., Kumar, S., Abrams, R. A., & Mehta, R. (2009). Using movement and intentions to understand human activity. *Cognition*, 112(2), 201-216. <https://doi.org/10.1016/j.cognition.2009.03.007>
- Zacks, J. M., Tversky, B., & Iyer, G. (2001). Perceiving, remembering, and communicating structure in events. *Journal of experimental psychology: General*, 130(1), 29. <https://www.tc.columbia.edu/faculty/bt2158/faculty-profile/files/grememberingandcommunicatingstructureinevents.pdf>
- Zamani, M., Sadri, A., Ghafoori, Z., Moshtaghi, M., Salim, F. D., Leckie, C., & Ramamohanarao, K. (2020). Unsupervised online change point detection in high-dimensional time series. *Knowledge and Information Systems*, 62(2), 719-750. <https://doi.org/10.1007/s10115-019-01366-x>
- Zeiner-Henriksen, H. T. (2016). Moved by the groove: Bass drum sounds and body movements in electronic dance music. In A. Danielsen (Ed.), *Musical rhythm in the age of digital reproduction* (pp. 121-140). Routledge.

- Zelechowska, A., Gonzalez Sanchez, V. E., Laeng, B., Vuoskoski, J. K., & Jensenius, A. R. (2020). Who moves to music? empathic concern predicts spontaneous movement responses to rhythm and music. *Music & Science*, 3, 2059204320974216. <https://doi.org/10.1177/2059204320974216>
- Zentner, M., & Eerola, T. (2010). Self-report measures and models. In P. N. Juslin & J. A. Sloboda (Eds.), *Handbook of music and emotion: Theory, research, applications* (pp. 187–221). Oxford University Press.
- Zentner, M., Grandjean, D., & Scherer, K. R. (2008). Emotions evoked by the sound of music: characterization, classification, and measurement. *Emotion*, 8(4), 494. <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=f60b550a2918bd4fb83882ee4784c4a7ff4df826>
- Zhou, F., De la Torre, F., & Hodgins, J. K. (2013). Hierarchical aligned cluster analysis for temporal clustering of human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3), 582-596. <https://doi.org/10.1109/TPAMI.2012.137>
- Zweigenhaft, R. L. (2008). A do re mi encore: A closer look at the personality correlates of music preferences. *Journal of Individual Differences*, 29(1), 45–55. https://d1wqtxts1xzle7.cloudfront.net/83198635/80804a3e65b3cef8a328391771d36edb307f-libre.pdf?1649084727=&response-content-disposition=inline%3B+filename%3DA_Do_Re_Mi_Encore_A_Closer_Look_at_the_P.pdf&Expires=1678973342&Signature=Mbu6ltWZ~kjvdQkZFdJ2dNzRs6WcEW8XofsSAzV5MeCfT06SWDE~Z4-JAZSt7cpGotNBH7Oh9Mvy2pdd9NMXG6sANIfpgp9IkQJBeiCMFDg9~8yNwir9ZpW~EjJDP~RstFWo1mUKuGudS4N2bZn7XqqcUEzFTgvgw2s~0xeMLr4WkNecEWbrh18ev5~5GkntbyvQjZbpUp640Ce6yuzeSyTISv8Uoe2EfxP4fomCRA~Ymc0ZFthk7aK1P7wE~8PifZfjTf35S9x2xKQEaxGiZIM4uhB00PIQIV0q5TcL0hh6aXifEIoeSVv-OJ9w79odwtgaOE7JaPMNw9B3-nvEg__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA



ORIGINAL PAPERS

I

GESTURAL AGENCY IN HUMAN-MACHINE MUSICAL INTERACTION

by

Juan Ignacio Mendoza & Marc Richard Thompson, 2017

The Routledge Companion to Embodied Music Interaction, pp. 412-419.

<https://doi.org/10.4324/9781315621364-45>

Gestural Agency in Human-Machine Musical Interaction

Juan Ignacio Mendoza and Marc Richard Thompson

This is the retyped authors' manuscript.

The abstract and some typographic details are different from the published version.

Abstract

Musical technologies are evolving in such a way that they start to resemble the people that use them. In this vision it seems pertinent to abandon the conception of musicians as users of musical instruments in favor of machines and humans interacting to make music. Both the machine and the human can be modeled as embodied cognitive agents that comprise a network of musical gestures. These gestures are multi-modal signals that allow the agents to exert influence upon each other towards making music. This standpoint integrates traditional and novel music-making technologies, towards a better understanding of musical interaction.

Introduction

In this article we discuss the interaction between humans and machines with the purpose of making music. We first present a generalized model of a musical instrument, to which we simply refer as a *machine*. Then we make an analogy between the model of the musical machine and a model of the human with whom it interacts. Finally, we examine the human and machine models acting within a system of musical interaction.

The discussion is formulated in light of recent scholarly literature mostly related to the development of electronic musical instruments, as they have seen great development and are a logical step after non-electronic musical instruments. Take for instance the guitar followed by the electric guitar and then the electric guitar as a controller of software. The design of electronic

musical instruments has progressively diverted from the design of non-electronic musical instruments and the interaction of human beings with these musical machines has changed in turn, however still retaining some of its fundamental characteristics.

We define the concept Human-Machine Musical Interaction as a holistic view of the interactions between humans and machines with the goal of making music. The concept applies to any human, not necessarily one that has musical training or one that practices music professionally. Likewise, we consider machines as a generalization of musical instruments, whether they have computational capabilities or not.

The dynamics of Human-Machine Musical Interaction described in the following sections are based on the assumption that musical machines can be considered agents, whose behavior and inner workings have evolved to the point of resembling their human users. In this conception a musical instrument becomes an entity coexisting with the human, instead of a tool used by the human. Machines and humans communicate via various signals produced by different modalities. With the most obvious mode being auditory, we will explore how interactive systems also integrate visual and haptic signals for communication. We also remark that the information that those signals carry can be encapsulated under the broad term *gesture* (Cadoz & Wanderley, 2000; Jensenius et al., 2009). The discussion describes the process in which signals emanating from humans and machines affect each other in ways to produce music. We call this process *gestural agency*, expanding upon previous uses of the term (Cumming, 2000; Hatten, 2012; Robb, 2015).

Generalized model of musical machines

We start by analyzing electronic musical instruments, the most recent evolutionary stage of musical instruments. They can be the most complex kind of musical instruments, offering possibilities unforeseen in non-electronic musical instruments. We assume that very sophisticated realizations of electronic musical instruments could offer the same musical interaction possibilities of non-electronic musical instruments. Therefore, a model of electronic

musical instruments shall also serve for non-electronic musical instruments. This generalized model will allow us to analyze the interaction of musical machines with humans. Wanderley (2001) proposed to study the interaction of musicians and electronic musical instruments by dividing the inquiry into four parts: 1) definition and typologies of gesture; 2) gesture acquisition and input device design; 3) mapping of gestural variables to synthesis variables; and 4) sound synthesis. Broadly speaking, musical gesture has been defined as “human body movement that goes along with music” (Jensenius et al., 2009, p. 13). In Wanderley’s proposal, gesture refers to the bodily movement that is communicated to the musical instrument to produce musical sound. Further in this text we elaborate on musical gesture, redefining it as a concept beyond bodily movement. Previous to that we build on the latter three parts of Wanderley's proposal, which have been often presented as an electronic musical instrument’s main constituting modules (Wessel & Wright, 2002; Hunt, Kirk & Neighbour, 2004; Armstrong, 2006; Magnusson, 2010; Fabiani, Friberg & Bresin, 2013; de Campo, 2014).

The first module (gesture acquisition and input device) is comprised of an interface called *controller*, which is a device or several devices, for example a keyboard, knobs, buttons and so forth. The second module (mapping of gestural variables to synthesis variables) involves the connections between the first and the last modules. The last module (sound synthesis) is a device or system to produce a result from the instrument. This result is at least sound, but sometimes might also be a visual display, for example to inform the user about the state of the instrument. The human user (e.g., a musician) is connected to the first module by means of the controller. Using this interface the human transmits an action, also called gesture, to the instrument. This action is a signal, which is transduced from its original form into electrical signals, analog or digital, flowing towards the final module.

Caution is needed when interpreting the human-machine musical interaction as being modular. Modularity implies that the steps involved in the interaction are categorical and discrete. However, the true nature being represented is continuous and often the boundaries of the modules are not evident. This continuous nature can be explained by looking at some non-electronic instruments. In a piano, organ, or harpsichord the model is quite explicit as the first module represents the keyboard and additional commands to modify the sound. The module in

the middle represents the connections between keys and other commands (e.g., pedals, levers, and knobs) to the mechanism that produces sound, which are represented by the final module. In these instruments each module is associated with a physical part of the instrument. Now, how can we describe the violin with this modular approach? It can be argued that the first module corresponds to the bow and strings, the final module to the body, and the module in the middle to the bridge as it connects the strings to the body (Chadabe, 2002). However, all the parts of a violin play a role in each of the three modules. For example, the bow and strings are not only related to the communication with the musician's body, but also with sound production. In this case the categorical modular approach becomes problematic. In the clarinet, the control module can be thought of as the mouthpiece and keys, mapping given by the key mechanism and sound production at the body. Again, a modular approach poses important overlaps as, for example, the mouthpiece, apart from controlling the sound, is also producing the sound. These two instruments, as do probably many other non-electronic instruments, require a model that overcomes the rigid association of the model's modules with a physical part of the instrument.

The problem vanishes when the modules are understood as processes rather than physical parts of instruments. Therefore, in what follows we will call the first module simply Input and the last module Output, the module in between remaining as Mapping. For example, the Input for the violin is the bow and strings; while Mapping is strings, fretboard, bridge, and body; and Output is bow, strings, fretboard, and body. This model, now more general, also serves for electronic musical instruments in which the mapping can be designed to from simple to quite complex. One connection might be directly made from the controller to the synthesizer, whereas other connections might undergo some further processes in between, for example learning the connection that the user prefers (Caramiaux & Tanaka, 2013). To this extent, the incorporation of machine learning into electronic musical instruments has been a crucial evolution. It allows a user to build instruments quickly, for example just by mapping parameters between controller and synthesis, a paradigm that has been called *composed instrument* (Schnell & Battier, 2002; Fiebrink et al., 2010).

In this scenario, the signals are not always flowing in the direction from the controller to the sound production. In a musical instrument, most possibly an electronic one, there might be

internal feedback signals going from the Output module back to the Input module or into the Mapping module without leaving the Output module, enabling the musical machine to monitor its own internal behavior. Theoretically an unlimited amount of internal feedback signals can be mapped in any direction. Likewise, the signals entering the machine, the signals exiting the machine and all internal feedback signals can be of any kind. Obviously, the one kind of signal that is required to be output from a musical instrument is an auditory signal. Returning to non-electronic musical instruments we can observe that visual and haptic signals aid in the playing. For example, looking at the piano keyboard and feeling the keys or visually finding the approximate location of a position when playing the double bass and feeling the weight of the string at that point. Even olfactory signals are present in the interaction with musical instruments, although they are often less influential in making music. In the case of electronic musical instruments, output other than audio has been used more or less in the same fashion as in non-electronic musical instruments, as an aid to the production of sound. An exception is the case of electronic musical instruments that incorporate a visual component that is not an aid to playing but a signal with the same valuation as the auditory output signal. In the same way, the electronic musical instrument could produce haptic and olfactory signals.

Clearly the generalized model of musical machines turns aside from a more conservative conception in which the design and use of an electronic musical instrument is more or less bound to the same rules as for non-electronic musical instruments.

Human and machine embodied music cognition

The human being that interacts with a machine can be represented using the same logic as for the machine. The Input, Mapping and Output modules that constitute a musical instrument can be seen as parallel to Perception, Cognition, and Action in the human, from a Cartesian point of view (Armstrong, 2006). Likewise, if we consider a machine in which the Mapping module has a component that can learn, we can easily notice that it parallels the human module between Perception and Action: Cognition. In this view, the intermediate modules Cognition and

Mapping are the ones actively modifying the signals flowing in an enactive direction: towards Action in the human and towards Output in the machine.

Although it might seem useful to represent musical interactions between humans and machines, this approach holds a caveat. It places cognition between perception and action, like a sandwich (Hurley, 2002, p. 401), as if cognition was a process concerned solely with the brain. This idea has been rendered obsolete in general by the viewpoint of Embodied Cognition (Anderson, 2003) and in particular by the viewpoint of Embodied Music Cognition (Leman, 2008) which, generally speaking, consider cognition as a process involving not only the brain but also the body and its environment.

The embodied quality of a model of a musical human being can be illustrated in the same way as we described a model of musical machines in the previous section: in terms of processes rather than physical units. This means that its modules do not necessarily represent areas of the brain, limbs, or sensory organs. In this way each module can represent functions performed by more than one organ. It can also represent one organ associated with more than one module. It becomes clear that the module between Perception and Action, if we think of it as the human brain, cannot account for the cognitive processes that occur throughout the whole body (Wheeler, 2005). Therefore, we prefer to call the human's intermediate module Connections, allowing the term cognition to refer to a process occurring throughout the three modules. Also, we take into account that there are internal feedback signals connecting the modules without being sent outside. These signals represent, for example, the modification of a perception organ's behavior triggered by an action, with or without the mediation of thinking.

On the side of the machine, we can state similarly that *machine cognition* takes place. The concept of machine cognition, as in the human counterpart, comprises perception, understanding and learning by a machine (Haikonen, 2007). In the model of a musical machine that we have described so far, these areas are more or less represented by the modules Input and Mapping. Yet, in the design of machines that interact with their environment, it has been necessary to consider cognition as an enactive process that leads to consciousness, being embodied as the process involves the awareness of the machine's body and its surroundings (Holland, 2007). The awareness of the machine is related to its internal feedback signals, as

described in the previous section. Akin to the musical human model, we shall consider that the cognitive capacity of a musical machine takes part not only in the Mapping module but also at the Input and Output modules.

The generalized model of musical machines and the model of the musical human being now appear highly resemblant of each other. When human and machine interact, the Action module of the human is connected to the Input module of the machine and the Output module of the machine is connected to the Perception module of the human. In the resulting loop, the signals are continuously updated by the human and the machine. Moreover, these connections are not exclusive. Human and machine “catch” signals with their first modules, whether these signals are their own (e.g., to listen to their own sound) or coming from elsewhere (see FIGURE 1).

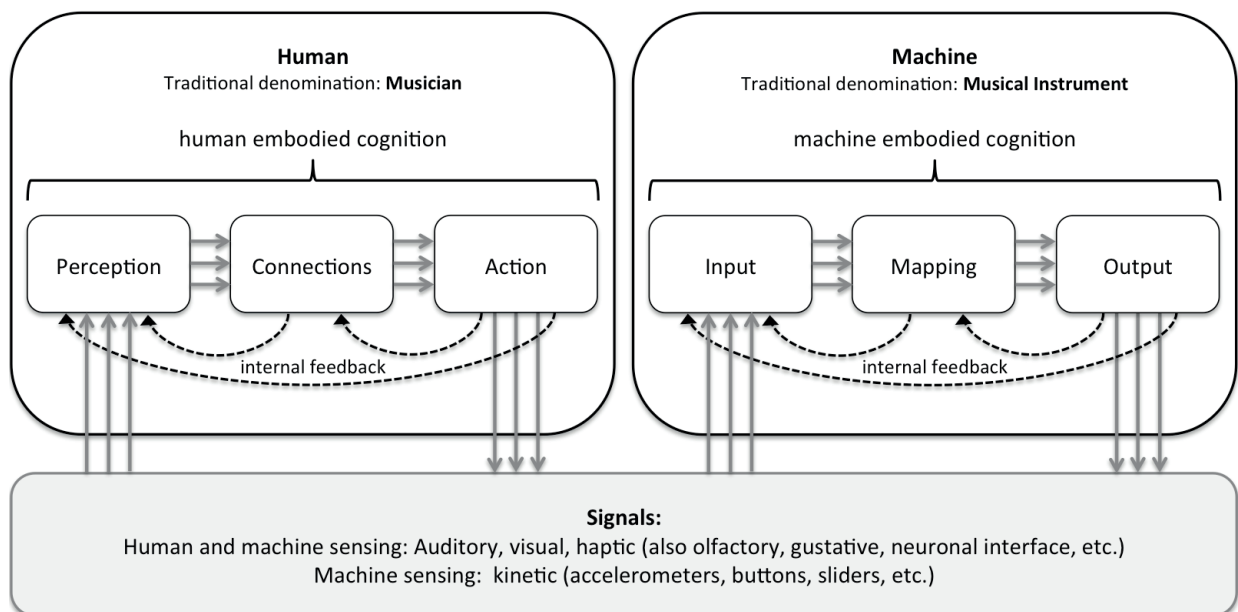


FIGURE 1: Human and machine are agents connected by signals. The model allows for the incorporation of more agents into a network.

Gestural Agency

Now we examine the interactions between the human and machine models that we have described. In this interaction system the flow of signals is enactive, in other words, they are directed towards Action in the Human and Output in the Machine. This implies that there is a target receiver of an outgoing signal. Thus, both human and machine have the potential to become agents exerting equal or unequal influence upon each other to produce music. These agents' behavior is affected by the conditions imposed, demanded or proposed by other agents (Gurevich & Treviño, 2007; Bown, Eldridge & McCormack, 2009). The process starts with exploration and discovery of the intentions of the other agents, gradually turning into an objective-based task as a musical aesthetic emerges (Caramiaux et al., 2014). In this way the machine resembles an entity, more a musician than a musical instrument (Van Nort, 2011).

The evolution of these agents requires some internal adaptation. In the model of the human that we described, the Connections module is dynamically affected by reasoning and experience, to satisfy demands from the musical environment or to accomplish individual musical goals. The machine can also go through such a process, as the mapping module can be affected by generative algorithms and machine learning.

This idea seems to challenge a more traditional conception of a musical instrument as an extension of the musician's body (Nijs, Lesaffre & Leman, 2009; Simoens & Tervaniemi, 2013). In that conception the flow of signals is bi-directional, as the instrument can affect the way the musician plays as much as the musician affects the sound of the instrument. We integrate this notion with the idea of musical machines being agents, regarding the relationship of a musical machine and a human being as two-fold. While the musical machine can be seen as external to the human body, it is still connected to the human body by the gestural signals, both comprising a greater organism. This organism is the interaction system that we have been describing, where not only the musical machine is an extension of the human body but also the human body is an extension of the musical machine. Rather than bi-directional, the signals in this system are multi-directional, as the output or action of one agent can be "caught" by the same agent or by other agents. Following this reasoning, musical interaction becomes a network of musical agents in a

musical ecosystem.

For example, one agent proposes a musical idea like a rhythmical pattern or a tonic note, which is acted upon by the rest of the agents in the system. This could be the case of musicians playing together, a single musician interacting with a musical instrument, or intelligent machines playing together. Agents have limitations and thus impose constraints, to which the interacting agent has to respond. A violin cannot play a very low drone of Tuvan chant as would a group of vocal performers or an electronic musical instrument. The one who plays the violin adapts to take most of the instrument. In the same way it would be possible for a musician to learn how to interact with a musical instrument that allows to shape complex drones but does not allow for such a detailed command of pitch and dynamics as the violin does.

As suggested in the previous section, the agents in the musical ecosystem are connected by *multimodal* signals (Françoise, Schnell & Bevilacqua, 2013). These signals act as interfaces between agents (Di Scipio, 2003) carrying information that we call *gesture*, regardless of the nature of the signal. In this context, the concept of gesture is not bound to movement of the human body. Gesture can be visual, kinetic, auditory, or of any other kind, as long as it is information that could be perceived by an agent as meaningful and thus having musical influence. In this sense we conform to the idea that gestures are information that can carry more information within them (Godøy, Jensenius & Nymoen, 2010). The signals carrying gesture are the pathways that conform the network of agents, which can be abstracted as a topological space (Mazzola & Andreatta, 2007; Van Nort, Wanderley & Depalle, 2014). This space can be called *gestural space*, as the substance of the topology is gesture. Gesture can be observed at any point of the gestural space. In less abstract terms, there is always gesture, and therefore agency, in the various communication channels between those who are making music.

The dimensions of the gestural space can be multiple as they are given by the parameters that agents can modify in the signals they process. For example, parameters can be pitch, level, filter cutoff and so forth. The production of musical sound can be seen as navigating through this space (Choi, Bargar & Goudeseune, 1995; Chadabe, 2002; Schwarz, 2012). The amount of dimensions in gestural space could be large, providing the musician with a rich set of possibilities. This could be the case of a symphonic orchestra or an electronic musical system.

However, humans can only handle a limited amount of dimensions in real-time and thus the rich set of possibilities remains feasible only in non-real-time music making such as composing or rehearsing (Tubb & Dixon, 2015).

In greater or lesser amount there are correspondences between the signals of the system, whether they are auditory, visual, or of any other kind. Several investigations have pointed out the resemblances between bodily movement and music (Godøy, Jensenius & Nymoen, 2010; Toiviainen, Luck & Thompson, 2010; Cox, 2011). For instance, the movement of the human body that produces a sound has remarkable similarities to bodily movement as a response to listening to that sound (Altavilla, Caramiaux & Tanaka, 2013). In the same way the bodily action of playing an instrument has similarities with the sound produced by the instrument (Haga, 2008). Another example is movement helping to visually convey the meaning of the music (MacRitchie, Buck & Bailey, 2013). As phenomena correlates among the gesture space, it is possible to reduce dimensionality to bring the topology closer to a human experience of real-time performing (Arfib et al., 2002; Zappi & McPherson, 2014). An electronic musical instrument that offers a highly dimensional gestural space will implement dimensionality reduction to respond to the ability of a human to handle a limited amount of dimensions in real-time music making. This is a demand imposed by the human agent upon the machine agent.

As a consequence of the exposed analysis, we call *gestural agency* the influence that an agent exerts over other agents within a musical ecosystem. The extent of this influence is a means of power that an agent has on shaping its musical environment, including the behavior of other agents. A conservative view on this has the human in possession of most of the power but we can see that in this system the human is more a participant than a user (Kaipainen et al., 2011). In this way the whole musical ecosystem is enactive as the production of signals is linked to a function, a role of each agent in relation to the other agents (Matyja & Schiavo, 2013).

As the interaction within a gestural agency system demands action to be taken in order to reach a musical goal, agents are necessarily presented with challenges. It has been argued that a musical instrument needs to pose a balanced challenge that is neither too much to be frustrating nor too little to be unappealing in order to be interesting (Wanderley & Orio, 2002; Wessel & Wright, 2002; Levitin, McAdams & Adams, 2002; McDermott et al., 2013; Fabiani, Friberg &

Bresin, 2013). This notion conforms to the concept of *flow*, defined as an optimal state of well-being achieved by an activity that provides challenges or opportunities for action accommodating the skill level of an individual (Nakamura & Csikszentmihalyi, 2014). As skills grow, *flow* is maintained by newer and slightly bigger challenges.

In sum, a musical machine demands a certain gesture from the human, to which the human has to respond, thereby changing internal connections to adapt to this requirement. This can take the form of learning to use an instrument, but can also be to learn how to interact with a more independent musical agent. In the same way the machine can be challenged by the human, leading to the reconfiguration of the machine. The *flow* state can be achieved by the regulation of the expectations and tasks by the human. Also, this regulation could be carried out on the side of the machine, especially in the case of electronic musical instruments incorporating learning algorithms based on reward (Smith & Garnett, 2012).

Conclusion

We have presented a generalized model for a musical machine that resembles a model for human embodied cognition. Both models depict enactive agents interacting in a musical ecosystem. We have observed that these agents are connected by signals that carry gesture. Gesture is the means that an agent has to exert influence over other agents to produce a musical result. This understanding of musical interaction between humans and machines fits well to traditional and newer technologies for making music. In practice this serves as a framework to analyze musical interactions that integrate humans, traditional musical instruments, and newer electronic musical instruments.

References

- Altavilla, A., Caramiaux, B., & Tanaka, A. (2013). Towards gestural sonic affordances. In *Proceedings of the International Conference on New Interfaces for Musical Expression* (pp. 61–64). Daejeon, Republic of Korea: Graduate School of Culture Technology, KAIST.
- Anderson, M. L. (2003). Embodied cognition: A field guide. *Artificial intelligence*, 149(1), 91–130.
- Arfib, D., Couturier, J. M., Kessous, L., & Verfaillie, V. (2002). Strategies of mapping between gesture data and synthesis model parameters using perceptual spaces. *Organised Sound*, 7(02), 127–144.
- Armstrong, N. (2006). *An enactive approach to digital musical instrument design* (Doctoral dissertation). Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.115.5347&rep=rep1&type=pdf>
- Bown, O., Eldridge, A., & McCormack, J. (2009). Understanding interaction in contemporary digital music: from instruments to behavioural objects. *Organised Sound*, 14(02), 188–196.
- Cadoz, C., & Wanderley, M. M. (2000). Gesture-music. In M. M. Wanderley & M. Battier (Eds.), *Trends in gestural control of music* (pp. 71–94). Paris, France: IRCAM - Centre Pompidou.
- Caramiaux, B., & Tanaka, A. (2013). Machine learning of musical gestures. In *Proceedings of the International Conference on New Interfaces for Musical Expression* (pp. 513–518). Daejeon, Republic of Korea: Graduate School of Culture Technology, KAIST.
- Caramiaux, B., Françoise, J., Schnell, N., & Bevilacqua, F. (2014). Mapping through listening. *Computer Music Journal*, 38(3), 34–48.
- Chadabe, J. (2002). The limitations of mapping as a structural descriptive in electronic instruments. In *Proceedings of the International Conference on New Interfaces for Musical Expression* (pp. 38–42). Dublin, Ireland.

- Choi, I., Bargar, R., & Goudeseune, C. (1995). A manifold interface for a high dimensional control space. In *Proceedings of the 1995 International Computer Music Conference, Banff Centre for the Arts, Canada* (pp. 385–392).
- Cox, A. (2011). Embodying music: principles of the mimetic hypothesis. *Music Theory Online*, 17(2).
- Cumming, N. (2000). *The sonic self: Musical subjectivity and signification*. Bloomington: Indiana University Press.
- de Campo, A. (2014). Lose control, gain influence – Concepts for Metacontrol. In *Proceedings of the 2014 International Computer Music Conference, Athens, Greece* (pp. 217–222).
- Di Scipio, A. (2003). ‘Sound is the interface’: from interactive to ecosystemic signal processing. *Organised Sound*, 8(03), 269–277.
- Fabiani, M., Friberg, A., & Bresin, R. (2013). Systems for interactive control of computer generated music performance. In A. Kirke & E. Miranda (Eds.), *Guide to Computing for Expressive Music Performance* (pp. 49–73). London, United Kingdom: Springer-Verlag.
- Fiebrink, R., Trueman, D., Britt, C., Nagai, M., Kaczmarek, K., Early, M., ... Cook, P. (2010). Toward understanding human-computer interaction in composing the instrument. In *Proceedings of the 2010 International Computer Music Conference, New York, NY, USA* (pp. 135–142).
- Françoise, J., Schnell, N., & Bevilacqua, F. (2013). A multimodal probabilistic model for gesture-based control of sound synthesis. In *Proceedings of the 21st ACM International Conference on Multimedia, Barcelona, Spain*. (pp. 705–708). New York, NY, USA: ACM.
- Godøy, R. I., Jensenius, A. R., & Nymoen, K. (2010). Chunking in music by coarticulation. *Acta Acustica united with Acustica*, 96(4), 690–700.
- Gurevich, M., & Treviño, J. (2007). Expression and its discontents: Toward an ecology of musical creation. In *Proceedings of the International Conference on New Interfaces for Musical Expression* (pp. 106–111). New York City, NY, United States.

- Haga, E. (2008). *Correspondences between music and body movement* (Doctoral dissertation). Retrieved from <https://www.duo.uio.no/handle/10852/26916>
- Haikonen, P. O. (2007). *Robot brains: circuits and systems for conscious machines*. Chichester: John Wiley & Sons.
- Hatten, R. (2012). Musical Forces and Agential Energies: An Expansion of Steve Larson's Model. *Music Theory Online*, 18(3).
- Holland, O. (2007). A strongly embodied approach to machine consciousness. *Journal of Consciousness Studies*, 14(7), 97–110.
- Hunt, A., Kirk, R., & Neighbour, M. (2004). Multiple media interfaces for music therapy. *IEEE MultiMedia*, 11(3), 50–58.
- Hurley, S. L. (2002). *Consciousness in action*. Cambridge, MA, USA: Harvard University Press.
- Jenselius, A. R., Wanderley, M. M., Godøy, R. I., & Leman, M. (2009). Musical gestures. In R. I. Godøy & M. Leman (Eds.), *Musical gestures: Sound, movement, and meaning* (pp. 12–35). New York, NY, USA: Routledge.
- Kaipainen, M., Ravaja, N., Tikka, P., Vuori, R., Pugliese, R., Rapino, M., & Takala, T. (2011). Enactive systems and enactive media: embodied human-machine coupling beyond interfaces. *Leonardo*, 44(5), 433–438.
- Leman, M. (2008). *Embodied music cognition and mediation technology*. Cambridge, MA, USA: MIT Press.
- Levitin, D. J., McAdams, S., & Adams, R. L. (2002). Control parameters for musical instruments: a foundation for new mappings of gesture to sound. *Organised Sound*, 7(02), 171–189.
- MacRitchie, J., Buck, B., & Bailey, N. J. (2013). Inferring musical structure through bodily gestures. *Musicae Scientiae*, 17(1), 86–108.
- Magnusson, T. (2010). Designing constraints: Composing and performing with digital musical systems. *Computer Music Journal*, 34(4), 62–73.

- Matyja, J. R., & Schiavio, A. (2013). Enactive music cognition: Background and research themes. *Constructivist Foundations*, 8(3), 351–357.
- Mazzola, G., & Andreatta, M. (2007). Diagrams, gestures and formulae in music. *Journal of Mathematics and Music*, 1(1), 23–46.
- McDermott, J., Gifford, T., Bouwer, A., & Wagdy, M. (2013). Should music interaction be easy?. In *Music and Human-Computer Interaction* (pp. 29–47). London, United Kingdom: Springer-Verlag.
- Nakamura, J., & Csikszentmihalyi, M. (2014). The concept of flow. In *Flow and the Foundations of Positive Psychology: the collected works of Mihaly Csikszentmihalyi* (pp. 239-263). Dordrecht, Netherlands: Springer.
- Nijs, L., Lesaffre, M., & Leman, M. (2009). The musical instrument as a natural extension of the musician. Retrieved from <https://biblio.ugent.be/publication/844863/file/944424>
- Robb, H. J. (2015). Imagined, Supplemental Sound in Nineteenth-Century Piano Music: Towards a Fuller Understanding of Musical Embodiment. *Music Theory Online*, 21(3).
- Schnell, N., & Battier, M. (2002). Introducing composed instruments, technical and musicological implications. In *Proceedings of the International Conference on New Interfaces for Musical Expression* (pp. 156–160). Dublin, Ireland.
- Schwarz, D. (2012). The sound space as musical instrument: Playing corpus-based concatenative synthesis. *Proceedings of the International Conference on New Interfaces for Musical Expression* (pp. 250-253). Ann Arbor, Michigan, USA: University of Michigan.
- Simoens, V. & Tervaniemi, M. (2013) Musician–Instrument Relationship as a Candidate Index for Professional Well-Being in Musicians. In *Psychology of Aesthetics, Creativity, and the Arts*, 7(2), 171-180.
- Smith, B. D., & Garnett, G. E. (2012). Reinforcement learning and the creative, automated music improviser. In *Evolutionary and Biologically Inspired Music, Sound, Art and Design* (pp. 223–234). Berlin; Heidelberg, Germany: Springer-Verlag.

- Toiviainen, P., Luck, G., & Thompson, M. R. (2010). Embodied meter: hierarchical eigenmodes in music-induced movement. *Music Perception*, 28(1), 59–70.
- Tubb, R., & Dixon, S. (2015). An evaluation of multidimensional controllers for sound design tasks. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, Seoul, Republic of Korea (pp. 47–56). New York, NY, USA: ACM.
- Van Nort, D. (2011). Human: Machine: Human: Gesture, sound and embodiment. *Kybernetes*, 40(7/8), 1179–1188.
- Van Nort, D., Wanderley, M. M., & Depalle, P. (2014). Mapping control structures for sound synthesis: Functional and topological perspectives. *Computer Music Journal*, 38(3), 6–22.
- Wanderley, M. M. (2001). Gestural control of music. In *International Workshop Human Supervision and Control in Engineering and Music*, Kassel, Germany. Retrieved from <http://recherche.ircam.fr/equipes/analyse-synthese/wanderle/pub/kassel/>
- Wanderley, M. M., & Orio, N. (2002). Evaluation of input devices for musical expression: Borrowing tools from HCI. *Computer Music Journal*, 26(3), 62–76.
- Wessel, D., & Wright, M. (2002). Problems and prospects for intimate musical control of computers. *Computer Music Journal*, 26(3), 11–22.
- Wheeler, M. (2005). *Reconstructing the cognitive world: The next step*. Cambridge, MA: MIT press.
- Zappi, V., & McPherson, A. (2014). Dimensionality and appropriation in digital musical instrument design. In *Proceedings of the International Conference on New Interfaces for Musical Expression* (pp. 455–460). London, United Kingdom: Goldsmiths, University of London.



II

MODELLING PERCEIVED SEGMENTATION OF BODILY GESTURES INDUCED BY MUSIC

by

Juan Ignacio Mendoza & Marc Richard Thompson, 2017

In Proceedings of the 25th Anniversary Conference of the
European Society for the Cognitive Sciences of Music

<http://urn.fi/URN:NBN:fi:jyu-201711024121>

Modelling Perceived Segmentation of Bodily Gestures Induced by Music

Juan Ignacio Mendoza¹, Marc Richard Thompson²

Department of Music, Art and Culture Studies; University of Jyväskylä, Finland

¹juigmend@student.jyu.fi, ²marc.thompson@jyu.fi

ABSTRACT

This article presents an ongoing investigation whose goal is to model perceived segmentation of music-induced bodily gestures. The investigation consists of three stages. The first stage is a database of multimodal recordings of people moving to music. The data of these recordings are video and motion-capture (acceleration and position at several points of the body). In the second stage the videos produced in the first stage are manually segmented. This is regarded as ground truth for the evaluation of the performance of an automatic gesture segmentation system developed in the third stage of the study. This system extracts kinetic features from motion-captured data. Then a novelty score is computed from the kinetic features. The peaks of the novelty score indicate segmentation boundaries. So far the kinetic features that have been evaluated are composed of only one windowed statistical function. None of them yields a reasonable similarity between computed and perceived boundaries. However, different functions of the kinetic features yield considerably similar results between perceived and computed boundaries at isolated regions of the data. This suggests that each of these functions performs best on a specific kind of gesture. Further work will consider evaluating kinetic features composed of combinations of functions.

I. INTRODUCTION

A. Background

In line with the Embodied Music Cognition train of thought (Leman, 2008), it has been argued that a person's spontaneous movement when listening to music can reflect the person's perception of the music. Qualitative investigation has observed, for example, that music teachers explain musical sound with bodily movements, especially with their hands (Clayton & Leante, 2011). Quantitative investigation has shown that bodily movement induced by music relates to features of the music, such as periodicity and kinetic energy (Toiviainen, Luck & Thompson, 2010) or tonality (MacRitchie, Buck & Bailey, 2013). The correspondence between music and bodily movement has been studied under the term *musical gesture* (Schneider, 2010). It has been noted that human beings have a remarkable ability to perceive and understand musical gestures by visual observation (Camurri & Moeslund, 2010). The first stage in perception of a gesture is the identification of when and where it starts and ends, a process called *segmentation* (Kahol, Tripathi & Panchanathan, 2004). Further phenomenological inquiry has observed that musical gestures are perceived in different time scales and that the grouping of shorter-scale gestures into larger entities depends on musical structure, a phenomenon called *co-articulation* (Godøy et al., 2016).

Several studies have observed the relation between bodily movement of people making music and moving to music (e.g., dancing) using qualitative analysis of video recordings (Wanderley et al., 2005; King & Ginsborg, 2011; Luck, 2011; Clayton & Leante, 2011; Trevarthen, Delafield-Butt & Schögler, 2011). Because the careful observation of video is a

time-consuming task, these studies have focused in a few examples. Therefore their results, while being important for advancing knowledge, are not appropriate for generalization. In contrast, a large-scale experimental investigation that could yield statistically relevant results, would take the effort of people watching many videos. These videos should show a range of individuals moving to different kinds of music. The observation of videos should include precise annotations of where gestures occur and a description of them. Such an endeavor appears to be prohibitive in terms of human resources. Thus, it seems reasonable to automate the process, which requires first to model human perceived segmentation of gestures.

B. Aim

The purpose of this study is to model perceived segmentation of music-induced bodily movement.

II. METHODS

This section presents the three-stage methodology used in this investigation project. Data is periodically added and methods are refined as the investigation advances. What follows corresponds to the state of the project as of April, 2017.

A. Multimodal Database

1) *Aim*. This stage of the investigation consists in collecting multimodal data, which allows to observe people's spontaneous movement to music. The data modalities are:

- Tri-axial position
- Tri-axial acceleration
- Video

2) *Participants*. N = 12, of which 7 (58.3%) are female and 6 (41.7%) are male. Their range of ages is 23 to 53, median 33. All of them are either degree students, researchers or other staff at the University of Jyväskylä. None of them is associated with the Music, Art and Culture Studies department or with research in musicology. All participants sign a document giving consent to the use of recorded data for research and communication thereof, including audio and video recordings.

3) *Apparatus*. Data is collected at the motion capture laboratory of the Music, Art and Culture Studies department of the University of Jyväskylä. The apparatus is composed by the following measurement processes:

- **Optical Motion Capture**: An array of 8 Qualisys Oqus cameras track the position of reflective markers attached to a tight suit that the participant wears. Markers are placed on every articulation and ending point of limbs, as well as on the head. Optical motion capture data is recorded using the Qualisys Track

Manager software running in a personal computer. This system is synchronised to an SMPTE signal emitted by a second computer. Also the Qualisys system sends back a synchronisation audio signal to the second computer.

- Tri-axial accelerometers: The participant wears a Thalmic Myo armband on one forearm beneath the motion capture suit. Also the participant holds a Nintendo Wii-remote (“wiimote”) controller with the hand of the arm that wears the Myo armband. Data from these devices are simultaneously recorded at a rate of 100 Hz in the second computer, using software made with the Pure Data programming environment (Puckette, 1997). This software also simultaneously records audio.
- Audio: Stimuli is presented to the participant using two Genelec 8030-A studio loudspeakers with their base at 110 cm. from the floor. A microphone hanging from the ceiling is connected to the audio system of the second computer, which simultaneously records this audio stream (i.e., room audio) in one audio channel and the audio synchronization signal from the optical motion capture system in a second channel. The starting and ending of the audio recording is set to be at the same time of the accelerometer devices’ data recording. The audio signal is used later to set a common starting time for accelerometer, optical motion capture and room audio.
- Video: Two small digital cameras (Vivitar DVR-786 and Sony DSC-W610) on flexible portable tripods record video and room audio. They are placed together, pointing perpendicular to the wall. The room shape is a rectangle. The image shows the participant’s full-body against a white wall. Redundancy of video recordings serves as a backup strategy. Later the video stream is synchronized to the accelerometer and optical motion capture using the room audio. This method allows flexibility when positioning the cameras, opposed to having cameras fixed to the wall or mounted on cumbersome rigging.

4) *Stimuli*. The list below shows the excerpts of music that have been used and a brief description that explains the choice.

- “Bouzouki Hiphop” (Tetarto Hood, 2014) from the beginning to 45.7 s. with no fade-in or fade-out. This is Rembetiko instrumental music mixed with Hip-hop bass and drums, published on the Internet by an independent artist. Tempo is 90 BPM and meter is 4/4. All participants declared to not know this piece.
- “Minuet in G Major” (Petzold, ca. 1725). MIDI rendition with piano sound, from beginning to end (104 bars, 93 s.) with no fade-in or fade-out. Tempo is ca. 128 BPM and meter is 3/4. All participants declared to know this piece.

- “Ciguri” (Otondo, 2008) from 56 to 180 s. with fade-out the last 5 s. This is an electroacoustic piece that has no perceivable beat that indicates tempo and that has “*an insistent and virtually isochronic rapid percussion attack, together with one or more streams of sustained electroacoustic sound with somewhat clear pitch structure*” (Olsen, Dean & Leung, 2016). All participants declared to not know this piece.
- “Stayin’ Alive” (Bee Gees, 1977) from the beginning to 108 s. with fade-out the last 2.3 s. Tempo is 104 BPM and meter is 4/4. All participants declared to know this piece.

5) *Procedure*. Data recording is done with one participant at a time. Participants are asked to move spontaneously to the stimulus when it starts sounding through the loudspeakers. They are not asked to dance as it was observed in pilot experiments that if they are asked to dance they feel inhibited because they are afraid to fail. This fear derives from the association of the word “dance” with movements that have to be done correctly, as inferred from participants’ accounts. However, if participants are asked to *move to music* this inhibition disappears. In fact, participants usually ask “*Do I have to dance?*”. When they do ask this question, they are explained that they can dance if they want, otherwise they can move freely.

Each stimulus is presented twice. Participants are asked on the first presentation to move to the music without any constraint other than an area of approximately 9m², which corresponds to the bounds of the Optical Motion Capture and Video Capture systems. The second time participants are asked to hold the Wii-mote with one hand and *dance* only with that arm (this arm is also wearing the Myo armband). In this condition participants are asked to remain at the center of the area facing to a corner of the room. This is done to get in the video recording the most complete visualization of the arm’s movement. In this condition participants are allowed to move the rest of the body naturally as long as the previous constraints are not violated. This procedure (called “trial”) is repeated for each stimulus.

Stimuli are presented in the order of the list above (4. *Stimuli*). However, participants were told that the first stimulus (Bouzouki...) was *just for practice*. Indeed that trial was intended to be a practice so that the participant could get familiarity with the procedure. Still, data for this stimulus is recorded and kept. Participants are allowed to rest as much as needed between trials.

B. Ground Truth

1) *Aims*. In this stage the videos from the Multimodal Database are manually segmented in two conditions. In each condition the time location of segmentation boundaries is recorded. This task is called *annotation*.

- Real-time annotation: Videos with their corresponding audio are segmented as they are watched.

- Non-real-time annotation: Videos without audio are segmented as they are watched, with the option of scrolling back and forth to refine the annotation.

2) *Participants and Stimuli*. Participants of this experiment are called *annotators*, to differentiate them from the participants in data collection for the Multimodal Database. So far two annotators have performed only the Non-real-time task upon the video corresponding to single-arm movement to the “Stayin’ Alive” stimulus. These annotators are doctoral students of musicology, one of them the first author of this article. This data has been regarded as preliminary.

3) *Apparatus*.

- Real-time annotation: A personal computer running a custom-made piece of software made with the Pure Data programming environment, which automatically presents the video and records the elapsed time when depressing a key of the computer’s keyboard. These times are recorded in a comma-separated-values text file.
- Non-real-time annotation: A personal computer running the Reaper digital audio editing software (Cockos Reaper, 2010). This system allows video playback at different speeds, scrolling through the video and accurately placing markers, which can be assigned different colors. These markers are exported as a comma-separated-values text file.

4) *Procedure*.

- Real-time annotation: The participant is presented with a video of the Multimodal Database and asked to depress a key when noticing “a change of movement”. This wording is meant to indicate a change in bodily gesture without giving an extensive explanation of the concept.
- Non-real-time annotation: The participant is asked to place markers where there is a change of movement. Additionally, the participant is asked to group the annotated markers into larger structures, without further explanation of what this means. To indicate the boundaries of these bigger structures a new set of markers is placed on top of the existing ones, with a different colour.

5) *Data Analysis*. Responses by all participants are summarized into a single compound response for each condition. This is done using Kernel Density Estimation, which produces a curve of density. The peaks of this curve, over a threshold, indicate the segmentation boundaries of the annotators as a group. Additionally, the digital audio file of the corresponding stimulus is segmented using Music Information Retrieval techniques (Lartillot, Toivainen & Eerola, 2008).

C. Automation

1) *Aim*. In this stage an automated system is developed with the goal of predicting human perceived boundaries. The system takes as input the accelerometer or optical motion-

capture data from the Multimodal Database. Performance of the system is assessed by comparing its output with the corresponding annotations obtained in the Ground Truth stage. The main challenge is to find an appropriate combination of kinetic features and their parameters that are consistent and distinct for each gesture.

2) *Procedure*. For now only accelerometer data from the Wii-mote is being considered. This means that data consists of tri-axial acceleration of a single moving point. This is a starting point and it is thought that the same method could be applied for data of any of the optical-motion-capture markers individually or in combination. The core of the system was developed by Foote and Cooper (2003) for media segmentation. This method has been adapted and expanded to be used in this investigation for the segmentation of kinetic data. The procedure involves the choice of multiple *free variables*, which determine the system’s performance. In its current state of development, the procedure is as follows:

- Downsample raw acceleration data from 100 Hz to 10 Hz. This sampling rate is enough to achieve satisfactory results at a lower computational cost than using full resolution.
- Compute magnitude (Euclidean norm). This is a free variable, here called “Input Data Type”, as either the tri-axial acceleration signal or its magnitude may be used as input for the next step.
- Compute windowed functions. A set of statistical functions is computed individually over a sliding window with hop of a single sample. The functions currently used are a subset of functions evaluated by previous investigation on medical surveying of physical activity using accelerometers (Lara, & Labrador, 2013; Machado et al., 2015). To minimize distortion at the borders, the signals are extended at the beginning with the value of the first sample and at the ending with the value of the last sample. The length of each of these extensions is half of the sliding window. The width of the window is a free variable. Also the choice of functions is a free variable.

The functions currently used are the following:

- kurtosis
 - skewness
 - mean
 - root mean square
 - standard deviation
 - mean absolute deviation
 - interquartile range
 - centered zero-crossings count
- Convolve the output of the previous step with a Gaussian kernel and rescale to a range between 0 and 1. The same extension procedure of the previous step is applied to the input of this step before convolution. The window of the kernel is a free variable. If the

window length is set to zero, then convolution is not done but only rescaling.

- Compute a distance matrix of a single function or combined functions. Here the outputs of one or more functions are dimensions of a matrix. Euclidean distance between each point with all the other points is computed to obtain the distance matrix. Additionally, for each function output there is a scaling factor $C \{0 < C \leq 1\}$, which determines the contribution (i.e., “weight”) of a function to the computed distances.
- Compute a Novelty Score by convolving a Gaussian-smoothed Checkerboard Kernel with volume $V=1$, along the diagonal of the distance matrix. Before performing the convolution, the matrix is extended to half the length of the kernel. The extension section at the beginning is set to the mean value of the section of the kernel that is in the non-extended distance matrix. The same procedure is done at the ending. These extensions with mean values help to reduce the distortion at the beginning and ending. Here the free variable is the length of the kernel.
- Extract peaks from the novelty score over a threshold. Here the free variable is the factor of the threshold $T \{0 < T \leq 1\}$. These peaks indicate the computed segmentation boundaries.

Computed segmentation boundaries are then compared with perceived segmentation boundaries (i.e., ground truth) of the corresponding videos, by means of a similarity measure. An earlier version of this measure was used to assess similarity of computed and perceived segmentation boundaries of electroacoustic music (Mendoza, 2014). In this study an updated version is used, which is computed as follows:

- a and b are vectors containing indexes (i.e., time location) of segmentation boundaries, at the downsampled rate. One of them contains perceived boundaries (ground truth) and the other contains computed boundaries (novelty peaks).
- L is the length of the downsampled data. $L_a=L_b$
- N is the amount of indexes. $N_a \geq N_b$
- Compute a distance matrix M_{jk} of vectors a and b :

$$M_{jk} = |a_j - b_k|$$
- Find the minima (m) of rows (r) and columns (c):

have highest similarity with ground truth are manually inspected to find constraints that would facilitate the search by a genetic algorithm. A mixed-integer constrained genetic algorithm has previously been used for a similar problem by an investigation oriented to find the audio features that yield a novelty score that has highest correlation with Kernel Density Estimation of perceived audio segmentation (Hartmann, Lartillot & Toivianen, 2016).

$$\begin{aligned} m_r(j) &= \operatorname{argmin} M_{jk} & k \in [1, n] \\ m_c(k) &= \operatorname{argmin} M_{jk} & j \in [1, n] \end{aligned}$$

The values of a and b at the intersection minima become vectors a' and b' , the closest paired elements from a and b .

- Find the mean distance d from the intersection of minima:

$$d(a, b) = \operatorname{mean}(m_r \cap m_c)$$

- Compute average closeness (c) of paired elements:

$$c = 1 - \frac{d}{L}$$

- Compute fraction of paired elements:

$$f(a, b) = \frac{N^*}{N''}$$

N^* is the least amount of unique elements and N'' is the largest amount of unique elements, in either vector a' or b' .

- Compute similarity (S):

$$S(a, b) = c \cdot f$$

This measure is used because it gives a single value that encompasses the hit and misses given by the fraction of paired elements and closeness of those elements. In the context of this study these elements are the time locations of segmentation boundaries. In this way it is not necessary to specify a vicinity of annotated boundaries in which a computed boundary has to be considered a match, as is the case of the method used by the MIREX structural segmentation evaluation (MIREX Structural Segmentation, 2016; Turnbull et al., 2007; Levy & Sandler, 2008). The MIREX 2016 structural segmentation evaluation considered a vicinity of 0.5 s. This is problematic as the transition from one gesture to another might take different times at different time-scales. Therefore the vicinity should be adjusted to those transition times. It is not clear how this can be done, so the similarity measure described above avoids the problem. However, it has the disadvantage that a visual comparison of very high values of S (e.g., over 0.8) might not appear to be reasonably similar and a very small difference in S might be visually perceived as a considerably different. This drawback is only a perceptual scaling problem that does not affect the computational effectiveness of the similarity measure. The selection of features (i.e., combinations of free variables) that yield results most similar to the ground truth is an optimization problem in a highly dimensional space. The amount of possible combinations is astronomical and an extensive search (i.e., by brute force) for the highest S value is therefore impractical. To overcome this difficulty, the solution space is explored by brute-force with constraints that reduce the free-variable space. Then the computed boundaries that

III. RESULTS

Data collected so far for the ground truth has been deemed not enough to make the analysis that compares real-time perceived segmentation, non-real-time perceived segmentation and computed audio segmentation. Nonetheless, the available non-real-time grouped annotated boundaries of

gesture recognition. Such a system will be useful for studying relationships between musical sound and bodily movement. Furthermore, a real-time implementation of this system could be integrated into the design of electronic musical instruments, as a high-level feature for mapping movement to sound. Overall, this automated system provides a cost-effective solution as it can take advantage of cheap accelerometer sensors and computing technology.

REFERENCES

- Bee Gees (1977). Stayin' Alive. On *Saturday Night Fever, The Original Movie Soundtrack*. RSO.
- Clayton, M., & Leante, L. (2011). Imagery, melody and gesture in cross-cultural perspective. In A. Gritten & E. King (Eds.), *New Perspectives on Music and Gesture*, 203. Farnham, England: Ashgate.
- Cockos Reaper [Computer software] (2010). Retrieved from <http://www.cockos.com/reaper>
- Foote, J. T., & Cooper, M. L. (2003). Media segmentation using self-similarity decomposition. In *Electronic Imaging 2003* (pp. 167-175). International Society for Optics and Photonics.
- Godoy, R. I., Song, M., Nymoen, K., Haugen, M. R., & Jensenius, A. R. (2016). Exploring Sound-Motion Similarity in Musical Experience. *Journal of New Music Research*, 45(3), 210-222.
- Hartmann, M., Lartillot, O., & Toiviainen, P. (2016). Interaction features for prediction of perceptual segmentation: Effects of musicianship and experimental task. *Journal of New Music Research*, 1-19.
- Kahol, K., Tripathi, P., & Panchanathan, S. (2004). Automated gesture segmentation from dance sequences. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004 Proceedings*. (pp. 883-888). IEEE.
- King, E., & Ginsborg, J. (2011). Gestures and glances: Interactions in ensemble rehearsal. In A. Gritten & E. King (Eds.), *New Perspectives on Music and Gesture*, 177-201. Farnham, England: Ashgate.
- Lara, O. D., & Labrador, M. A. (2013). A survey on human activity recognition using wearable sensors. *IEEE Communications Surveys and Tutorials*, 15(3), 1192-1209.
- Lartillot, O., Toiviainen, P., & Eerola, T. (2008). A matlab toolbox for music information retrieval. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme, & R. Decker (Eds.), *Data analysis, Machine Learning and Applications* (pp. 261-268). Berlin, Heidelberg: Springer.
- Leman, M. (2008). *Embodied Music Cognition and Mediation Technology*. Cambridge, MA: MIT Press.
- Levy, M., & Sandler, M. (2008). Structural segmentation of musical audio by constrained clustering. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2), 318-326.
- Luck, G. (2011). Computational analysis of conductors' temporal gestures. In A. Gritten & E. King (Eds.), *New Perspectives on Music and Gesture* (159). Farnham, England: Ashgate.
- Machado, I. P., Gomes, A. L., Gamboa, H., Paixão, V., & Costa, R. M. (2015). Human activity data discovery from triaxial accelerometer sensor: Non-supervised learning sensitivity to feature extraction parametrization. *Information Processing & Management*, 51(2), 204-214.
- MacRitchie, J., Buck, B., & Bailey, N. J. (2013). Inferring musical structure through bodily gestures. *Musicae Scientiae*, 17(1), 86-108.
- Mendoza, J. I. (2014). *Self-report Measurement of Segmentation, Mimesis and Perceived Emotions in Acoustic Electroacoustic Music*. Master's Thesis. University of Jyväskylä. Retrieved from <http://urn.fi/URN:NBN:fi:juyu-201406192112>
- MIREX Structural Segmentation (2016). Retrieved from http://www.music-ir.org/mirex/wiki/2016:Structural_Segmentation
- Otondo, F. (2008). Ciguri. On *Tutuguri*. Sargasso.
- Olsen, K. N., Dean, R. T., & Leung, Y. (2016). What constitutes a phrase in sound-based music? A mixed-methods investigation of perception and acoustics. *PLoS One*, 11(12): e0167643.
- Petzold, C. (ca. 1725). Minuet from The Anna Magdalena Bach Notebook, Anh. 114.
- Puckette, M. (1997). Pure Data. *International Computer Music Conference*. Thessaloniki, Greece: Michigan Publishing
- Schneider, A. (2010). Music and gestures. In R. I. Godoy & M. Leman (Eds.), *Musical Gestures: Sound, Movement, and Meaning*. London: Routledge.
- Tetarto Hood (2014). Bouzouki Hiphop Instrumental - Rempetila. Retrieved on the 23 August of 2016 from <https://www.youtube.com/watch?v=mMWMS6VqXTg>
- Toiviainen, P., Luck, G., & Thompson, M. R. (2010). Embodied meter: Hierarchical eigenmodes in music-induced movement. *Music Perception*, 28(1), 59-70.
- Trevarthen, C., Delafield-Butt, J., & Schögler, B. (2011). Psychobiology of musical gesture: Innate rhythm, harmony and melody. In A. Gritten & E. King (Eds.), *New Perspectives on Music and Gesture*, 11-43. Farnham, England: Ashgate.
- Turnbull, D., Lanckriet, G. R., Pampalk, E., & Goto, M. (2007, September). A Supervised Approach for Detecting Boundaries in Music Using Difference Features and Boosting. In *ISMIR* (pp. 51-54).
- Wanderley, M. M., Vines, B. W., Middleton, N., McKay, C., & Hatch, W. (2005). The musical significance of clarinetists' ancillary gestures: An exploration of the field. *Journal of New Music Research*, 34(1), 97-113.



III

SEGMENTATION BOUNDARIES IN ACCELEROMETER DATA OF ARM MOTION INDUCED BY MUSIC: ONLINE COMPUTATION AND PERCEPTUAL ASSESSMENT

by

Juan Ignacio Mendoza & Marc Richard Thompson, 2017

Human Technology, 18(3), 250–266

<https://doi.org/10.14254/1795-6889.2022.18-3.4>

SEGMENTATION BOUNDARIES IN ACCELEROMETER DATA OF ARM MOTION INDUCED BY MUSIC: ONLINE COMPUTATION AND PERCEPTUAL ASSESSMENT

Juan Ignacio Mendoza G.
University of Jyväskylä
Department of Music, Art and Culture Studies
Finland

Abstract: *Segmentation is a cognitive process that serves to the understanding of information perceived through the senses. Likewise, the automatic segmentation of data captured by sensors may be used for the identification of patterns. This study is concerned with the segmentation of dancing motion captured by accelerometry and its possible applications such as pattern learning and recognition, or gestural control of devices. To that effect, an automatic segmentation system was formulated and tested. Two participants were asked to ‘dance with one arm’ while their motion was measured by an accelerometer. The performances were recorded in video, and later manually segmented by six annotators. The annotations were used to optimize the automatic segmentation system, maximizing a novel similarity score between computed and annotated segmentations. The computed segmentations with highest similarity to each annotation were then manually assessed by the annotators, resulting in Precision between 0.71 and 0.89, and Recall between 0.82 to 1.*

Keywords: *gestural interface, perceptual evaluation, temporal segmentation, accelerometer, bodily motion, similarity*



INTRODUCTION

The advancement in miniaturization of accelerometers, gyroscopes and magnetometers, has made it possible to develop portable and wearable systems that sense the movement of the human body. This has opened doors for many applications in a vast range of domains. Many such applications require identifying segmentation boundaries within movement; that is, where data changes from one regime to another. Following this, the detected segments can be classified or clustered. Some methods detect segmentation boundaries in the same process that performs classification or clustering. Examples of applications that use these processes include systems for detecting, recognizing and monitoring activities for clinical diagnosis or assisting in sports training (Cornacchia, Ozcan, Zheng, & Velipasalar, 2017).

The focus of the current study was to identify segmentation boundaries within the movements of a dancing person. In a practical application, the detected segmentation boundaries may be used to control playback of sound, music or lighting, for example. The movement of the dancer may be sensed in a number of different ways, but this study focuses on the use of a single triaxial accelerometer. The dancer would hold with one hand a device that has a built-in accelerometer, or would have the device attached to one of their limbs. Data from the accelerometer would be streamed to a machine that computes segmentation boundaries in real time. The output would be the time when a segmentation boundary has occurred, with respect to real time. Then, this information may be used for the control of a separate process (e.g., triggering events) or for machine-learning processes such as clustering or classification of the found segments.

It is desirable that the result of the segmentation system is produced fast enough for near-real-time interaction. Also, it is necessary that the motion segments are meaningful to an observer. In other words, motion segments produced by the system should match the segments perceived by an observer. The meaningfulness of motion segments would additionally facilitate the learning of motion patterns and mappings to audio or visual effects. To that extent, it must be acknowledged firstly, that human perception of bodily movement is highly subjective (Bläsing, 2015; Kahol, Tripathi, & Panchanathan, 2004; Zacks, Kumar, Abrams, & Mehta, 2009) and is hierarchically structured such that short patterns are grouped into larger ones (Bernard, Dobermann, Vögele, Krüger, Kohlhammer, & Fellner, 2017; Dreher, Kulp, Mandery, Wächter, & Asfour, 2017; Krüger, Kragic, Ude, & Geib, 2007; Lin, Karg, & Kulić, 2016). Also, it must be taken into consideration that dance patterns may or may be not repetitive. Thus, the system must be capable of detecting repetitive and non-repetitive patterns, and must allow the user to make adjustments to obtain perceptually meaningful results.

The algorithm described by Foote (2000) for segmentation of digital audio was found to be an appropriate candidate for segmentation of dance movement. This algorithm has subsequently been used for segmentation of video (Foote & Cooper, 2003), and of dance motion based on speed extracted from video (Tardieu et al., 2009). It has also been used to identify boundaries between activities such as walking, jogging and sitting, in single-axis accelerometer data (Rodrigues, Probst, & Gamboa, 2021). While most published implementations are online (i.e., data is processed serially as it is input to the algorithm), Schätti (2007) described an online implementation for segmentation of an audio signal. Also these implementations have been tested on data whose segments span several seconds or

minutes (e.g., sections of a song, walking). Therefore, the current study has focused on the adaptation of an online version of the algorithm to work with a triaxial accelerometer signal, and the assessment of its capability to meet the requirements of the intended application. The contributions of the present study are, first, the application and testing of the segmentation algorithm at a smaller time-scale (i.e., short dancing patterns spanning a few seconds), and a more robust perceptual assessment than those used in previous work. The second contribution is a novel measure to evaluate the similarity between computed and perceived segmentation boundaries.

This report is structured as follows: The remainder of the introduction presents a succinct review of the state-of-the-art methods that most closely meet the requirements stated above, including unsupervised near-real-time detection of segmentation boundaries, boundaries of self-similarity checkerboard patterns, and assessments of effectiveness. In favor of a timely report, a comprehensive comparison of different techniques is out of the scope of this study. Following this, the Methods utilized and the Results so obtained are reported. Finally, the Conclusion provides a summary of the study, including directions for future work.

Unsupervised Near-Real-Time Detection of Segmentation Boundaries

Several algorithms that detect segmentation boundaries and give results in near-real-time have been tested with data from accelerometers. For example, Gharghabi et al. (2019) described a method that evaluates the similarity in shape –but not in statistical properties– between all fixed-length windows within a bigger window, the length of which is specified by the user. A segmentation boundary is recorded where the similarity is minimal. This method assumes that each segment will be composed of at least two instances of a periodic motion.

Another approach is to pose the task as a multivariate change-point detection problem (Endres, Christensen, Omlor, & Giese, 2011; Gong, Medioni, & Zhao, 2014; Krüger et al., 2017; Zhou, De la Torre, & Hodgins, 2012). Essentially, a change-point indicates a difference in statistical properties of the data within a sliding window (Aminikhanghahi & Cook, 2017; Fathy, Barnaghi, & Tafazolli, 2018; Liu, Yamada, Collier, & Sugiyama, 2013; Patterson et al., 2016). The sliding window is a free parameter that adjusts time-scale (i.e., granularity). Depending on the method, other free parameters may need to be adjusted. Zameni et al. (2020) described a method that efficiently identifies segmentation boundaries in signals that can be highly dimensional. This method has initialization parameters, but no parameters that can be used to explicitly adjust time-scale or relevance. The cited systems were tested with various types of data. When the test data had been recorded by triaxial accelerometers, the tests aimed to segment activities that take at least a few seconds to complete. However, segments of dancing motion may range from less than a second to more than a few seconds.

Boundaries of Self-Similarity Checkerboard Patterns

The detection of change-points in motion data can be seen as equivalent to novelty detection, which is the identification of abrupt changes in data by a system, without training of the system (Markou & Singh, 2003). Foote (2000) described a method suitable for finding segmentation boundaries in musical audio signals. This method exploits the characteristic

checkerboard patterns that can be observed in a self-similarity distance matrix of audio features through time, by correlating a checkerboard kernel along the diagonal of the matrix. This results in a novelty score that indicates the rate of change in the data. The peaks of the novelty score indicate change-points that correspond to perceived changes in the music. The granularity of the novelty score is adjusted with the width of the kernel and relevant peaks can be selected over a threshold.

Assessment of Effectiveness

To measure the effectiveness of segmentation algorithms, most published studies have relied at least to some extent on classic measures of precision, recall and accuracy, by comparing human-annotated ground truth boundaries annotated by one or more people with computed boundaries. These measures work well for classification problems in which the options are either “match” or “not a match” between a computed boundary and a ground truth boundary. Dreher et al. note that a computed segmentation boundary being only slightly different to the ground truth should be counted as a match. This is usually solved by establishing a window around each ground truth boundary. A computed point is deemed to be a true positive if it lies within that window. This approach was used in the study by Zameni et al., for example. Dreher et al. proposed a method that involves a window weighted with a normal distribution. However, the problem with this approach is that the window’s width is fixed while there is no certainty that any given width will correspond to the true probability distribution for the occurrence of a boundary, for all boundaries. It is not possible to generalize the temporal length of the transition from one motion to another. In contrast, the evaluation method used by Gharghabi et al. consists of a score that measures the temporal distance between each computed boundary and the closest boundary in the ground truth. All the distances are added and then divided by the total time. However, this score does not penalize extra or missing computed boundaries, which is problematic as there is no certainty that the number of annotated and computed boundaries will always be the same. Lin et al. (2013) describe another approach for evaluation of results, in which all frames in the ground truth segments are labelled and the number of frames in the computed segments corresponding to the ground truth-labels constitute the measure of similarity. This last method might be appropriate for classification of segments but it might be too restrictive for evaluating only the boundaries. This is because boundaries of short false-positive computed segments (e.g., transitions between motions) will break the continuity of parallel labelling resulting in a very high dissimilarity score. Mendoza (2014), and also Mendoza and Thompson (2017), proposed similarity scores that measure the distance between ground truth and computed boundaries as in the method by Gharghabi et al., but also penalize missing or extra computed boundaries.

The Present Study

The following section describes the implementation of Foote’s algorithm for the segmentation of accelerometer data. Then, an experimental assessment is described in which ground truth is used to tune the algorithm’s free parameters using a revised version of the similarity measure by Mendoza and Thompson. In contrast to previous studies, the computed results are not assessed by means of a similarity measure but manually by the same annotators who provided the ground truth.

METHODS

Detection of Segmentation Boundaries

This subsection describes the method for finding temporal segmentation boundaries, focusing on its online implementation and its adaptations to work with accelerometer data. A succinct description of the original offline version is provided. For details of the algorithm in general and the offline version, the reader is directed to the original source (Foote, 2000).

The offline version of the algorithm has as input data stored in memory, which has been sampled at regular intervals. This data is represented by the matrix $M \in \mathbb{R}$, so that $M_{1:m} = [F_1, F_2 \dots F_m]^T$. Each frame F at time-index $t \in \{1 \dots m\}$ contains data for each sample. A distance matrix $D \in \mathbb{R}^{m \times m}$ is computed for all data in M . D is a self-similarity matrix. A two-dimension checkerboard kernel is produced by the Kronecker product of checkerboard matrix C and only-ones matrix J of width n as follows:

$$C = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \quad (1)$$

$$K = C \otimes J \quad (2)$$

K is then tapered by multiplying it element-wise with a two-dimensional Gaussian (i.e., a normal distribution). Next, K is correlated along the diagonal of D . The result of this correlation is novelty score N , the peaks of which indicate the locations of segmentation boundaries. The peaks can be selected by a threshold θ , discarding peaks of lower values that might be irrelevant. Hence, n and θ are free parameters for granularity and peak relevance, respectively.

The online version of the algorithm consists in M being a stream of data frames $F_t = (f_x^t, f_y^t, f_z^t)$, sampled at regular intervals, containing the three axes of the accelerometer. A window of n frames is stored in a buffer W_{nov} (Figure 1a). For each incoming frame, the last frame in the buffer is removed while the current frame is stacked in the first position, and distance matrix $D \in \mathbb{R}^{n \times n}$ is computed for W_{nov} (Figure 1b). In this study, Euclidean distance was used. Then, the inner product between Gaussian-tapered checkerboard kernel K and D is computed, resulting in a new point in novelty score N (Figure 1c).

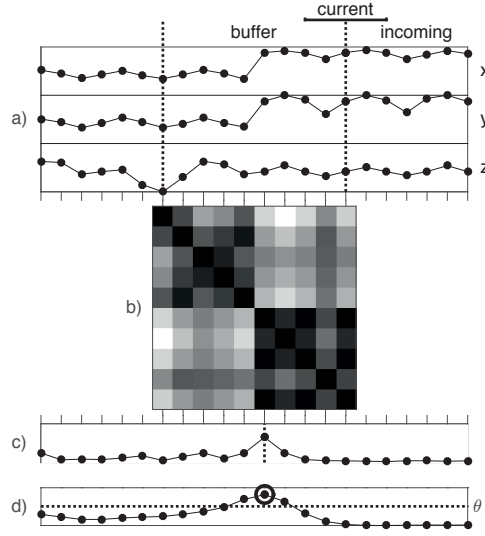


Figure 1. Online detection of temporal segmentation boundaries. Horizontal axes represent time. (a) is triaxial accelerometer data. (b) is self-similarity matrix D of data in the buffer W_{nov} , where lighter shades represent more distance. (c) is novelty score N , where the vertical dotted line indicates the current result. (d) is the smoothed novelty score N' , where θ is a threshold and the point in a circle is the selected peak indicating a segmentation boundary. Note that this visualization shows N and N' aligned in time, but in practice there will be a lag due to the low-pass Gaussian filter and the test for a peak.

When tested, N contained many irrelevant peaks. Therefore a low-pass filter was applied. The filter used in this study was a one-dimension Gaussian kernel with minima zero and unit area to prevent artefacts at borders and to preserve scale, respectively. This filter is computed upon a second buffer W_{filt} having the size of the one-dimensional Gaussian n_{filt} , resulting in a smooth novelty score N' . Finally, if the current novelty score value is a peak over threshold θ , it is considered a segmentation boundary (Figure 1d). Identification of peaks requires another buffer of only three samples to test a local maximum. Hence, the identification of a novelty peak has lag

$$l = \frac{n + n_{filt}}{2} + 3 \quad (3)$$

with respect to the current incoming frame.

Since self-similarity matrix D is symmetric, it is necessary to compute only half of it, either the upper or lower triangle, without the diagonal. Also there is no need to compute the whole triangle for each new frame. It is only needed to initialize matrix D with allocation values (e.g., zeros), then compute the distance between the current frame and all the other frames in the buffer. Then, compute the inner product of the upper or lower triangle of D and the corresponding triangle of K . This will output the current novelty value. Then the values within D are shifted, discarding the distances between the oldest frame and the newer ones. This operation reallocates memory indexes, which takes much less computation time than redundant computation of distance.

The time-scale of the segments may be adjusted dynamically with parameters n and θ . This may be accomplished by fixing the ratio between parameters n_{nov} and n_{filt} , so that parameter n modifies the size of buffers W_{nov} and W_{filt} at the same time. When changing n , a new checkerboard kernel may be computed, or a kernel may be selected from many that might have been previously computed and stored in memory. Because of the operations on D and K , the asymptotical computational complexity is $O(n^2)$. However, in practice n may not grow too much to present a concern, as its size would be limited to the intended granularity and may be reduced by reducing the sampling rate.

Accelerometer Data Collection

Two participants, one female and one male, provided motion data to test the segmentation method. This data was collected at the motion-capture laboratory of the department of Music, Art and Culture Studies at the University of Jyväskylä. These participants are referred to as *dancers* to differentiate them from the participants that provided data for the ground truth and perceptual assessment (see subsection “Ground truth annotation”).

In individual sessions, the dancers were asked to “dance with one arm” while holding with the corresponding hand a Nintendo Wii-remote controller. They were asked to move to the music, without displacement of the body, and always facing one corner of the room. While these conditions may not generalize to all dancing scenarios, they provided a clear view of the moving arm to a video camera. Video recordings were later used for manual annotation. The elimination of the random variable of orientation facilitated the annotation task. Also it simplified the analysis, thus making it possible to focus on first solving the segmentation problem in a simple condition before embarking on a more complex scenario. The dancers were told that other than these constraints, they could move as they wanted.

Three musical stimuli were presented through loudspeakers:

1. “Minuet” (Petzold, ca. 1725) MIDI rendition with piano sound, from beginning to end (104 bars, duration 92.5 s.) with no fade-in or fade-out. It has a ternary metre (3/4, or three beats per bar). Both participants declared to know this piece.
2. “Ciguri” (Otondo, 2008) from 56 to 183.7 s. (duration 122.7 s.) with fade-out the last 5 s. This is an electroacoustic piece that has no perceivable beat and therefore no metre. Both participants declared to not know this piece.
3. “Stayin’ Alive” (Gibb, Gibb, & Gibb, 1977) from the beginning to 108.5 s. with fade-out the last 2.3 s. It has a binary metre (4/4, or four beats per bar). Both participants declared to know this piece.

The number of performances amounted to six. This was deemed enough for this study as they provided variety: musical genre, metre, familiarity and the gender of the participants. These characteristics would permit to observe to some extent their effect on the test. Furthermore, later these performances were used for the task described in the next section (“Ground truth annotation”). More performances would have extended the annotation task implying the risk of abandonment or fatigue, the latter reducing the reliability of results.

Stimuli were presented in the order listed above and each stimulus was presented twice. During the first presentation, participants were asked to move freely within an area of about 4m², to familiarize themselves with the stimulus. For the second presentation, participants were asked to dance with one arm as described above. Data of the performances were recorded as follows:

- *Accelerometer*: The Nintendo Wii-remote has a triaxial accelerometer, which transmits data in real-time via Bluetooth. This stream was received and recorded by a computer at a rate of 100 Hz, using custom-made software.
- *Video*: A digital video camera recorded video showing the participant's whole body against a white wall. Both participants used their right arm, and were recorded so the image clearly showed the moving arm.
- *Audio*: Digital audio was captured by the microphone of the video camera and by a microphone hanging from the ceiling. The latter was recorded to a digital audio workstation synchronized with the recording of accelerometer data. These signals were subsequently used to synchronize video and accelerometer data.

Ground Truth Annotation

Six participants (3 male, 3 female) were recruited to identify segmentation boundaries in the one-arm-dancing videos. None of them had participated in the data collection described in the previous section. Their ages ranged from 26 to 34 years, with a median age of 27. All were non-Finnish international students at the University of Jyväskylä. All had completed at least an introductory course in music psychology, covering an introduction to perception and segmentation. These participants are referred to as *annotators*, to differentiate them from the *dancers* who performed the one-arm dance (see subsection "Accelerometer data collection").

Each annotator, in an individual session, was asked to watch the videos and identify segmentation boundaries in two conditions. In the first condition, the videos with audio were presented by a computer running custom-made software. The annotators were instructed to press a key when a boundary was identified, in real time. The time of the key relative to the video was recorded by the computer. They had only one chance to perform the task. It was thought that the music in the video may influence the responses as auditory cues, such as pitch or rhythm, and could be used to judge the existence of a boundary. For the second condition, the videos without audio were presented by the computer running a digital audio editor software. In this condition, participants could freely play the video, pause, scroll forward and backwards, place markers and adjust the location of the markers until they were satisfied. In this condition, the annotators did not have a limit of time for the task and the annotation was based solely on visual information.

The following were the instructions to the annotators, common for both conditions:

"You will be presented with six videos, each lasting around two minutes. Each video shows a person 'dancing' with an arm. When doing this, the person does distinct patterns with the arm. A pattern is composed by one distinct movement or several repetitions of the same movement. When the video is playing press the space bar to indicate a change in pattern. Focus in the movement of the arm holding the white device (it is a sensor)."

The two annotation conditions represented different approaches for perceived segmentation. To assess their suitability, the annotators were interviewed after completing the tasks. They were asked to verbally express what they considered to be difficult or easy about the tasks. All participants mentioned that, in the real-time annotation task, their responses might have been influenced by the music and they were less precise than in the non-real-time condition. The reasons mentioned for this included that in the real-time condition the responses might have been anticipated as an effect of the music. Also, it was mentioned that, in the real-time task, it was more difficult to press the button exactly at the intended time, thus preventing a response to be recorded accurately or in some cases at all. All participants expressed that the non-real time condition allowed for more precise responses, as they could take time to revise them. Because of this, the data relating to real-time audiovisual annotation was deemed inappropriate for use as a ground truth. Thus, non-real-time visual annotation was chosen as ground truth for perceived segmentation boundaries.

Optimization using similarity based on distance and rate of paired elements

A grid search was performed to maximize the similarity between annotated (ground truth) and computed segmentation boundaries, by modification of parameters n and θ . This search was performed independently for each accelerometer recording and their corresponding annotations, mimicking the adjustment that might be achieved manually by an end-user or automatically by a machine-learning procedure. Similarity was evaluated by distance and penalization of extra or missing boundaries, improving previous work (Mendoza, 2014; Mendoza & Thompson, 2017).

Consider vectors a and b containing the time indexes of annotated and computed segmentation boundaries, respectively. L is the length, in samples, of the corresponding recorded data, from the start to the end of the musical stimulus. n_a and n_b are the number of boundaries, or length, of a and b respectively. In any case $n_a \geq n_b$ or vice-versa. Each element in a is paired to the closest element in b , so that a' and b' are vectors containing only the paired elements and have equal lengths n_p (equivalent to the shortest between n_a and n_b). Then, the following measures are computed:

Closeness:

$$c = 1 - \frac{1}{L} \sum_{i=1}^{n_p} |a'_i - b'_i| \quad (4)$$

Rate of paired elements:

$$p = \frac{2n_p}{n_a + n_b} \quad (5)$$

Similarity:

$$S = c \cdot p, 0 \leq S \leq 1 \quad (6)$$

The distance between paired boundaries is the absolute time difference, as shown in equation 4. Note that two boundaries of either sequence (a or b) may be paired with a single boundary in the other sequence if their distances are equal. Also, if n_a and n_b are not equal

and there are no equidistant boundaries to compensate for that inequality, then some boundaries will be not paired and this will be penalized by the rate of paired elements (equation 5). A Monte Carlo simulation was computed with pseudo-random a and b , for $L = 1000$, with n_a and n_b in the range $\{1 \dots L - 1\}$, and 10^4 iterations. The distribution for the resulting S values has an upper p -value of 0.05 at $S = 0.66$.

Perceptual Assessment

The perceptual assessment was made by the same annotators that provided the ground truths. For each annotator, the annotated and computed boundaries with highest similarity were selected. This means that the assessment is for the 'best case scenario'. For each of these sequences of boundaries a video was produced embedding a scrolling timeline with consecutive numbers for boundaries into the corresponding video that was annotated (Figure 2).

Three videos were produced for each annotator. One had markers for their original annotation, to measure the extent of agreement they would have with the annotation they had previously made. A second video had markers for the computed boundaries. A third video had a confounding sequence of boundaries produced by placing a marker in the middle of the segments bounded by the average point for each pair of paired annotated and computed boundaries. The videos with confounding boundaries were intended to reduce the chance of annotators realizing that one of the sequences was their own annotation, and the responses to those videos were not analyzed.

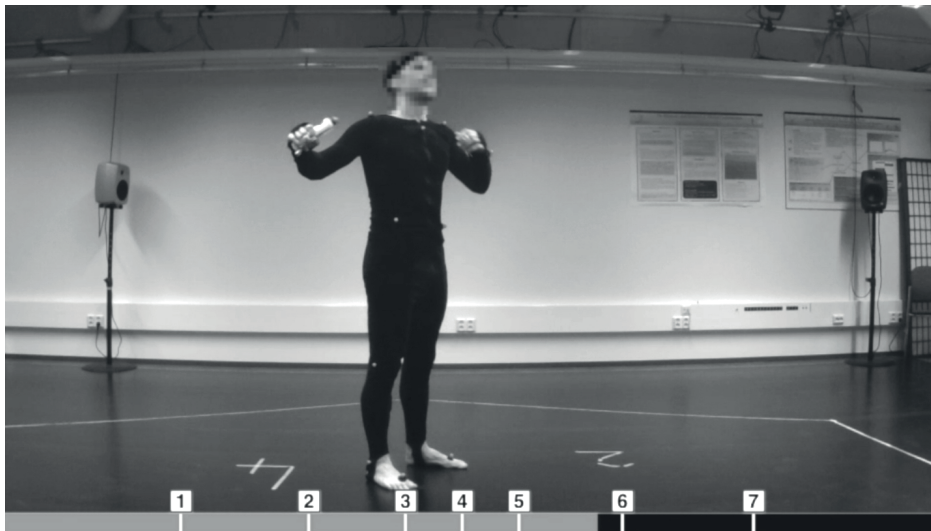


Figure 2. Example frame of a video shown to an annotator for perceptual assessment. The same video without the numbered markers had been used for annotation.

The videos contained no audio, as the annotations used in the computation of boundaries corresponded to video without audio. Each video was embedded in a webpage and had on-screen controls that could be activated with a pointing device (e.g., mouse, trackpad) to play, stop, scroll forward and backwards. The pages were presented in random order by an automatic system that also recorded responses. Each page consisted of instructions, the video and a list of numbered items, one for each marker. Each item in the list had two buttons that could be selected by clicking on them. One button was to answer “yes, there is a change in pattern” and was recorded as a *confirmed boundary*. The other button was to answer “no, there is no change in pattern” and was recorded as a *rejected boundary*. This assessment is used in replacement of the paradigm used in previous studies that considered a computed boundary to be correct if it is within a window around a ground truth boundary. It has the advantage of not needing to specify a fixed window.

The definition of the task was identical to the one given for the annotation task. One distinct questionnaire was produced for each annotator with the corresponding videos. This questionnaire did not reveal how the segmentation sequences were produced. After completing each page all responses were recorded and options were shown to immediately continue to the next page or to continue later. The annotators were asked to complete the questionnaire in their own space and time, using their own computers and to take as much time as they needed.

The decision to assess the best-case-scenario boundaries was made after testing the questionnaire. This test was done with different participants who would take up to 50 minutes to complete a questionnaire with three videos. It was decided that the questionnaire should not exceed three videos, to prevent fatigue and abandonment.

The data obtained from the questionnaires was processed to obtain the following relevance measures:

$$Precision(computed) = \frac{n_{cb}}{n_b} \quad (7)$$

$$Recall(computed) = \frac{n_{cb}}{n_{cb} + n_{ca} - n_p} \quad (8)$$

$$Precision(annotated) = \frac{n_{ca}}{n_a} \quad (9)$$

where n_{cb} is the number of confirmed computed boundaries (true positives), n_b is the number of computed boundaries (true and false positives), n_{ca} is the number of confirmed annotated boundaries, n_p is paired annotated and computed ($n_{ca} - n_p$ is false negatives), and n_a is the number of annotated boundaries (true and false positives). $Precision(annotated)$ may be considered as an indication of the assessment’s reliability. It is not possible to obtain $Recall(annotated)$ as false negatives would require the possibility of adding new boundaries, which was not part of the assessment task.

RESULTS AND DISCUSSION

Computation of the grid search was performed with the recorded accelerometer data downsampled to 25 Hz. The standard deviation σ for the two-dimensional Gaussian that tapers K and the one-dimensional Gaussian smoothing filter for N' were set to $\sigma = n/5$. The length of the one-dimensional Gaussian was set to n ; that is, to the width of K and D . The standard deviation of both Gaussians was searched within $\sigma = \{0.5, 0.6, \dots, 2\}$ seconds. Since recorded accelerometer data was used, computation was performed in non-real-time. Therefore, the filtered novelty score was rescaled to $0 \geq N \geq 1$ and the threshold for peak selection was searched within $\theta = \{0, 0.1, \dots, 0.5\}$. For real-time computation, these values would yield a lag time of $l = \{0.22, 0.24, \dots, 0.52\}$ seconds. Note that lag time does not consider computation time, which depends on the specific computing device used.

The highest lag time among the results is 0.5s, for the segmentation corresponding to Annotator 2, of Dancer 1, to "Minuet". The median lag time was 0.35s. Considering this time scale, this system is not suitable for any practical application that requires immediate perceptual real-time response (i.e., up to about 10 to 50 milliseconds). However, this lag time is suitable for applications in which the occurrence of a segmentation boundary is not to be acted upon immediately. For example, this delayed response may be mapped to a procedure that changes the stimulus music in such a way that it prompts the dancer to change the motion pattern, thus creating a feedback loop. Another use of this delayed response is to record the segments' times, then compute statistics (e.g., mean, standard deviation) and use those for a larger time-scale control of music, lights or other actionable medium. Furthermore, the segmentation result may be used to produce a near-real-time visual or sonic display that may be useful in clinical applications and research in biomechanics, for example.

Tables 1 and 2, respectively, show values for maximum distance d and similarity (S) obtained in the grid search, where $d = |a' - b'|$. The distance is expressed in seconds. The minimum similarity value ($S = 0.56$) has a p -value of 0.39, while the minimum mean similarity value ($S = 0.62$) has a p -value of 0.17. These minimum values represent the worst performance of the automatic segmentation. The greatest mean S values were found for the musical stimuli "Minuet" and "Stayin' Alive", which both have a clear beat and were familiar to the dancers. Conversely, similarity is lower for "Ciguri", which is a piece that has no clear beat and was not familiar to the dancers. This suggests that the effectiveness of the method may be directly related to both or either of these conditions: the presence of a clear beat, and the familiarity the dancers might have with the musical stimulus. Also the table shows that most maxima d seem too large to indicate corresponding paired boundaries. Although this may be considered a limitation of the method, it is still possible that the highly distant computed boundaries are confirmed in the perceptual assessment.

Table 1. Maximum Distance (d) in seconds, between Annotated and Computed Boundaries.

Annotator	Dancer 1			Dancer 2		
	Minuet	Ciguri	Stayin' Alive	Minuet	Ciguri	Stayin' Alive
1	3.80	2.52	2.62	2.33	6.07	2.11
2	4.11	3.59	1.69	5.67	6.74	2.22
3	4.53	7.87	1.96	3.60	5.31	2.84
4	4.44	3.10	3.74	1.15	5.78	3.29
5	3.82	6.31	2.79	2.13	2.27	0.73
6	1.59	2.86	2.72	2.47	1.90	1.56
mean	3.71	4.38	2.52	2.89	4.68	2.13

Table 2. Similarity (S) between Annotated and Computed Boundaries.

Annotator	Dancer 1			Dancer 2		
	Minuet	Ciguri	Stayin' Alive	Minuet	Ciguri	Stayin' Alive
1	0.64	0.66*	0.74*	0.71*	0.75*	0.83*
2	0.76*	0.63	0.68*	0.82*	0.63	0.80*
3	0.71*	0.60	0.68*	0.71*	0.68*	0.91*
4	0.61	0.57	0.74*	0.82*	0.67*	0.74*
5	0.66*	0.68*	0.73*	0.64	0.63	0.71*
6	0.56	0.60	0.70*	0.64	0.61	0.74*
mean	0.66*	0.62	0.71*	0.72*	0.66*	0.79*

* $p \leq 0.05$ (not adjusted for multiple comparisons)

Table 3 contains relevance values for the case of maximum similarity for each annotator. The corresponding sequences of annotated and computed boundaries are visualized in Figure 3. The fifth and sixth boundaries of Annotation 2 seem to be too far for any of them to correspond to the fifth computed boundary. However, this boundary was confirmed in the perceptual assessment. It is not possible to conclude whether this boundary corresponds to any of the annotated boundaries, or if it is a new boundary that was unseen at the annotation task (i.e., serendipity effect) or if it was a mistake made by the annotator in the assessment task.

Table 3. Perceptual Assessment of Annotated and Computed Segmentation with Highest Similarity (S) for each Annotator.

Annotator	Stimulus	Dancer	S	<i>Precision</i> (computed)	<i>Recall</i> (computed)	<i>Precision</i> (annotated)
1	Stayin' Alive	2	0.83	0.75	0.82	1
2	Minuet	2	0.82	0.86	0.86	0.75
3	Stayin' Alive	2	0.91	0.89	1	0.67
4	Minuet	2	0.82	0.71	1	0.71
5	Stayin' Alive	1	0.73	0.71	0.88	0.95
6	Stayin' Alive	2	0.74	0.80	0.92	0.86
mean			0.81	0.79	0.91	0.82

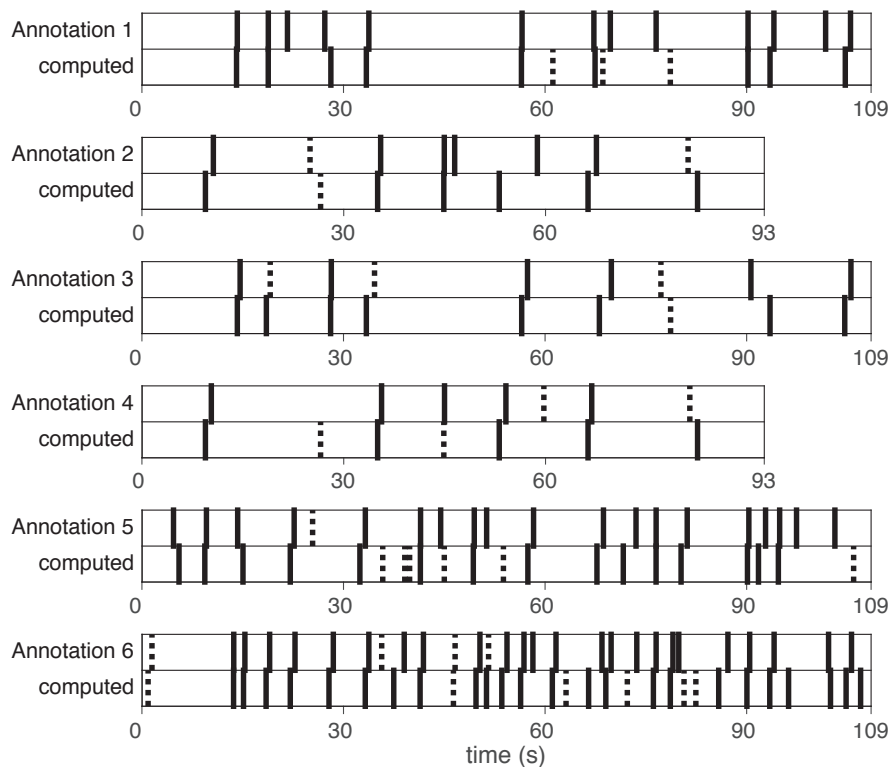


Figure 3. Annotated and closest computed segmentation boundaries for each annotator, corresponding to Table 3. Full lines indicate confirmed and dotted lines indicate rejected.

Another problem is that most annotators rejected boundaries that they had previously annotated, as shown by measure $Precision(annotated)$. While these values are fairly high, some assessment responses look counter-intuitive. For instance, the third boundary of Annotation 4 is evidently close enough to its computed counterpart to be considered an exact match. However, the computed boundary was rejected as shown by the dotted line. Another example that may cast doubt on the perceptual task is the second and fourth boundaries of Annotation 3. These were rejected but their computed counterparts, even being noticeably very near, were confirmed. These odd assessment responses are not the norm, but they raise questions about the reliability of the perceptual tasks.

The two aforementioned assessment problems may be solved by a revised questionnaire including a task that shows both annotated and computed boundaries in the same time line, thus making evident to the annotator the distance between them. In addition, the task would require the annotator to explicitly indicate the corresponding annotated boundary for each computed boundary and vice-versa, if such correspondence exists. Despite the drawbacks of the segmentation and assessment methods, the best-case scenario reveals very high Precision and Recall values. This is relevant as the best-case scenario is akin to the best possible re-tuning that a user could make in a practical application scenario.

A further limitation of this study is that the annotation and assessment tasks were done at different times. This might explain the odd responses mentioned above. A possible solution would be to integrate annotation, automatic segmentation, optimization, and assessment, into one procedure.

CONCLUSIONS

This article has presented an adaptation, testing and perceptual assessment of a method to compute segmentation boundaries in accelerometer data. The method is based on an algorithm widely used for segmentation of digital audio (Foote, 2000). Experimental testing of the adapted and extended algorithm used accelerometer data of subjects moving their arm to music, as a simplistic form of dance, from which segmentation boundaries were computed. The fine tuning of the algorithm's parameters was based on annotators' responses, using a novel measure of distance of paired elements between computed and annotated boundaries, combined with penalization for missing or extra boundaries. Perceptual assessment, consisting of rejection or confirmation of computed boundaries, resulted in fairly high values for measures of relevance *Precision* and *Recall*. The segmentation procedure requires a context-dependent minimum time to produce a response, which in this study was maximum about half a second. This is suitable for systems that do not require an immediate response.

Future work on the perceptual assessment of segmentation boundaries should include a task to pair computed and annotated boundaries, in combination with the task to reject or confirm boundaries. It would also be useful to evaluate more and different input data modalities for computing segmentation, as well as manually or automatically learned features that might improve effectiveness. Furthermore, after the segmentation and assessment methods presented in this article are improved as mentioned, they should be incrementally tested on more complex motion and more realistic conditions. Possible next steps might be to attempt segmentation of dancing motion using both arms, legs, the full body, allow free displacement, different musical stimuli and so forth.

IMPLICATIONS FOR RESEARCH AND APPLICATION

This study has developed and tested a system to produce near-real-time segmentation sequences of accelerometer data. This system may be useful for proposing segmentation to a final user, making the process faster than manually. For example, the system could produce several sequences at different granularity levels, out of which the user selects the most appropriate. Likewise, a matrix of multigranular segmentation sequences may be used without any further screening by the user. As such, the system may see a number of practical applications, for example the inspection of data (e.g., identification of daily activity events in data recorded by a wearable accelerometer) or mapping the segmentation results to actionable processes (e.g., gestural control of music, lights, etc.). An important contribution of this study is the formulation of a novel non-parametric similarity measure based on distance and rate of paired elements. Although the measure was developed to assess similarity of segmentation sequences, it may be used to assess the similarity between any pair of sequences of ordered numbers.

REFERENCES

- Aminikhanghahi, S., & Cook, D. J. (2017). A survey of methods for time series change point detection. *Knowledge and information systems*, 51(2), 339-367. <https://doi.org/10.1007/s10115-016-0987-z>
- Bernard, J., Dobermann, E., Vögele, A., Krüger, B., Kohlhammer, J., & Fellner, D. (2017). Visual-interactive semi-supervised labeling of human motion capture data. *Electronic Imaging*, 2017(1), 34-45. <https://doi.org/10.2352/ISSN.2470-1173.2017.1.VDA-387>
- Bläsing, B.E. (2015). Segmentation of dance movement: effects of expertise, visual familiarity, motor experience and music. *Frontiers in psychology* 5, 1500. <https://doi.org/10.3389/fpsyg.2014.01500>
- Cornacchia, M., Ozcan, K., Zheng, Y., & Velipasalar, S. (2017). A survey on activity detection and classification using wearable sensors. *IEEE Sensors Journal* 17(2), 386–403. <http://doi.org/10.1109/JSEN.2016.2628346>
- Dreher, C. R., Kulp, N., Mandery, C., Wächter, M., & Asfour, T. (2017). A framework for evaluating motion segmentation algorithms. In *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)* (pp. 83-90). IEEE. <https://doi.org/10.1109/HUMANOIDS.2017.8239541>
- Endres, D., Christensen, A., Omlor, L., & Giese, M.A. (2011). Emulating human observers with bayesian binning: Segmentation of action streams. *ACM Transactions on Applied Perception (TAP)*, 8(3), 1-12. <https://doi.org/10.1145/2010325.2010326>
- Fathy, Y., Barnaghi, P., & Tafazolli, R. (2018). An Online Adaptive Algorithm for Change Detection in Streaming Sensory Data. *IEEE Systems Journal*, 13(3), 2688-2699. <https://doi.org/10.1109/JSYST.2018.2876461>
- Foote, J. (2000). Automatic audio segmentation using a measure of audio novelty. In *2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings.* (Vol. 1, pp. 452-455). IEEE. <https://doi.org/10.1109/ICME.2000.869637>
- Foote, J. T., & Cooper, M. L. (2003). Media segmentation using self-similarity decomposition. In *Storage and Retrieval for Media Databases 2003* (Vol. 5021, pp. 167-175). International Society for Optics and Photonics. <https://doi.org/10.1117/12.476302>
- Gharghabi, S., Yeh, C.C.M., Ding, Y., Ding, W., Hibbing, P., LaMunion, S., Kaplan, A., Crouter, S.E., & Keogh, E. (2019). Domain agnostic online semantic segmentation for multi-dimensional time series. *Data Mining and Knowledge Discovery*, 33(1), 96–130. <https://doi.org/10.1007/s10618-018-0589-3>
- Gibb, B., Gibb, R., & Gibb, M. (1977). Stayin' alive. In *Saturday Night Fever, The Original Motion Picture Soundtrack*. Germany: RSO.
- Gong, D., Medioni, G., & Zhao, X. (2014). Structured time series analysis for human action segmentation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7), 1414–1427. <https://doi.org/10.1109/TPAMI.2013.244>
- Kahol, K., Tripathi, P., & Panchanathan, S. (2004). Automated gesture segmentation from dance sequences. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings.* (pp. 883–888). IEEE. <https://doi.org/10.1109/AFGR.2004.1301645>
- Krüger, B., Vögele, A., Willig, T., Yao, A., Klein, R., & Weber, A. (2016). Efficient unsupervised temporal segmentation of motion data. *IEEE Transactions on Multimedia*, 19(4), 797-812. <https://doi.org/10.1109/TMM.2016.2635030>
- Krüger, V., Kragic, D., Ude, A., & Geib, C. (2007). The meaning of action: A review on action recognition and mapping. *Advanced robotics*, 21(13), 1473-1501. <https://doi.org/10.1109/TMM.2016.2635030>
- Lin, J.F.S., Karg, M., & Kulić, D. (2016). Movement primitive segmentation for human motion modeling: A framework for analysis. *IEEE Transactions on Human-Machine Systems* 46(3), 325–339. <https://doi.org/10.1109/THMS.2015.2493536>
- Liu, S., Yamada, M., Collier, N., & Sugiyama, M. (2013). Change-point detection in time-series data by relative density-ratio estimation. *Neural Networks*, 43, 72-83. <https://doi.org/10.1016/j.neunet.2013.01.012>

- Markou, M., & Singh, S. (2003). Novelty detection: a review—part 1: statistical approaches. *Signal processing*, 83(12), 2481-2497. <https://doi.org/10.1016/j.sigpro.2003.07.018>
- Mendoza, J.I. (2014). Self-report measurement of segmentation, mimesis and perceived emotions in acousmatic electroacoustic music. Master's thesis. University of Jyväskylä. <http://urn.fi/URN:NBN:fi:jyu-201406192112>
- Mendoza, J. I., & Thompson, M. (2017). Modelling Perceived Segmentation of Bodily Gestures Induced by Music. In *ESCOM 2017: Conference proceedings of the 25th Anniversary Edition of the European Society for the Cognitive Sciences of Music (ESCOM)*. Ghent University. <http://urn.fi/URN:NBN:fi:jyu-201711024121>
- Otondo, F. (2008). Ciguri. In *Tutuguri*. Sargasso.
- Patterson, T., Khan, N., McClean, S., Nugent, C., Zhang, S., Cleland, I., & Ni, Q. (2016). Sensor-based change detection for timely solicitation of user engagement. *IEEE Transactions on Mobile Computing*, 16(10), 2889-2900. <https://doi.org/10.1109/TMC.2016.2640959>
- Petzdold, C. (ca. 1725). Minuet in G major. *The Anna Magdalena Bach Notebook*, Anh. 114.
- Rodrigues, J., Probst, P., & Gamboa, H. (2021). TSSummarize: A Visual Strategy to Summarize Biosignals. In *2021 Seventh International conference on Bio Signals, Images, and Instrumentation (ICBSII)* (pp. 1-6). IEEE. <https://doi.org/10.1109/ICBSII51839.2021.9445154>
- Schätti, G. (2007). Real-Time Audio Feature Analysis for Decklight3. <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.85.7916&rep=rep1&type=pdf>
- Tardieu, D., Chessini, R., Dubois, J., Dupont, S., Hidot, S., Mazzarino, B., ... & Visentin, A. (2009). Video Navigation Tool: Application to browsing a database of dancers' performances. *on Multimodal Interfaces eNTERFACE '09*, 35. <http://citeseerx.ist.psu.edu/viewdoc/download?jsessionid=0249E27EDBD8D12E8FF58DE4F9ABC18A?doi=10.1.1.159.3151&rep=rep1&type=pdf>
- Zacks, J. M., Kumar, S., Abrams, R. A., & Mehta, R. (2009). Using movement and intentions to understand human activity. *Cognition*, 112(2), 201-216. <https://doi.org/10.1016/j.cognition.2009.03.007>
- Zameni, M., Sadri, A., Ghafoori, Z., Moshtaghi, M., Salim, F. D., Leckie, C., & Ramamohanarao, K. (2020). Unsupervised online change point detection in high-dimensional time series. *Knowledge and Information Systems*, 62(2), 719-750. <https://doi.org/10.1007/s10115-019-01366-x>
- Zhou, F., De la Torre, F., & Hodgins, J. K. (2012). Hierarchical aligned cluster analysis for temporal clustering of human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3), 582-596. <https://doi.org/10.1109/TPAMI.2012.137>

Authors' Note

This study was partially funded by the Finnish Foundation for Technology Promotion (Tekniikan edistämissäätiö). All correspondence should be addressed to Juan Ignacio Mendoza, at the University of Jyväskylä, email: juigmend@student.jyu.fi



IV

RELATIONSHIPS BETWEEN AUDIO AND MOVEMENT FEATURES, AND PERCEIVED EMOTIONS IN MUSICAL PERFORMANCE

by

Marc Richard Thompson, Juan Ignacio Mendoza,
Geoff Luck, & Jonna Katariina Vuoskoski, 2023

Music and Science, 6

<https://doi.org/10.1177/20592043231177871>

Relationships Between Audio and Movement Features, and Perceived Emotions in Musical Performance

Marc R. Thompson^{1,2} , Juan Ignacio Mendoza², Geoff Luck^{1,2}
and Jonna K. Vuoskoski³

Abstract

A core aspect of musical performance is communicating emotional and expressive intentions to the audience. Recognition of the musician's intentions is constructed from a combination of visual and auditory performance cues, as well as compositional features. The current study attempted to quantify these contributions by measuring relationships between ratings of perceived emotion, and motion and auditory performance features. A pianist and violinist with advanced degrees in music performance individually performed four short western tonal pieces. The musicians were tasked with performing the pieces while invoking different expressive intentions: sad, happy, angry, and as a control, deadpan. To examine how different expressive intentions influenced performance behavior, the musicians' body movements were tracked using optical motion capture and rendered into point-light animations. Participants rated perceived emotions (happiness, sadness, tenderness, anger) in audio-only, video-only, and audiovisual rating conditions. We first explored how compositional aspects of the music and performers' expressive intentions contributed to ratings across the three viewing conditions. Through a series of analyses of variance, we found that participants successfully decoded the performers' expressive intentions based on visual information alone and auditory information alone. In the rating conditions in which audio was present, compositional aspects had a stronger effect on participant ratings than performers' expressive intentions. Next, we quantified relationships between the ratings and both motion and auditory performance features. Of the features investigated, musical mode had the greatest impact on ratings. Additionally, perceived emotion ratings were more consistent among responders in conditions with audio than without. These results suggest that, in music performance, auditory information is conceptualized by most responders in a similar way, while visual information might be open to a variety of interpretations.

Keywords

Motion capture, music, perception, emotions, performance, embodiment

Submission date: 25 August 2022; Acceptance date: 9 May 2023

A core aspect of musical performance is communicating emotional and expressive intentions to the audience. Performance research of the last quarter-century has stressed that musicians not only communicate intentions aurally but also visually (Behne & Wöllner, 2011; Davidson, 1993; Dahl & Friberg, 2007). The visual channel presents the audience with bodily gestures, which observers use to identify how expressive a performance is, or which emotions the musician intends to express. Previous work has found that kinematic cues from motion captured performances play a significant role in communicating different levels of musical expressivity (Vuoskoski

¹ Centre of Excellence in Music, Mind, Body and Brain, Department of Music, Art & Culture Studies, University of Jyväskylä, Jyväskylä, Finland

² Department of Music, Art & Culture Studies, University of Jyväskylä, Jyväskylä, Finland

³ RITMO Center for Interdisciplinary Studies in Rhythm, Time and Motion, Department of Musicology & Department of Psychology, University of Oslo, Oslo, Norway

Corresponding author:

Marc R. Thompson, Centre of Excellence in Music, Mind, Body and Brain, Department of Music, Art & Culture Studies, University of Jyväskylä, Jyväskylä, Finland.

Email: marc.thompson@jyu.fi



et al., 2014; Vuoskoski et al., 2016). In the current study, we extended this work by asking participants to rate perceived discrete emotions (e.g., happiness, sadness, anger, and tenderness) in piano and violin performances, and further explored the relative contributions of compositional aspects and performers' expressive intentions across different presentation conditions. Additionally, we used regression to quantify the relationships between participants' ratings of emotions, and movement and audio cues derived from the performances.

Musical expression has been defined by nuances in timing, intensity, timbre, and pitch that give a musical performance its unique character and distinguish it from other renditions of the same piece (Palmer, 1997). From an audience's perspective, these creative alterations might act as acoustic cues to identify musicians' intentions and the emotional qualities of music. In western classical music, musicians rarely alter notated pitch and duration. Rather, expressivity is borne from emphasizing ambiguous aspects of the compositional structure such as micro-timing and dynamics (Clarke, 2005). These creative choices enable musicians to communicate ideas and intentions to audiences, who, in turn, make aesthetic judgements as to whether those choices are stylistically appropriate or successful (Akkermans et al., 2019; Gabriellson & Juslin, 1996).

The study of musical expression has extended to encompass psychological and biological movement aspects (Juslin, 2003). Early work by Davidson (1993) introduced a paradigm in which musicians perform various renditions of the same piece of music while employing different levels of musical expressivity, and observers evaluate the visual impact using ratings of perceived expression. This study influenced a broad area of music and movement research, including the study of ancillary movements uninformed in the production of sound (Wanderley, 2002) and, more generally, musical gestures. Musical gestures are generally defined as bodily movements or gestures that have meaning. Various theoretical frameworks have been used to explain how gestures evoke musical ideas. Applying Peirce's Theory of Signs, gestures may be indices illustrating a causal relationship between movement and expression (Clarke, 1995). From an ecological perspective, it is argued that musical gestures contain affordances that can be perceived by individuals with specific histories related to the context of music performance. In Gibsonian understanding, affordances contain calls to action (e.g., a chair affords sitting, and music affords dancing; Gibson, 1979). The question of what physical action is being called for in a music-listening situation might be viewed as a limitation of the notion of affordance. However, the enactive and embodied views of perception hold that perception is an active process. From this perspective, the call to action might be related to sensorimotor processing of perceptual input, such as the memory of past experiences with musical expressions evoked through listening to music (see also Shapiro, 2014; Wilson, 2002)

For an enculturated listener, the music's auditory stream is embedded with rich signifiers the listener uses to draw

meaning from performances. Musicians perform expressive gestures regardless of whether the audience sees them. Windsor (2011) has evocatively described this as the performer leaving 'traces' in the environment, to be picked up by listeners. A person listening to music might imagine physical gestures used to perform the music based on the cues contained within the auditory stream. Whether a feature of movement, such as its kinematics (e.g., speed and acceleration), has a clear relation to the perception of emotional intention in a listening situation, remains an empirical question.

The relationship between the perception of emotions and musical auditory content has been studied using quantitative methods such as regression. Various studies have found that happy emotional content is associated with fast tempo, major mode (Dalla Bella et al., 2001; Juslin, 2000; Peretz et al., 1998), high pitch, increased sound level (Lange & Frieler, 2018), and soft timbre (Juslin & Lindström, 2010), while sadness is generally linked to features inversely associated to happiness. Anger is associated with fast tempo, increased sound level, high-frequency content (Juslin, 2000), and sharp timbre and minor modes (Juslin & Lindström, 2010; Lange & Frieler, 2018). Fear has been found to be related to reduced sound level, staccato articulation, large articulation variability, soft timbre (Juslin, 2000), and minor mode (Juslin & Lindström, 2010; Lange & Frieler, 2018). Tenderness is related to slow tempo, reduced sound level (Lange & Frieler, 2018), low pitch, major mode, soft timbre (Juslin & Lindström, 2010), and reduced changes in dynamics (Eerola et al., 2009). The most significant feature may be mode (i.e., major or minor) (Eerola et al., 2013). Battcock and Schutz (2019) observed that mode predicted the most variance for perceived valence, which is the degree of perceived positiveness or negativeness (e.g., sadness, anger, and fear have negative valence, while happiness and tenderness have positive valence).

Relationships between the perception of emotions and musicians' movements have also been studied. Dahl and Friberg (2007) presented marimba, bassoon, and soprano saxophone performances to participants, who rated them for perceived emotional intentions happiness, sadness, anger, and fear under three conditions (audiovisual, video-only, audio-only). All intended emotions were recognized except fear regardless of condition. Participants also rated movement content, and significant relations were found between emotions and movement features: happiness was associated with slow speed (bassoon), and large amounts of movement (marimba and saxophone); sadness with small amount of movement (marimba), slow speed (all), and smooth fluency of movement (marimba); anger with large amount of movement (marimba), fast speed (marimba), and jerky fluency of movement (all); and fear with small amount of movement (marimba and saxophone). Crucially, facial expression was not presented to the participants. A variant of this paradigm is to use point-light skeleton animations produced with motion-capture data of music performances, which allow observers to view broad

movement patterns without the influence of facial expressions (Burger et al., 2013; Eaves et al., 2020; Vuoskoski et al., 2014).

Previous work by Vuoskoski et al. (2014) and Vuoskoski et al. (2016) has influenced the aims and set-up of the current study. Vuoskoski et al. (2014) reported that visual kinematic performance cues were more important than auditory performance cues when making ratings of perceived expressivity in audiovisual excerpts of piano playing. A novel and balanced manipulation of stimuli, in which motion-capture videos of piano performances were time-warped to fit to non-corresponding audio, enabled the authors to quantify the respective contributions of visual and auditory cues in self-report ratings of perceived musical expressivity. In contrast, Vuoskoski et al. (2016) explored the contributions of visual and auditory cues in self-reports of felt emotions in reaction to musical performances. Again, results highlighted the important role of visual cues for observers' experience of musical performances.

The current study differs from previous similar work in three important ways. First, we added violin performances to contrast previous findings. Our aim was not to produce results generalizable to all instrument groups and situations, but rather to bring attention to differences in the way emotional communication is expressed between two important western instruments. Second, while other studies have looked at emotional engagement or induction of emotions when viewing or listening to musicians' instrumental performances (Camurri et al., 2004; Castellano et al., 2008; Vuoskoski et al., 2014; Vuoskoski et al., 2016), these are different questions than emotion perception or emotion recognition. Instead of focusing on the notion of musical expressivity, participants rated the performances with respect to perceived discrete emotions: tenderness, sadness, happiness, and anger (see Eerola & Vuoskoski, 2011). The music performed on both piano and violin consisted of short pieces that had been validated to express specific emotions (details in the Methods section). Musicians performed each piece while expressing emotional intentions either congruent or incongruent with the validated emotion (e.g., a happy piece performed in a happy, sad, tender, angry, or manner, etc.). Third, the analysis examines relationships between participant ratings and features computed from the motion-capture data, as well as acoustic and musical features extracted from the audio signal and the musical score. Performances were presented to participants in three modes: audiovisual, video-only, and audio-only.

We expected to find cross-modal relationships between auditory and visual features when perceiving musical expressivity. As suggested by Windsor (2011), music presented in one modality can give the perceiver cues as to information from another modality. For instance, in a listening condition, louder sounds might evoke images of faster gestures. Evidence for this proposition was measured by correlating all presentation conditions with both audio and motion features, as well as using a mixture of audio and motion features as predictors in multiple regression models. Finally, regarding

the contribution of composition to ratings of perceived emotion, we hypothesized that musical mode would have a significant effect on perceived emotion, even when the pieces were performed incongruently (e.g., a happy piece in major mode performed angrily).

Methods

Piano and Violin Performances

A violinist and a pianist were recruited to record solo performances of four short musical pieces, each with four different kinds of emotional expressions. Both musicians were advanced conservatory students with more than 15 years of formal training on their respective instruments. The decision to record only two musicians was made to limit the number of performances presented to the participants.

The musicians performed short pieces taken directly or inspired from a database of musical compositions used by Vieillard et al. (2008). Their aim had been to validate musical excerpts that conveyed four intended musical emotions (happiness, sadness, scare, and peacefulness) that could be distinguished on the dimensions of valence and arousal (Russell, 1980), and were composed to match film music clichés (e.g., happiness denoted by major mode and fast tempo; scare denoted by minor mode with dissonances, etc.). These musical pieces were composed for keyboard instruments, but our study required music suitable for piano and violin. In the case of happiness, sadness, and scare, we selected three pieces whose melodic part could be adapted for violin. The corresponding pieces from the database are G03 (happy, in d-major), T01 (sad, in d-minor), and P02 (scary, in d-minor). Because the pieces in the database labeled as peaceful could not be easily transferred to violin (because of being composed by mostly arpeggiated figures and intricate interplay between the treble and bass parts), we created a piece, entitled *Tenderness*, by transposing T01 to D-major (see Supplementary Material Figure S1). The label *Tenderness* was used for consistency with emotion labels used in a wider selection of literature (see Juslin & Laukka, 2003).

The musicians performed each piece with four different types of expression: happy, angry, sad, and (as a control) deadpan, resulting in 16 performances with congruent and incongruent composition and expression pairings. Regarding the use of the term "angry" over "scary", the term "scary" pertains to the response of a listener rather than being an expression on its own, and our aim was to emphasize the emotions conveyed by musicians (where "angry" would be a more suitable choice in this regard). The musicians were instructed to convey each of the target emotions (happiness, anger, and sadness) as best as they could through their performance, although they were asked to avoid extreme variations in tempo between the different performances. For the deadpan performance, the participants were instructed to play without any expression. The reason for not including a tender expression was that

a tender performance was expected to look and sound highly like a sad performance, leading to very limited variability between the conditions. Crucially, the musicians were not given any instruction regarding how they should move while performing, and movement was not discussed during the recording sessions. The posture and movement of the musicians was recorded with a marker-based motion-capture system, the details of which are given in the Results section.

Stimulus Generation

Audio: The pianist played a digital piano and the performances were recorded in MIDI format. For a more realistic piano sound, the MIDI data was imported into GarageBand (Apple, Inc., Cupertino, CA), running on Mac OS X. The “Grand Piano” software instrument with 50% reverb was used to generate high-quality renditions of the performances. The violin was recorded with a microphone and the performances were presented without modification or extra audio editing.

Video: The videos shown to participants were created by rendering the motion-capture data into stick-skeleton animations using MATLAB and the Motion Capture Toolbox (Burger & Toiviainen, 2013). To make the point-light animations clearer to interpret, the number of markers was reduced through a procedure that included both eliminating some of the markers and creating new synthetic markers located at the midpoint between two original markers. This marker reduction process approximates a similar method employed by Burger et al. (2013). The marker configurations viewed for both piano and violin performers can be seen in Supplementary Material Figures S2 & S3.

Participants

A total of 92 Finnish university students aged 18–65 ($M = 25.66$, $SD = 7.95$; 63 female) participated in this study. Forty-five of the participants (49%) reported having received at least some musical training on an instrument (ranging from 1 to 50 years; $M = 10.58$, $SD = 8.21$). Participants were placed into one of three rating condition groups (see Procedure section below). There were no significant differences between the three groups in terms of age; $F(2,89) = 1.09$, $p = .342$, years of musical training; $F(2,89) = 0.166$, $p = .847$, or gender; $\chi^2(2) = 0.146$, $p = .929$. Due to technical issues, the data of two of the participants was not saved, resulting in a final sample of 90 participants. The participants received a free cinema ticket (value €9.75) as a reward for taking part in the study.

Procedure

Participants were randomly placed into one of three rating groups that differed only in terms of the type of stimuli presented. There were three rating conditions: audiovisual (AV), video-only (V), and audio-only (A). Participants in

Group 1 ($n = 31$) rated the A and V of the piano performances (note that one participant’s audio-only ratings were not saved due to a technical issue). Participants in Group 2 ($n = 34$) rated the A and V versions of the violin performances. Participants in Group 3 ($n = 25$) rated the AV versions of both the piano and violin performances. In all groups, the different types of stimuli (A, V, AV) were presented in respective blocks, and the stimuli within each block were presented in a different random order to each participant. Furthermore, the order of the blocks was balanced across participants.

The data collection sessions were conducted in a laboratory setting using a computer interface (see Supplementary Material Figures S2 and S3). The Max/MSP (version 5.1.9; Cycling 74, Walnut, CA) graphical programming environment (running on Mac OS X) was used to present the stimuli and collect the data. The point-light animations were presented with a resolution of 800×600 pixels and a frame rate of 30 fps. The audio was presented in WAV format through high-quality headphones (AKG K141 Studio). The participants were told they would hear and see short musical performances expressing different emotions, and their task would be to evaluate the degree to which the performances convey certain emotions. In the A and AV rating conditions, the participants were instructed to base their ratings of perceived emotion on what they *heard*. They were asked to “evaluate how tender, sad, happy, or angry the performances SOUND”. Similarly, in the V rating condition (without any sound), the participants were asked to “IMAGINE how tender, sad, happy, or angry the performances would sound”. The evaluations were made using four horizontal scales labeled *tenderness*, *sadness*, *happiness*, and *anger*, ranging from *Not at all* to *Very much*. The participants could use as many of the scales as they found applicable to any given performance (i.e., there was no forced choice). The outputs of the scales, coded using MIDI note numbers, provided data in the range 0–127.

The data collection sessions started with two practice trials using audiovisual excerpts that were like—but not part of—the actual stimulus set to which participants were instructed to respond. These responses were not included in the data. After completing the task, participants completed a short demographic questionnaire (including questions about their musical training) and were fully debriefed.

Results

Inter-Rater Agreement

The first step in the analysis was to investigate inter-rater agreement among the responses to musical performances. To this end, two analyses were performed on each subset of rated emotions: Interclass Correlation (two-way random, average measures, absolute agreement; abbreviated ICC; see Shrout & Fleiss, 1979) and Krippendorff’s alpha (Krippendorff, 2011). Table 1 lists the discrepancies between these methods. For Krippendorff’s alpha, the ranges varied from low to moderate, indicating a high

Table 1. Inter-rater agreement for perceived emotion.

Rating Condition	Rated Emotion	Piano		Violin	
		Krippendorff's alpha	ICC(2,k) ^a	Krippendorff's alpha	ICC(2,k) ^a
Audiovisual	Tenderness	0.38	0.94	0.33	0.93
	Sadness	0.45	0.96	0.46	0.96
	Happiness	0.49	0.96	0.61	0.98
	Anger	0.6	0.98	0.35	0.94
Video-only	Tenderness	0.26	0.92	0.2	0.91
	Sadness	0.24	0.91	0.26	0.93
	Happiness	0.2	0.89	0.16	0.87
	Anger	0.22	0.9	0.15	0.86
Audio-only	Tenderness	0.47	0.97	0.29	0.94
	Sadness	0.41	0.96	0.39	0.96
	Happiness	0.61	0.98	0.57	0.98
	Anger	0.63	0.98	0.4	0.96

^aShrout & Fleiss (1979).

variance in the responses. Conversely, ICC values were high, indicating consistency in variation across responses, even though their means may substantially differ. Despite this, both measures show a general pattern indicating that agreement was lower among responders in the video-only rating condition for performances of both instruments. Owing to the role of individual differences in emotional experiences, studies on music and emotion generally yield ratings with low agreement among responders (Vuoskoski et al., 2022; Zentner et al., 2008). Nonetheless, the averages of the ratings can be considered a suitable metric as they cancel out the differences while emphasizing the characteristics that multiple participants agree upon.

Comparing Differences Between Presentation Conditions

To explore the relative contribution of compositional aspects and performers' expressive intentions to participants' ratings of perceived emotion, and how this might vary across the three presentation conditions (A, V, & AV), a series of two-way repeated-measures ANOVAs were carried out. The two within-subjects factors were Type of Expression (deadpan, sad, happy, or angry) and Type of Composition (Tenderness, Sadness, Happiness, and Scare). The main aim of these analyses was to compare the magnitude of effect sizes (generalized eta squared; Bakeman, 2005) across presentation conditions and instruments. In addition, we explored the degree to which participants were able to accurately decode performers' expressive intentions in the A and V presentation conditions. Analyses were carried out separately for each presentation condition, emotion scale (perceived *tenderness*, *sadness*, *happiness*, and *anger*) and instrument (piano and violin). The results are summarized in Table 2, and the mean ratings are visualized in Supplementary Material Figures S4–S6.

For the audio-only condition, the main effect of Type of Composition was larger than the main effect of Type of Expression, suggesting that compositional aspects accounted for more variance in participants' ratings of perceived emotion. The mean effect size (generalized eta squared; Bakeman, 2005) of Type of Composition was $\eta_G^2 = .47$ for the piano, and $\eta_G^2 = .33$ for the violin performances, while the mean effect size of Type of Expression was $\eta_G^2 = .21$ for the piano, and $\eta_G^2 = .14$ for the violin (combined mean effects: Type of Expression $\eta_G^2 = .17$; Type of Composition $\eta_G^2 = .40$; see also Figure 1 for an illustration of the effect sizes across the three presentation conditions).

Post-hoc tests (paired t-tests with Holm–Bonferroni correction for multiple comparisons) revealed that in the piano performances, the target emotional expressions received the highest ratings on the corresponding rating scales but did not always differ significantly from all other expressive intentions: For perceived *sadness*, the sad expression was rated as the saddest, but was not significantly different from the deadpan expression. With respect to perceived *tenderness* (which did not correspond to any specific expressive intention), the sad expression was rated as the most tender, followed by deadpan, happy, and angry expressions. For the violin performances, the target emotional expressions also received the highest ratings on the corresponding rating scales, although *happiness* ratings did not differ significantly between the happy and angry expressions. These findings demonstrate that the participants were quite successful in decoding the expressive emotional intentions of the musicians based on auditory cues alone.

In the Video-only rating condition, Type of Expression played a more central role: The main effect of Type of Expression was substantially larger than the main effect of Type of Composition. The mean effect size of Type of Expression was $\eta_G^2 = .22$ for the piano, and $\eta_G^2 = .15$ for the violin performances, while the mean effect size of Type of Composition was $\eta_G^2 = .02$ for the piano, and $\eta_G^2 = .05$

Table 2. Summary results of the two-way repeated-measures ANOVAs exploring the relative effects of performers' expressive intention and composition on participants' ratings of perceived emotion.

Rating Condition	Rated Emotion	Instr.	Main effect of Type of Expression	Main effect of Type of Composition	Interaction
Audiovisual	Tenderness	Piano	$F(3,72) = 47.9^{***}, \eta_G^2 = .27$	$F(3,72) = 23.2^{***}, \eta_G^2 = .18$	***
		Violin	$F(3,72) = 70.9^{***}, \eta_G^2 = .26$	$F(3,72) = 9.5^{***}, \eta_G^2 = .09$	***
	Sadness	Piano	$F(3,72) = 33.2^{***}, \eta_G^2 = .13$	$F(3,72) = 49.9^{***}, \eta_G^2 = .41$	***
		Violin	$F(3,72) = 42.1^{***}, \eta_G^2 = .24$	$F(3,72) = 50.5^{***}, \eta_G^2 = .35$	***
	Happiness	Piano	$F(3,72) = 28.4^{***}, \eta_G^2 = .09$	$F(3,72) = 69.8^{***}, \eta_G^2 = .48$	***
		Violin	$F(3,72) = 46.7^{***}, \eta_G^2 = .24$	$F(3,72) = 127.5^{***}, \eta_G^2 = .54$	***
Anger	Piano	$F(3,72) = 30.1^{***}, \eta_G^2 = .17$	$F(3,72) = 132.1^{***}, \eta_G^2 = .58$	***	
	Violin	$F(3,72) = 36.6^{***}, \eta_G^2 = .19$	$F(3,72) = 34.1^{***}, \eta_G^2 = .23$	***	
Audio-only	Tenderness	Piano	$F(3,87) = 88.1^{***}, \eta_G^2 = .37$	$F(3,87) = 44.2^{***}, \eta_G^2 = .21$	***
		Violin	$F(3,99) = 49.7^{***}, \eta_G^2 = .23$	$F(3,99) = 18.4^{***}, \eta_G^2 = .07$	***
	Sadness	Piano	$F(3,87) = 28.2^{***}, \eta_G^2 = .13$	$F(3,87) = 69.2^{***}, \eta_G^2 = .40$	ns
		Violin	$F(3,99) = 33.9^{***}, \eta_G^2 = .11$	$F(3,99) = 67.4^{***}, \eta_G^2 = .34$	***
	Happiness	Piano	$F(3,87) = 23.9^{***}, \eta_G^2 = .13$	$F(3,87) = 166.3^{***}, \eta_G^2 = .62$	***
		Violin	$F(3,99) = 30.3^{***}, \eta_G^2 = .10$	$F(3,99) = 124.9^{***}, \eta_G^2 = .54$	***
Anger	Piano	$F(3,87) = 47.0^{***}, \eta_G^2 = .20$	$F(3,87) = 169.1^{***}, \eta_G^2 = .63$	***	
	Violin	$F(3,99) = 33.2^{***}, \eta_G^2 = .10$	$F(3,99) = 57.5^{***}, \eta_G^2 = .35$	***	
Video-only	Tenderness	Piano	$F(3,90) = 42.3^{***}, \eta_G^2 = .27$	ns, $\eta_G^2 = .01$	ns
		Violin	$F(3,99) = 36.6^{***}, \eta_G^2 = .17$	$F(3,99) = 10.3^{***}, \eta_G^2 = .03$	***
	Sadness	Piano	$F(3,90) = 30.5^{***}, \eta_G^2 = .22$	$F(3,90) = 6.2^{***}, \eta_G^2 = .03$	**
		Violin	$F(3,99) = 56.6^{***}, \eta_G^2 = .22$	$F(3,99) = 15.0^{***}, \eta_G^2 = .06$	***
	Happiness	Piano	$F(3,90) = 28.5^{***}, \eta_G^2 = .20$	$F(3,90) = 3.75^*, \eta_G^2 = .02$	ns
		Violin	$F(3,99) = 26.0^{***}, \eta_G^2 = .09$	$F(3,99) = 19.4^{***}, \eta_G^2 = .08$	ns
Anger	Piano	$F(3,90) = 35.7^{***}, \eta_G^2 = .20$	$F(3,90) = 4.73^{**}, \eta_G^2 = .03$	*	
	Violin	$F(3,99) = 26.9^{***}, \eta_G^2 = .12$	ns, $\eta_G^2 = .01$	***	

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

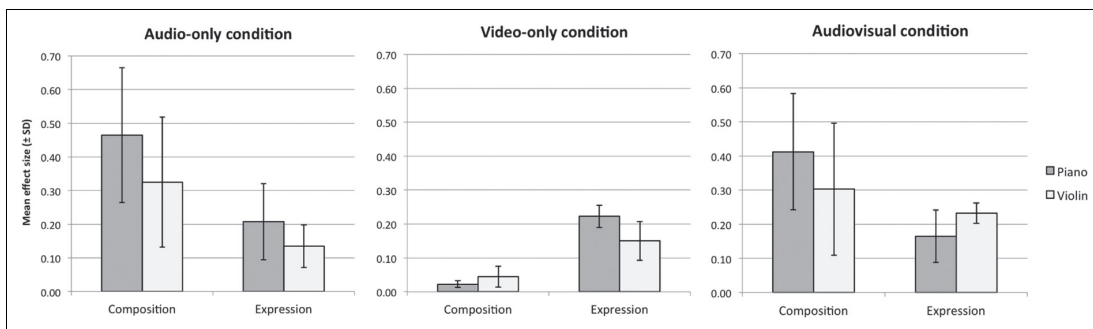


Figure 1. The mean effect sizes (generalized eta squared; Bakeman, 2005) \pm standard deviations of Type of Composition and Type of Expression on ratings of perceived emotion across the three presentation conditions and two instruments.

for the violin (combined mean effects: Type of Expression $\eta_G^2 = .19$; Type of Composition $\eta_G^2 = .03$). In other words, in the absence of auditory information, the type of expressive intention accounted for substantially more of the variance in participants' ratings.

Post-hoc tests (paired t-tests with Holm–Bonferroni correction for multiple comparisons) revealed that in the piano performances, the target emotional expression was always rated as significantly higher than any other emotion. Similarly to the audio-only condition, the sad expression was rated as the most tender, followed by deadpan, happy, and angry expressions. For the violin

performances, the target emotional expressions received the highest ratings on the corresponding rating scales but did not always differ significantly from all other expressive intentions: For perceived sadness, the sad expression was rated as the saddest, but was not significantly different from the deadpan expression. For perceived happiness, the happy expression was rated as the happiest, but did not differ significantly from the angry expression. These findings show that participants were able to decode the expressive emotional intentions of the musicians based on visual kinematic cues alone, albeit with substantial imprecision.

In the Audiovisual condition, the differences between the mean effect sizes of Type of Expression and Type of Composition were slightly reduced: The mean effect size of Type of Expression was $\eta_G^2 = .17$ for the piano, and $\eta_G^2 = .23$ for the violin performances, while the mean effect size of Type of Composition was $\eta_G^2 = .41$ for the piano, and $\eta_G^2 = .30$ for the violin (combined mean effects: Type of Expression $\eta_G^2 = .20$; Type of Composition $\eta_G^2 = .36$). These findings suggest that, compared to the audio-only condition, visual kinematic information enhanced the perceptual salience of expressive intentions in relation to the compositional features.

Relationships Between Rated Emotions and Computed Features

Motion Features. For each rated performance, motion features (*mo*) were computed from the motion-capture data. To permit comparing piano and violin performances, we computed features related to the kinematic aspects of movement (see Dahl & Friberg, 2007). These measures correspond to the magnitudes of the numerical approximations of the time-derivatives velocity (*avbspeed*), acceleration (*avaccmag*), jerk (*avjrkmag*), and City-Block Total Distance (*cbtotdist*) as a measure of total amount of movement (Camurri et al., 2004; Thompson & Luck, 2012). These were computed for five marker groups for each musician. The marker groups for the pianist were the upper body (*ub*), torso (*t*), head (*h*), left finger (*lf*), and right finger (*rf*). The marker groups for the violinist were full body with violin and bow (*f*), torso (*t*), head (*h*), left finger (*lf*), and right finger (*rf*) (see Tables 3 and 4). The marker groups differ between pianist and violinist. For the pianist, the ‘upper body’ comprises markers on the torso, head, elbows, wrist, and middle fingers (no markers placed on the lower body). The violinist performed in a standing position, and markers were placed on the whole body, bow, and violin because they were included in the rated videos. Additionally, the functions of the right- and left-hand

fingers differ between instruments. In the case of the pianist, the right hand typically plays the melody, whereas the left-hand plays harmony. In the case of a violin, the right hand moves the bow, whereas the left-hand fingers depress the strings on the fingerboard. Each resulting motion feature is composed of 16 data points, meaning one for each performance.

Audio Features. As with motion features, the criterion for selecting audio features was that they should permit a comparison between instruments. Six audio features (*au*) were computed from each performance. Two features were derived from the manual annotations of the note onsets: average performance speed (*avpspeed*) and standard deviation (*varpspeed*). Next, the mode (*mode*) of each piece was annotated from the score (positive unit for the major mode and vice versa). The last three features, related to signal energy, were computed using MIRTtoolbox v. 1.6.3 (Lartillot et al., 2008). These were root-mean-square energy (*rms*), and sub-band flux of bands 3 (*avfluxsb3*) and 7 (*avfluxsb7*) (Alluri & Toiviainen, 2010). These bands account for variability of energy in the lower and higher portions of the frequency spectrum, respectively (see Table 3 for an overview). As with motion features, each resulting audio feature is composed of 16 data points.

Correlation Between Features. To assess the distinctiveness of each feature, Figure 2 displays the level of correlation between all motion and audio features. A lower correlation (indicated by white or lightly shaded cells) with other features indicates high distinctiveness. For motion features (*mo*), total distance (*cbtotdist*) is notably distinct, for all marker groupings and both instruments. The time-derivative features (*avbspeed*, *avaccmag*, and *avjrkmag*) are highly intercorrelated for all the marker groups of the violinist but not for the marker groups of the pianist. For audio features, mode is the most distinct. However, it must be acknowledged that this feature is identical for both instruments and a dichotomous feature treated as continuous, as opposed to the true continuous nature of the

Table 3. Summary of motion and audio features calculated from piano and violin performances. Motion features were calculated for each marker group.

	Full Name	Abbreviation
Motion Features (<i>mo</i>)	Average Speed	avbspeed
	Average Acceleration Magnitude	avaccmag
	Average Jerk Magnitude	avjrkmag
	City-Block Total Distance	cbtotdist
Audio Features (<i>au</i>)	Average Performance Speed	avpspeed
	Variability of Performance Speed	varpspeed
	RMS Energy	rms
	Average Flux of Sub-Band 3	avfluxsb3
	Average Flux of Sub-Band 7	avfluxsb7
	Mode (major, minor)	mode

Table 4. Summary of marker groups for violinist and pianist. From each marker group, four movement features were computed (see Table 3).

	Marker Group	Abbreviation
Pianist	upper body	ub
	torso	t
	head	h
	left finger	lf
	right finger	rg
Violinist	full body w/ violin and bow	f
	torso	t
	head	h
	left finger	lf
	right finger	rf

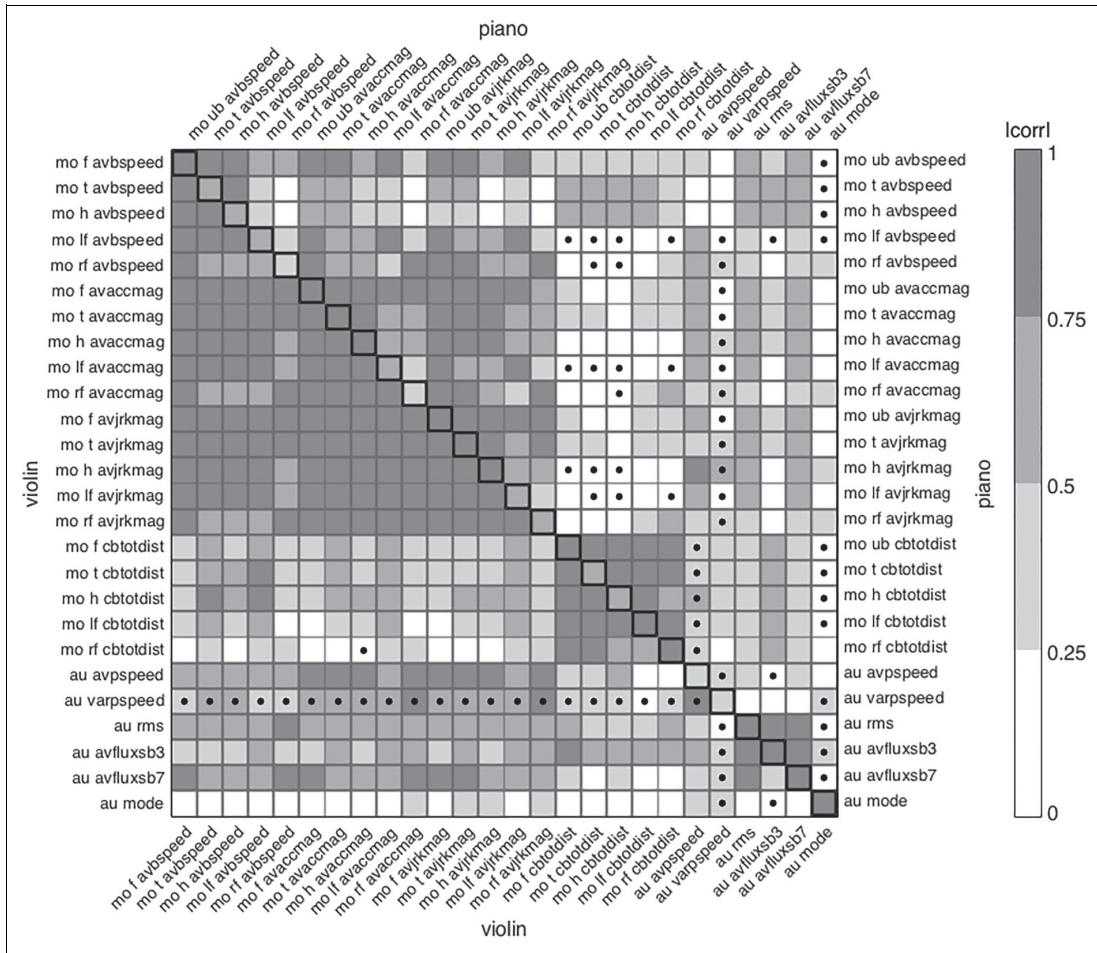


Figure 2. Pearson's correlation coefficients between features. The upper triangle shows correlation for piano performances and the lower triangle for violin performances. The diagonal shows correlation of features between piano and violin. A dot indicates negative coefficient.

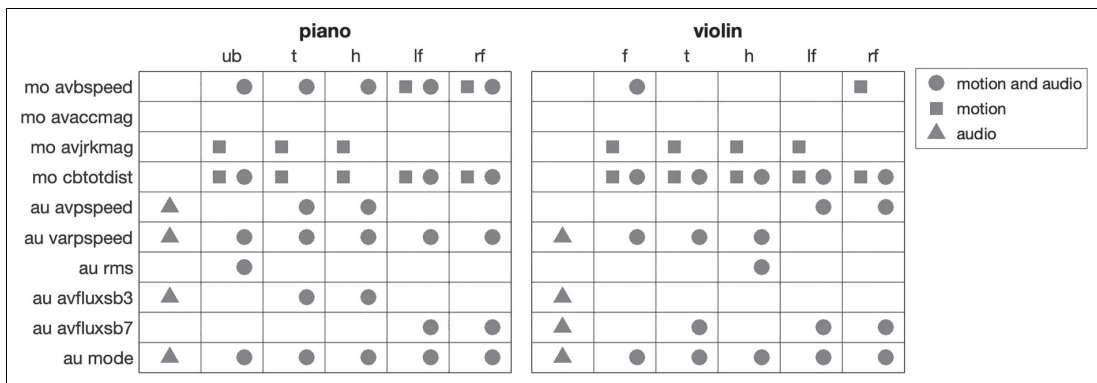


Figure 3. Features with low collinearity when among all features (motion and audio), only motion features, and only audio features.

other features. Energy and spectral flux are generally highly intercorrelated, but in the case of the violin, the lower spectral flux sub-band (*avfluxsb3*) differs from the higher spectral flux sub-band spectral flux (*avfluxsb7*). This may be due to this instrument only playing melody within a middle to high register, reaching the lower spectral band only occasionally, therefore having greater variability in this band. On the contrary, piano performances involve playing chords with the left hand, and thus the lower spectrum content is more homogeneous.

The inclusion of motion and audio features within the same figure highlights potential cross-modal relationships. For the violinist, average performance speed (*au avpseed*) was moderately to highly positively correlated with time-derivatives (*avbspeed*, *avaccmag*, and *avjrkmag*) and total distance (*cbtotdist*) of each marker group while variation in speed (*varpspeed*) was inversely correlated. Similar relationships appear in the piano performances, except that total distance (*cbtotdist*) was inversely correlated to variation in speed and positively correlated to average speed.

Feature Selection Based on Low Collinearity. Figure 2 provides a snapshot of collinearity within the feature set. To make the multiple regression models as reliable as possible, subsets of motion and audio features with low collinearity were compiled so that their Variance Inflation Factor would not exceed two. These subsets included either motion features, audio features, or a combination of both. Subsets with motion features contained only features for one marker group. Motion measures were found to be highly collinear, but at least one motion feature was retained in each feature subset. It was desirable to have at least one motion feature in the subsets with audio and motion features combined, so that they would ‘compete’ for inclusion in a model, being those most statistically significant (lower *p*-value) the ones included. Figure 3 shows the features retained, revealing distinct patterns for each instrument. For instance, among the motion features in the case of the piano, both hands (*lf* and *rf*) have different characteristics than the torso (*t*) and head (*h*), while in the case of the violin the bow hand (*rf*) is distinct.

Correlations Between Features and Ratings. Linear correlations were computed between each feature and mean ratings for each emotion. Figures 4 and 5 show these values for piano and violin performances, respectively. The most striking result is that, for both instruments, the time-derivatives of motion have greater correlations with perceived emotions when audio is absent in the stimulus. Also, the time-derivatives, for both instruments, are inversely correlated with the ratings of *sadness* and *tenderness*. This effect is greater in the violin ratings, showing clear positive correlation between motion time-derivatives and the ratings of *happiness* and *anger*. The correlations between time-derivatives and ratings for violin performances are stronger than for piano performances when the stimuli are audiovisual. However, in the audio-only

condition, all features have very low or no correlation with the ratings obtained for the violin performances. Conversely, for piano performances, the relations between emotion ratings and all features are remarkably similar in both conditions where audio is presented. Audio features that are highly correlated with the motion derivatives also have high correlations with emotion ratings. There is of course a clear relation between the physical energy used to produce a sound and the energy of the resultant sound, reflected, for example, in the features performance speed (*avpspeed*) and RMS energy (*rms*).

Calculation and Selection of Regression Models. A Simple Ordinary Least Squares (OLS) Linear Regression Model (LRM) was computed for each feature as an independent variable and the mean value of each rating subset as the dependent variable. Additionally, a Multiple OLS-LRM was computed for all the possible permutations of features within each subset of features previously screened. For example, the subset of audio features and torso motion features of piano performances has five low-collinearity features. Models including all permutations of two to five features were computed. All features were standardized so that the coefficients of a model can be used as an indicator of the contribution of their corresponding feature to the model. A regression model is expressed as an equation in the form.

$$Y = C + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \epsilon$$

where *Y* is the mean responses (perceived emotions) vector, *C* is a constant vector, β is a weight coefficient vector for each vector *X* of features $\{1, n\}$ included in the model and ϵ is the error vector. Of this equation, only the weight (β) coefficients and their corresponding *t*-test *p*-value are considered for analysis, as they provide information about the contribution of each feature in the model. The constant term does not provide any useful information for the purpose of this study. Also, since the number of data points is low ($n = 16$), assessment of the error term is irrelevant. The adjusted coefficient of determination (Adjusted R^2 or R^2_{adj}) was used to assess a model’s goodness-of-fit adjusting for the number of features included. All multiple regression models that had at least one weight coefficient having a *t*-test *p*-value equal or greater than 0.05 were discarded, but all the simple regression models were retained for further examination. Tables 5 and 6 summarize the retained models, with a single simple regression model (Table 5) and single multiple regression (Table 6) selected for each mean rating of perceived emotions. Additionally, the tabulated models had the lowest Corrected Akaike Information Criterion, which increases with a model’s fit but penalizes the addition of features, also adjusting for the small number of data points (Hurvich & Tsai, 1989).

Simple Linear Regression Models. Table 5 shows the selected simple linear regression models for motion audio features.

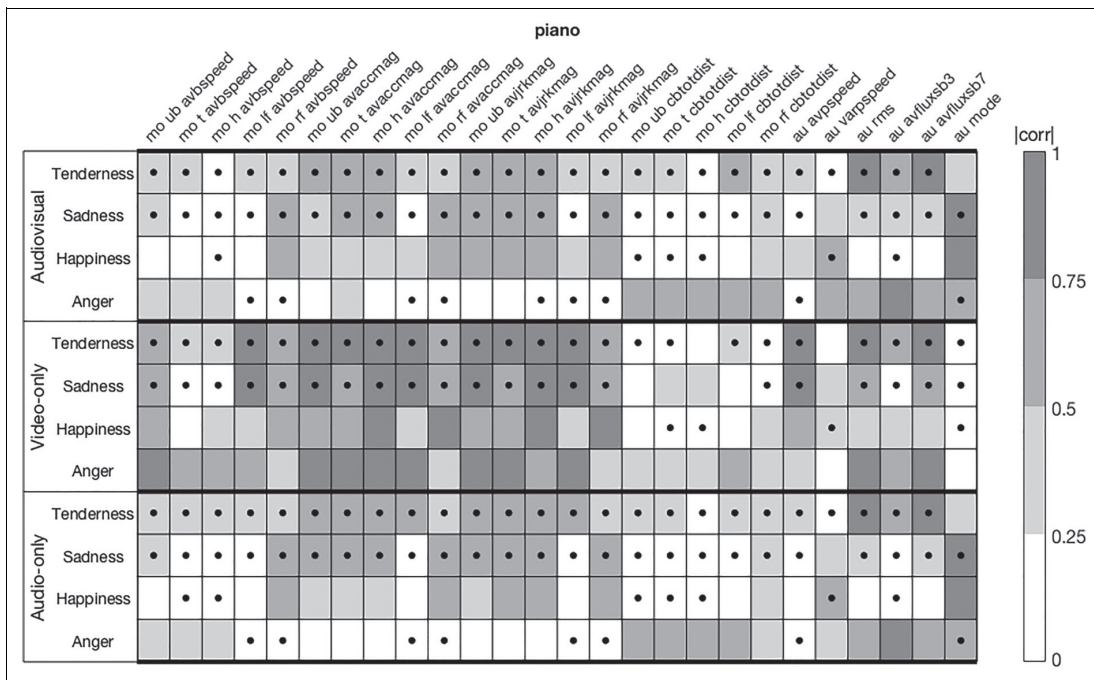


Figure 4. Pearson's correlation between mean ratings of perceived emotions and computed features of piano performances. A dot indicates negative coefficient.

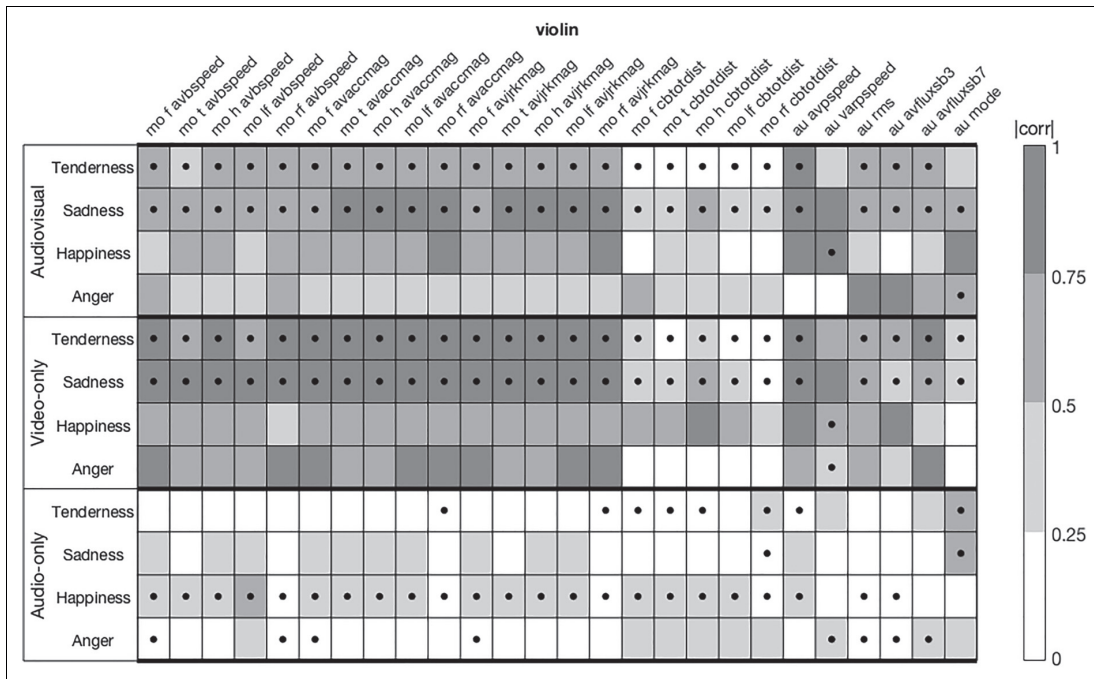


Figure 5. Pearson's correlation between mean ratings of perceived emotions and computed features of violin performances. A dot indicates negative coefficient.

Table 5. Simple linear regression results for motion (mo) and audio (au) features.

Rating Condition	Instrument	Rated Emotion	Simple linear regression models for Motion				Simple linear regression models for Audio			
			Adjusted R ²	Motion Feature	β sign	p	Adjusted R ²	Audio Feature	β sign	p
Audiovisual	piano	Tenderness	0.3	mo t avaccmag	-	0.02	0.64	au rms	-	<0.01
		Sadness	0.3	mo rf avbspeed	-	0.02	0.57	au mode	-	<0.01
		Happiness	0.34	mo rf avbspeed	+	0.01	0.74	au mode	+	<0.01
		Anger	0.35	mo t cbtodist	+	<0.01	0.76	au avfluxsb3	+	<0.01
	violin	Tenderness	0.43	mo f avjrkmag	-	<0.01	0.5	au avpspeed	-	<0.01
		Sadness	0.65	mo rf avjrkmag	-	<0.01	0.65	au avpspeed	-	<0.01
		Happiness	0.56	mo rf avjrkmag	+	<0.01	0.72	au varpspeed	-	<0.01
		Anger	0.24	mo f avbspeed	+	0.03	0.58	au rms	+	<0.01
Video-only	piano	Tenderness	0.85	mo h avaccmag	-	<0.01	0.66	au rms	-	<0.01
		Sadness	0.71	mo h avjrkmag	-	<0.01	0.55	au avpspeed	-	<0.01
		Happiness	0.5	mo h avjrkmag	+	<0.01	0.34	au avpspeed	+	<0.01
		Anger	0.64	mo t avaccmag	+	<0.01	0.68	au rms	+	0.01
	violin	Tenderness	0.84	mo f avjrkmag	-	<0.01	0.62	au avpspeed	-	<0.01
		Sadness	0.81	mo h avjrkmag	-	<0.01	0.8	au avpspeed	-	<0.01
		Happiness	0.48	mo h cbtodist	+	<0.01	0.61	au avpspeed	+	<0.01
		Anger	0.76	mo rf avbspeed	+	<0.01	0.69	au avfluxsb7	+	<0.01
Audio-only	piano	Tenderness	0.33	mo t avaccmag	-	0.01	0.67	au rms	-	<0.01
		Sadness	0.36	mo rf avbspeed	-	<0.01	0.62	au mode	-	<0.01
		Happiness	0.35	mo rf avbspeed	+	<0.01	0.76	au mode	+	<0.01
		Anger	0.32	mo t cbtodist	+	0.01	0.75	au avfluxsb3	+	<0.01
	violin	Tenderness	0.01	mo rf cbtodist	-	0.3	0.36	au mode	-	<0.01
		Sadness	0.08	mo f avjrkmag	+	0.14	0.3	au mode	-	0.02
		Happiness	0.17	mo lf avbspeed	-	0.06	0.04	au avpspeed	-	0.22
		Anger	0.12	mo t cbtodist	+	0.1	0.16	au mode	+	0.07

Note. avbspeed = average speed; avaccmag = average acceleration magnitude; avjrkmag = average jerk magnitude; cbtodist = city-block total distance; avpspeed = average performance speed; varpspeed = performance speed standard deviation; rms = root mean square; mode = mode (major or minor); avfluxsb3 = sub-band flux band 3; avfluxsb7 = sub-band flux band 7; t = torso; f = full body; h = head; lf = left finger; rf = right finger.

Table 6. Multiple linear regression results for motion and audio features.

Rating Condition	Instr.	Rated Emotion	Adjusted R^2	F-test p-value	Marker group	Motion avpspeed	Motion cbtotdist	Motion varpspeed	Audio avfluxsb3	Audio avfluxsb7	Audio mode	
Audiovisual	piano	Sadness	0.9	< 0.01	–	β	–	–	–	–17.1	–	–27.7
						p	–	–	–	<0.01	–	<0.01
		Happiness	0.83	< 0.01	rf	β	10.5	–	–	–	–	23.8
						p	0.01	–	–	–	–	<0.01
		Anger	0.93	< 0.01	–	β	–	–	8.3	23.4	–	–7.6
						p	–	–	<0.01	<0.01	–	<0.01
	violin	Tenderness	0.52	< 0.01	–	β	–	–	11.3	–	–8.5	13.3
						p	–	–	0.02	–	0.04	<0.01
		Sadness	0.72	< 0.01	–	β	–	–	–	–19.1	–	–18.8
						p	–	–	–	<0.01	–	<0.01
		Happiness	0.87	< 0.01	–	β	–	–	–23.4	–	–	15.8
						p	–	–	<0.01	–	–	<0.01
Anger	0.82	< 0.01	–	β	–	–	–	9.7	6.4	–9.4		
				p	–	–	–	<0.01	0.01	<0.01		
Video-only	piano	Anger	0.88	< 0.01	lf	β	6.5	6.7	–	–	9.1	4.2
						p	<0.01	<0.01	–	–	<0.01	0.02
	violin	Happiness	0.76	< 0.01	–	β	–	–	–8.8	11.1	–	–
						p	–	–	<0.01	<0.01	–	–
Audio-only	piano	Sadness	0.89	< 0.01	rf	β	–	–8.2	–	–	–10	–23.5
						p	–	<0.01	–	–	<0.01	<0.01
		Happiness	0.85	< 0.01	rf	β	11.4	–	–	–	–	25.8
						p	<0.01	–	–	–	–	<0.01
	violin	Anger	0.91	< 0.01	–	β	–	–	7.9	23.3	–	–8
						p	–	–	0.01	<0.01	–	0.01
		Anger	0.44	< 0.01	lf	β	–	11	–	–	–9.9	11.8
						p	–	0.03	–	–	0.05	0.02

These models indicate the degree of linear relation between features and ratings, measured by R^2_{adj} . In practice, it is a correlation analysis. However, the use of R^2_{adj} allows comparison with the models that have more than one feature. The table only displays the models' beta coefficient (β) sign, as the magnitude is irrelevant for models with a single feature. The sign indicates whether the relation is positive or negative. The inclusion of motion and audio features within the same table allows direct comparison between each rated emotion's selected features. Models displayed in bold characters have the higher fit for each rated emotion. Models with p-value ≥ 0.05 are deemed to be statistically insignificant but still worth of noting value is not much greater than 0.05. Models that have higher R^2_{adj} than their multiple regression counterparts (see Table 6) are shown with gray background. The same overall observations made for the correlation analysis apply to Table 5. However, these models are intended to observe the strongest relationships between ratings and features.

In the movement feature models, for *happiness* and *anger*, the relation of movement features is positive, meaning that higher movement feature values (or more activity) produce higher ratings for these emotions, except the case for perceived happiness in violin performances. Meanwhile, *sadness* and *tenderness* ratings had a negative relation with the movement features. This was the case for all piano-rating conditions (audiovisual, video-only,

audio-only) and for most violin-rating conditions (audiovisual, video-only). For piano ratings, the relations were strongest in the video-only condition, and the strongest relation ($R^2_{adj} = 0.85$) was between *tenderness* ratings and head acceleration (*mo h avaccmag*). The audio-only condition resulted in non-significant models for the violin performances. Interestingly, in the case of the piano, the audio-only models are similar to those for the audiovisual condition with R^2_{adj} values varying between 0.3 and 0.36.

Turning to the audio feature models, the relation between mode is negative for *sadness* and positive for *happiness* ratings, meaning that minor mode corresponds to perceptions of sadness and major mode corresponds to perceptions of happiness. The ratings of *anger* have a positive relation with the variation of lower spectral content (*avfluxsb3*). This can be observed in the ratings obtained in both audiovisual and audio-only conditions, which correspond to different responders. For ratings of violin performances presented in the audiovisual condition, the average performance speed computed from audio annotation (*avpspeed*) is negatively related to *tenderness* and *sadness*, while its standard deviation (*varpspeed*) is negatively related to *happiness*. Also in the case of the violin, audio energy (rms) is positively related to anger when presentation is audiovisual, but not when it is audio-only. Audio energy is positively related to *anger*, as is the case for the piano. Finally, it is worthwhile to remark that for

performances of both instruments, audio features dominate when the stimuli include audio.

Multiple Regression Models. None of the multiple regression models computed and selected with the described procedures for subsets of motion-only candidate features yielded higher R^2_{adj} than their simple regression counterparts. All the multiple regression models computed for subsets of audio-only candidate features and having higher R^2_{adj} than their simple regression counterparts appeared in the models selected from the ones computed with the subset of audio and motion candidate features, except one. The model for *sadness* ratings of audio-only piano performances composed only by audio features is slightly improved by adding total distance of the right finger and replacing variability of the lower part of the spectrum for variability of the higher part of the spectrum.

Hence, Table 6 contains the multiple regression models with highest R^2_{adj} of all the multiple regression candidate features subsets, and with higher R^2_{adj} than their simple regression counterparts. All models in Table 6 for ratings that included audio, have mode included. For positive valence mode is positive and vice versa for negative valence, as in the simple regression models. An exception is the model for *anger* when only video was presented, which seems to be spurious and suggesting an effect of serendipity, as at least theoretically and intuitively, mode does not contribute to visual perception. For the ratings of *sadness* in audiovisual condition, the models are the same in terms of features and the sign of their coefficients: lower spectrum and minor mode. A similar effect is observed for *anger* ratings: high variability of lower spectrum and minor mode. However, the model for piano performances includes variability of performance speed (*avpspeed*) and the model for violin performances also includes variability of higher spectrum (*avfluxsb7*). Models for *tenderness* and *happiness* ratings had more distinct characteristics for each instrument, variability of performance speed (*avpspeed*) and lower spectrum (*avfluxsb3*) had different signs for either instrument.

In the case of the piano, as it can be seen with simple regression, models for *happiness* and *anger* ratings in both conditions with audio are similar, meaning the same features with very close coefficient values. For *sadness* ratings there is an inverse relation with variation of lower spectrum, while for *anger* the relation is positive. This may be because the pianist played chords with less dynamics in the sadly intended pieces, while for the anger-intended emotion, the pianist may have hit the chords more forcefully.

A few models were improved over the simple regression, by including either or both average bodily speed and total distance. Notably, some models that correspond to ratings for performances presented without video have relevant contributions of motion features. Also, for ratings of piano performances, the marker of the right finger has the greatest contribution for *happiness* (average speed) and

sadness, while the left hand for *anger* (average speed and total distance). Presumably this is because the right hand plays the melody noticeably fast. Likewise, the left hand plays the chords and, as it has been said, they might have been hit more energetically in the pieces with higher ratings for *anger*. In the case of the violin, the model for *anger* ratings was improved with the inclusion of total distance and variability of high spectrum, to the existing mode that alone has very low correlation. Finally, neither simple nor multiple regression yielded a strong model for happiness ratings of violin performances presented as audio-only.

Discussion

Advancing previous work by Vuoskoski et al., (2014) and Vuoskoski et al. (2016), this study investigated cross-modal contributions of sight and sound in the perception of expressed emotion in musical performance. Participants rated piano and violin performances in which the musicians played four short pieces attempting to convey four emotional intentions. We also examined the effect of presentation modality (AV, A or V) on participant ratings of perceived emotion, and the relationships between ratings and motion and audio features derived from performance data (motion-capture and audio recordings). Linear relationships between the features and ratings were measured using Pearson's correlation and ordinary least-squares regression. Below, we outline the study's main results, reflect on its limitations, and suggest avenues for future research.

A general finding was that emotion ratings were more consistent among responders when audio was present (audiovisual and audio-only conditions). This suggests that music provided cues that most responders interpreted in more unified ways as opposed to a wider variety of interpretations within the video-only condition. However, the post-hoc tests of the ANOVA analyses revealed that participants were marginally able to decode performers' expressive intentions based on both visual and auditory information alone (Akkermans et al., 2019, and Gabrielsson & Juslin, 1996). In rating conditions where audio was present, compositional aspects had a stronger effect on participant ratings than performers' expressive intentions. The same pattern of results was observed in both the ANOVAs and the analysis of musical and acoustic features, where the musical mode (whether the piece was major or minor) was the dominant variable in predicting the ratings of perceived emotions. Thus, the compositional aspects of the performances were generally stronger than the expressive aspects when it came to evaluating performances for perceived emotion.

Relations Between Audio and Motion Features

A correlation analysis shows relationships between audio and motion features (see Figure 2). For violin performance, the average speed at which pieces were played was

positively correlated with the violinist's motion acceleration and jerkiness (i.e., faster performances were played with greater acceleration through all parts of the body). The relationship between speed, acceleration, and jerk appeared in piano performance, but to a lesser extent. Of all the features and for both violin and piano performances, the least correlated feature was mode. This indicates that mode (major or minor), being a compositional aspect, did not affect performance aspects (e.g., performance speed as indicated by speed or dynamics as indicated by RMS).

Relations Between Performance Features and Perceived Emotion Ratings

Correlations between performance features and perceived emotions were different for each instrument but exhibited some commonalities (details in Figures 4 and 5, and Tables 5 and 6). The correlation between ratings for perceived emotions in piano performances and computed features were very similar when audio was presented, the audio features being much stronger than motion features. The correlations between ratings for perceived emotions in violin performances ranged from nil to low when the performance was presented only as audio. For both instruments' performances, the correlations between perceived emotions and motion features were generally stronger when audio was not present, and mostly higher than audio features.

For both instruments, performances presented as video with audio or only video, the time-derivatives of motion correlated directly with high activity emotions (*happiness* and *anger*), and inversely with low activity emotions (*tenderness* and *sadness*). This effect was also observed for the piano performances presented without video and was much stronger for violin performances than for piano performances. In general, acceleration and jerk had the highest correlations among motion features with rated emotions. Multiple regression models made of low-collinearity features fit better to perceived emotions for most emotions and presentation conditions (details in Table 6). Crucially, all multiple regression models for ratings where audio was presented had better fit after including mode in the model. Also, when audio was presented, mode was always directly related to positive valence emotions (*tenderness* and *happiness*), and inversely related to negative valence emotions (*sadness* and *anger*). However, exceptions are the *anger* and *tenderness* ratings in the violin's performances, relating to major and minor mode, respectively. This suggests that while there are plenty of commonalities in piano and violin performances, each instrument had qualities that affected perceived emotion differently.

The relationships between performance features and perceived emotion ratings were generally consistent with previous research. For audio features, we found relationships between mode and emotional valence (whether a piece was happy or sad), which aligned with work cited in the introduction (e.g., Dalla Bella et al., 2001; Juslin, 2000; Peretz et al., 1998). Regarding the relationship between

amount of movement and *happiness* ratings being direct, and the relationship between amount of movement and *sadness* ratings being inverse, we found partial agreement with Dahl and Friberg (2007). Our analysis verified this for all the marker groups on the violinist when the video was presented and weak or inverse when only audio was presented. For piano performance ratings, this relationship only appeared in the movement of the right hand (*rf*) when audio was present. The relationships between ratings in the audio-only condition of violin performances, and all performance features (motion and audio) ranged from irrelevant to weak. Conversely, this relationship was observed to be substantially stronger for video-only and audiovisual conditions. While unclear why this occurred, it could be due to the melodic nature of the violin performances, which lacks the additional information provided by the harmonic accompaniment by the left hand in piano performances. Also, the most-correlated motion features for ratings of violin performances when video was presented, are performance speed and variability of performance speed (*avpspeed* and *varpspeed*, respectively), and energy (*rms*). These are moderately to strongly correlated with motion time-derivatives, suggesting that responders made their assessment of violin performances with more emphasis on movement, while assessing piano performances with more emphasis on sound, likely due to the presence of chords reinforcing the mode (major or minor). It should be noted that our analysis is meant to model and predict the dataset at hand, and not be read as being generalizable to other similar data. However, the models point towards some tentative conclusions, potentially useful for future research—the main one being that mode had a consistent and robust effect on ratings when audio was present.

Limitations of the Study. This study has some limitations that could be considered when designing follow-up work. First, participants viewed performances by only two musicians. This resulted from our choice to have one musician per instrument to limit the number of stimuli presented to the participants. Related to the choice of musical instrument, the performance features used for correlation and regression analysis were limited to features that would apply equally to pianos and violins. Using piano and violin performances of the same pieces, our results may not generalize to all musical instruments. Rather, the results evince differences in the types of expressive gestures made by different instrumentalists.

Another potential limitation is the inter-rater agreement. To ensure that the perceptual data would be suitable for regression analysis, we tested the agreement among the ratings using Intraclass Correlation (ICC) and Krippendorff's alpha. Although both tests showed a similar pattern between rating conditions (e.g., agreement was generally lower for the video-only condition), the overall results of Krippendorff's alpha were remarkably lower than the ICC. The artifacts resulting in high ICC

may be due to tau-inequivalence and non-normality, which are impractical to measure in a small sample size ($N = 31$ or 34 in this study). Conversely, Krippendorff's alpha is independent of n (it can be two raters or one million). The literature on music and emotion research has primarily focused on ICC measures. While precedence is not the best justification for using ICC, it must be acknowledged that the level of agreement in our participants is typical for music and emotion experiments, and that individual differences are a central part of emotional experiences. Our regression models explained quite a bit of the variance in the mean ratings by the motion and acoustic features. This demonstrates that the mean ratings reflect something salient about participants' emotion perception. Our goal was not to account for all the variance, but to identify which of the motion and acoustic features played a greater role in the participants' evaluations. We hope that the discrepancies between Krippendorff's alpha and ICC discussed here will be useful for future research in this area.

A more general limitation of the current study is that its findings are not scalable to music cultures outside the western classical tradition. The current study took for granted that the participants had been raised in circumstances that would have exposed them to western classical tonal music. This assumption was made solely since all participants were of Finnish nationality. Thus, it is inappropriate to claim that the findings of this study are valid for other musical cultures. Interestingly, the strongest signifier of perceived intention was the musical mode, with ratings strongly correlating with the traditional view that music in major keys is positively valenced and music in minor keys is negatively valenced. A recent study by Smit et al. (2022) found that the major/minor valence dichotomy is, by and large, something that is learned through culture as opposed to being universal across musical cultures. Therefore, our results should be kept within the scope of western enculturated music listeners.

Finally, let us consider the various theories mentioned in the introduction. Embodied and enactive approaches explaining the perception and cognition of music propose that gestures performed by musicians play a significant role in communicating musical expressivity. This may stem from linguistic communication, in which gestures convey clear messages and can substitute speech (McNeill, 2007). Generally, there is a sense that when musicians produce performance gestures with their body, observers are able to interpret expressive or emotional intentions. However, music performance gestures (outside of conducting gestures) lack the clear meaning analogous to linguistic gestures. In other words, they remain ambiguous, particularly when presented without the performance's auditory component. This appears to be what our data is pointing at, as compositional factors such as musical mode (major/minor) acted as clear predictors as to whether a piece was rated sad or happy, despite the musicians' performance intentions. While there is a strong semiotic component to music, it remains more salient in auditory

and compositional cues than movement cues. This does not mean that the ecological approach is not applicable to music performance, but our findings indicate that auditory affordances are more useful in communicating discrete musical emotions than movement cues. Regarding cross-modal recognition, although motion and audio features were correlated (e.g., louder music correlated with faster movements), our paradigm could have more strongly emphasized cross-modal recognition in the experimental design. For instance, to properly test Windsor's idea that the audio channel contains affordances that a listener would use to create images of the musician's movements, a future study could ask participants to specifically describe the gestures being performed while listening to music.

Conclusion

This study builds on previous research on how musical performance serves as a means of expressive and emotional communication. We found that performances rated high in happiness and anger were characterized by greater variation in musical dynamics, speed, and activity, whereas performances rated high in sadness and tenderness had more subtle dynamics and movements. Interestingly, when the presentation condition of the performance was audiovisual or audio-only, ratings of perceived emotions were mainly influenced by compositional elements, such as musical mode, rather than by musicians' emotional intentions. This suggests that compositional structure within the context of western classical music has a stronger impact on an audience's emotional response than the performer's gestures (see Laukka & Gabrielsson, 2000). However, when the presentation condition was video-only, participants were able to decode emotional intentions to a much lesser extent. Additionally, inter-rater agreement was lowest when the presentation condition was video-only, signifying the ambiguity of expressive gestures without the full context of the performance, which is a valuable starting point for future work on this subject. To expand this research, future studies could explore different musical instruments and non-western musical cultures while implementing the methodological suggestions made in this study.

Action Editor

Diana Omigie, Department of Psychology, Goldsmiths, University of London.

Peer Review

Two anonymous reviewers.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.


Ethical Approval

The experimental procedures followed the University of Jyväskylä's policy on the ethical conduct of research involving human participants, and informed consent was obtained from all participants.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Academy of Finland (Centre of Excellence in Music, Mind, Body and Brain) and the Kone Foundation (post-doctoral funding to Jonna Vuoskoski).

ORCID iD

Marc R. Thompson  <https://orcid.org/0000-0003-0662-1812>

Supplemental Material

Supplemental material for this article is available online.

References

- Akkermans, J., Schapiro, R., Müllensiefen, D., Jakubowski, K., Shanahan, D., Baker, D., Busch, V., Lothwesen, K., Elvers, P., Fischinger, T., Schlemmer, K., & Frieler, K. (2019). Decoding emotions in expressive music performances: A multi-lab replication and extension study. *Cognition and Emotion*, *33*(6), 1099–1118. <https://doi.org/10.1080/02699931.2018.1541312>
- Alluri, V., & Toiviainen, P. (2010). Exploring perceptual and acoustical correlates of polyphonic timbre. *Music Perception*, *27*(3), 223–242. <https://doi.org/10.1525/mp.2010.27.3.223>
- Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods*, *37*(3), 379–384. <https://doi.org/10.3758/BF03192707>
- Battcock, A., & Schutz, M. (2019). Acoustically expressing affect. *Music Perception*, *37*(1), 66–91. <https://doi.org/10.1525/mp.2019.37.1.66>
- Behne, K.-E., & Wöllner, C. (2011). Seeing or hearing the pianists? A synopsis of an early audiovisual perception experiment and a replication. *Musicae Scientiae*, *15*(3), 324–342. <https://doi.org/10.1177/1029864911410955>
- Burger, B., Saarikallio, S., Luck, G., Thompson, M. R., & Toiviainen, P. (2013). Relationships between perceived emotions in music and music-induced movement. *Music Perception*, *30*(5), 517–533. <https://doi.org/10.1525/mp.2013.30.5.517>
- Burger, B., & Toiviainen, P. (2013). *MoCap Toolbox—A Matlab toolbox for computational analysis of movement data*. 10th Sound and Music Computing Conference, SMC 2013, Stockholm, Sweden. <https://jyx.jyu.fi/handle/123456789/42837>
- Camurri, A., Mazarino, B., Ricchetti, M., Timmers, R., & Volpe, G. (2004). Multimodal analysis of expressive gesture in music and dance performances. In A. Camurri & G. Volpe (Eds.), *Gesture-based communication in human-computer interaction* (Vol. 2915, pp. 20–39). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-24598-8_3
- Castellano, G., Mortillaro, M., Camurri, A., Volpe, G., & Scherer, K. (2008). Automated analysis of body movement in emotionally expressive piano performances. *Music Perception*, *26*(2), 103–119. <https://doi.org/10.1525/mp.2008.26.2.103>
- Clarke, E. F. (1995). Expression in performance: Generativity, perception, and semiosis. In J. Rink (Ed.), *The practice of performance: Studies in musical interpretation* (pp. 21–54). Cambridge University Press.
- Clarke, E. F. (2005). *Ways of listening: An ecological approach to the perception of musical meaning*. Oxford University Press.
- Dahl, S., & Friberg, A. (2007). Visual perception of expressiveness in musicians' body movements. *Music Perception*, *24*(5), 433–454. <https://doi.org/10.1525/mp.2007.24.5.433>
- Dalla Bella, S., Peretz, I., Rousseau, L., & Gosselin, N. (2001). A developmental study of the affective value of tempo and mode in music. *Cognition*, *80*(3), B1–B10. [https://doi.org/10.1016/S0010-0277\(00\)00136-0](https://doi.org/10.1016/S0010-0277(00)00136-0)
- Davidson, J. W. (1993). Visual perception of performance manner in the movements of solo musicians. *Psychology of Music*, *21*(2), 103–113. <https://doi.org/10.1177/030573569302100201>
- Eaves, D. L., Griffiths, N., Burrige, E., McBain, T., & Butcher, N. (2020). Seeing a drummer's performance modulates the subjective experience of groove while listening to popular music drum patterns. *Musicae Scientiae*, *24*(4), 475–493. <https://doi.org/10.1177/1029864919825776>
- Eerola, T., Friberg, A., & Bresin, R. (2013). Emotional expression in music: Contribution, linearity, and additivity of primary musical cues. *Frontiers in Psychology*, *4*. <https://doi.org/10.3389/fpsyg.2013.00487>
- Eerola, T., Lartillot, O., & Toiviainen, P. (2009). Prediction of multidimensional emotional ratings in music from audio using multivariate regression models. In *10th International Society for Music Information Retrieval Conference (ISMIR 2009)*, 621–626.
- Eerola, T., & Vuoskoski, J. K. (2011). A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music*, *39*(1), 18–49. <https://doi.org/10.1177/0305735610362821>
- Gabrielsson, A., & Juslin, P. N. (1996). Emotional expression in music performance: Between the performer's intention and the listener's experience. *Psychology of Music*, *24*(1), 68–91. <https://doi.org/10.1177/0305735696241007>
- Gibson, J. J. (1979). *The ecological approach to visual perception*. L. Erlbaum.
- Hurvich, C. M., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, *76*(2), 297–307. <https://doi.org/10.1093/biomet/76.2.297>
- Juslin, P. N. (2000). Cue utilization in communication of emotion in music performance: Relating performance to perception. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(6), 1797–1812. <https://doi.org/10.1037/0096-1523.26.6.1797>
- Juslin, P. N. (2003). Five facets of musical expression: A psychologist's perspective on music performance. *Psychology of Music*, *31*(3), 273–302. <https://doi.org/10.1177/03057356030313003>
- Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels,

- same code? *Psychological Bulletin*, 129(5), 770–814. <https://doi.org/10.1037/0033-2909.129.5.770>
- Juslin, P. N., & Lindström, E. (2010). Musical expression of emotions: Modelling listeners' judgements of composed and performed features. *Music Analysis*, 29(1–3), 334–364. <https://doi.org/10.1111/j.1468-2249.2011.00323.x>
- Krippendorff, K. (2011). *Computing Krippendorff's Alpha-Reliability*. Retrieved from https://repository.upenn.edu/asc_papers/43
- Lange, E. B., & Frieler, K. (2018). Challenges and opportunities of predicting musical emotions with perceptual and automatized features. *Music Perception*, 36(2), 217–242. <https://doi.org/10.1525/mp.2018.36.2.217>
- Lartillot, O., Toiviainen, P., & Eerola, T. (2008). A Matlab toolbox for music information retrieval. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme, & R. Decker (Eds.), *Data analysis, machine learning and applications* (pp. 261–268). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-78246-9_31
- Laukka, P., & Gabrielsson, A. (2000). Emotional expression in drumming performance. *Psychology of Music*, 28(2), 181–189. <https://doi.org/10.1177/0305735600282007>
- McNeill, D. (2007). *Gesture and thought*. University of Chicago Press.
- Palmer, C. (1997). Music performance. *Annual Review of Psychology*, 48(1), 115–138. <https://doi.org/10.1146/annurev.psych.48.1.115>
- Peretz, I., Gagnon, L., & Bouchard, B. (1998). Music and emotion: Perceptual determinants, immediacy, and isolation after brain damage. *Cognition*, 68(2), 111–141. [https://doi.org/10.1016/S0010-0277\(98\)00043-2](https://doi.org/10.1016/S0010-0277(98)00043-2)
- Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6), 1161–1178. <https://doi.org/10.1037/h0077714>
- Shapiro, L. (2014). *The Routledge handbook of embodied cognition*. Routledge & CRC Press.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420. <https://doi.org/10.1037/0033-2909.86.2.420>
- Smit, E. A., Milne, A. J., Sarvasy, H. S., & Dean, R. T. (2022). Emotional responses in Papua New Guinea show negligible evidence for a universal effect of major versus minor music. *PLOS ONE*, 17(6), e0269597. <https://doi.org/10.1371/journal.pone.0269597>
- Thompson, M. R., & Luck, G. (2012). Exploring relationships between pianists' body movements, their expressive intentions, and structural elements of the music. *Musicae Scientiae*, 16(1), 19–40. <https://doi.org/10.1177/1029864911423457>
- Vieillard, S., Peretz, I., Gosselin, N., Khalifa, S., Gagnon, L., & Bouchard, B. (2008). Happy, sad, scary and peaceful musical excerpts for research on emotions. *Cognition & Emotion*, 22(4), 720–752. <https://doi.org/10.1080/02699930701503567>
- Vuoskoski, J. K., Thompson, M. R., Clarke, E. F., & Spence, C. (2014). Crossmodal interactions in the perception of expressivity in musical performance. *Attention, Perception, & Psychophysics*, 76(2), 591–604. <https://doi.org/10.3758/s13414-013-0582-2>
- Vuoskoski, J. K., Thompson, M. R., Spence, C., & Clarke, E. F. (2016). Interaction of sight and sound in the perception and experience of musical performance. *Music Perception*, 33(4), 457–471. <https://doi.org/10.1525/mp.2016.33.4.457>
- Vuoskoski, J. K., Zickfeld, J. H., Alluri, V., Moorthigari, V., & Seibt, B. (2022). Feeling moved by music: Investigating continuous ratings and acoustic correlates. *Plos One*, 17(1), e0261151. <https://doi.org/10.1371/journal.pone.0261151>
- Wanderley, M. M. (2002). Quantitative analysis of non-obvious performer gestures. *Gesture and Sign Language in Human-Computer Interaction: International Gesture Workshop, GW 2001 London, UK, April 18–20, 2001 Revised Papers*, 241–253.
- Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, 9(4), 625–636. <https://doi.org/10.3758/BF03196322>
- Windsor, W. L. (2011). Gestures in music-making: Action, information and perception. In A. Gritten & E. King (Eds.), *New perspectives in music and gesture* (pp. 45–66). Farnham.
- Zentner, M., Grandjean, D., & Scherer, K. R. (2008). Emotions evoked by the sound of music: Characterization, classification, and measurement. *Emotion*, 8(4), 494. <https://doi.org/10.1037/1528-3542.8.4.494>



V

**EXPLORING RELATIONS BETWEEN BIG FIVE
PERSONALITY TRAITS AND MUSICAL EMOTIONS
EMBODIED IN SPONTANEOUS DANCE**

by

Juan Ignacio Mendoza, Birgitta Burger, & Geoff Luck, 2022

Psychology of Music

<https://doi.org/10.1177/03057356221135355>

Exploring relations between Big Five personality traits and musical emotions embodied in spontaneous dance

Psychology of Music

1–19

© The Author(s) 2022



Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/03057356221135355

journals.sagepub.com/home/pom



Juan Ignacio Mendoza Garay¹ , Birgitta Burger²
and Geoff Luck¹

Abstract

We explored the hypothesis that musical emotions are embodied differentially by people according to their personality. Nine hundred and fifty two individuals completed the Big Five personality inventory. A subset of 60 participants were asked to spontaneously move to 30 short musical stimuli while being recorded with a motion-capture system. The musical stimuli were separately rated for perceived emotions. Embodied musical emotions were evaluated as the correlation between features derived from the motion-capture data and the mean ratings of perceived emotions. Correlations between embodied musical emotions and personality traits provided tentative support for our hypothesis. A series of linear regression analyses revealed that scores on Openness and Agreeableness were most strongly, and Neuroticism and Conscientiousness most weakly, predicted by embodied musical emotions. Overall, our results offer tentative support for the existence of differential relationships between embodied musical emotions and personality, and describe statistical models that might be empirically tested in future studies.

Keywords

music, emotion, perception, dance, movement, personality, individual differences

When someone spontaneously dances to music, their movement and posture may reflect characteristics of the music. In other words, dance may *embody* musical properties, beat, and rhythm usually being the most evident (Burger et al., 2014, 2018; Burger, Thompson, et al., 2013; Toiviainen et al., 2010). Likewise, more complex and abstract characteristics of music may be embodied in dance, such as emotional content (Van Dyck et al., 2017). However, it is likely that not everyone will embody

¹University of Jyväskylä, Jyväskylä, Finland

²Universität Hamburg, Hamburg, Germany

Corresponding author:

Juan Ignacio Mendoza Garay, University of Jyväskylä, 40014 Jyväskylä, Finland.

Email: juigmend@student.jyu.fi

musical emotions in the same way and to the same extent. One might ask, therefore, whether people have individual characteristics that affect how and how much they will embody musical emotions. Alternatively, might these individual characteristics be predicted by the way people embody musical emotions? These two questions are facets of the same relationship. Beyond scientific curiosity, acknowledging the effect of individual differences on bodily expression of musical emotions may be relevant to activities that involve music and dance with a variety of people, such as teaching music and dance, and the use of dance and music in therapeutic settings.

The relationships between people's individual characteristics and musical emotions have been studied in various ways. Usually, such individual characteristics are examined in terms of personality traits and measured with a questionnaire, even though other characteristics may be considered, such as gender or cultural background. Musical emotions have been observed in terms of perceived emotions in music and felt emotions when listening to music or emotions induced by music. These can be evaluated with a questionnaire or by measurement of physiological activity. Gerra et al. (1998), for example, reported an experiment in which participants were presented with classical and electronic dance music, while several physiological and psychological measurements were recorded. Results showed that after listening to both kinds of music there was a change in emotional state. However, only after listening to electronic dance music was it observed that changes toward a negative mood and release of stress hormones had a positive correlation with "harm-avoidance" and a negative correlation with "novelty-seeking" temperaments of Cloninger's personality scales (Cloninger, 1987). Another study, conducted by Park et al. (2013), looked at how "Big Five" personality traits (Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness) modulate neural correlates of musical emotion processing. Participants completed the NEO-FFI questionnaire of Big Five personality traits (McCrae & Costa, 2004) and, while being scanned by a Magnetic Resonance Imaging device, listened to music expressing different emotions. The results showed significant correlations between brain activity and both Neuroticism and Extraversion as a response to music expressing happiness and fear, respectively.

Other studies on the relationships between musical emotions and personality have evaluated musical emotions, perceived or felt, solely by means of self-report. Vuoskoski and Eerola (2011b) conducted an experiment in which participants completed the Big Five Inventory (BFI) personality questionnaire (John & Srivastava, 1999), the POMS-A questionnaire to evaluate mood (Terry et al., 2003), and rated music in terms of perceived discrete emotions (happiness, sadness, anger, fear, and tenderness). Ratings of perceived sadness correlated positively with Neuroticism and negatively with all other traits except Conscientiousness. Also, mood was associated with mood-congruent biases in perceived emotions, moderated by Extraversion. In another experimental study, Vuoskoski and Eerola (2011a) asked participants to complete the BFI and to rate emotions felt when listening to music. Ratings in terms of three-dimensional affect—Valence (i.e., positive vs. negative), Energy, and Tension—yielded more consistent and differentiated responses compared with discrete emotions. However, the relation between personality and music-induced emotions was stronger for discrete emotions. In addition, Extraversion was significantly correlated with experienced happiness, sadness, and tenderness. In a similar vein, Liljeström et al. (2012) asked participants to listen to music and indicate if it was familiar, how much they liked it, which emotions they felt and how intensely. Participants also completed the NEO-PI-R questionnaire for Big Five traits (Costa & McCrae, 1992). A positive correlation was observed between Neuroticism and experience of negative emotions, while for all other traits the correlation was negative. This is consistent with the results of Vuoskoski and Eerola (2011b). Furthermore, the correlation between personality traits and ratings of emotion intensity was moderately positive for Agreeableness, Extraversion, and Openness, negligible for Conscientiousness, and weakly negative for Neuroticism.

The studies mentioned in the previous paragraphs reveal distinct relationships between personality traits and the perception and feeling of emotions in music. Trait Openness is a special case as it has been suggested to be related to transient emotional responses (colloquially referred to as “chills”) to music and other expressions facilitating aesthetic experiences (McCrae, 2007). Nusbaum and Silvia (2011) tested this hypothesis in an experiment, and found that Openness was the only Big Five trait that significantly predicted such responses as an effect of music listening. Furthermore, Silvia et al. (2015), found a significant and moderate correlation between Openness and the feeling of a *profound experience* (also referred to as “awe”) when listening to music, while the correlation with the other traits was much lower.

Although perception and experience of musical emotions may be observed by means of physiological measures and self-report questionnaires, still other possibilities exist. For example, one might examine characteristics of spontaneous movement to music, such as how such movements embody the music’s emotional content. Burger, Saarikallio, et al. (2013) asked participants to spontaneously move to music (i.e., dance) while they were recorded with a motion-capture system. Bodily features were computed, for example, the torso’s tilt and rotation, floor area used and acceleration of different body parts. Another group of participants rated the perceived emotional content of the same music in terms of both dimensional affect—Arousal (i.e., emotional activation) and Valence—and discrete emotions Happiness, Anger, Sadness, and Tenderness. A correlational analysis between bodily features and emotion ratings revealed significant relations between them, even though the two datasets were collected independently of each other and from different groups of participants. Using the same data, Burger, Polet, et al. (2013) found a mediation effect of emotion ratings on the relation between bodily features and features of the music, such as energy and activity in the low- and high-frequency ranges, attack time, and note density. That study also used Big Five personality scores of the dancing participants and found a moderation effect of Extraversion on the relation between head acceleration and the activity of low-frequency audio (i.e., low-frequency spectral flux). Furthermore, Conscientiousness was found to be a significant moderator of the relation between note density and movement fluidity.

Using the same motion-capture and personality data as Burger, Polet, et al. (2013), Luck et al. (2010) found that Extraversion was directly related to the level of overall acceleration. This was later confirmed in a study with different data by Carlson et al. (2016). The latter study also found that responsiveness to changes in tempo correlated positively with Conscientiousness and negatively with Extraversion. This suggests that conscientious people were compelled to follow tempo accurately while extraverts preferred to divert and follow their own beat. Bamford and Davidson (2019) measured the time to entrainment (i.e., the alignment of the periodicity of the movement of the body to the beat of the music) of participants that had completed the BEAS Big Five questionnaire (DeYoung et al., 2007) and the Empathy Quotient questionnaire (Wakabayashi et al., 2006). Results showed that Empathy and Agreeableness correlated negatively with time to entrainment. In other words, the more Empathic or Agreeable a person is, the faster (and arguably more easily) they will align their dancing motion with the beat of the music.

While these studies have identified significant relations between dancing motion and personality, the predictive power of the produced models and correlations is at best modest. However, a more recent study by Agrawal et al. (2020) traded the interpretability of bodily features for greater prediction power. Instead of using bodily features extracted by manual selection (e.g., speed or acceleration of body parts, or the distance or angle between them) or by dimensionality reduction (e.g., vertical or lateral speed), they used the covariance among the speed of body parts. As a result, predictions for all Big Five personality traits were remarkably close to their scores as measured by a questionnaire.

To summarize, there exists ample evidence that embodied responses to music are related to personality traits and to musical emotions. What is less clear, however, is how embodied musical

emotions (i.e., the extent to which emotions are reflected in dancing) are related to personality. Consequently, we explored the hypothesis that musical emotions are embodied differentially by people according to their personality. We conducted a detailed and systematic analysis of relationships between personality traits and embodied musical emotions utilizing the following three different kinds of data: (1) Personality scores (measured by responses to the BFI) of individuals who moved spontaneously to music, (2) motion-capture data recorded from the same individuals, and (3) perceptual data concerning emotions perceived in the music they danced to. These data had been previously used in other studies as follows: Motion-capture data, personality data and ratings of perceived emotions had been used by Burger, Polet, et al. (2013), and Burger and Toiviainen (2020b); motion-capture and personality data had been used by Luck et al. (2010, 2014); motion-capture data and ratings of perceived emotions had been used by Burger, Saarikallio, et al. (2013); only motion-capture data had been used by Burger et al. (2014), Burger, Thompson, et al. (2013), Burger and Toiviainen (2020a), and Saarikallio et al. (2013).

Method

Participants

For the spontaneous dancing task, 60 participants took part (43 females, 17 males, ages from 19 to 32, $M = 24$, $SD = 3.3$). These individuals were selected from a total of 952 persons who had previously completed the BFI questionnaire. The scores of the selected participants were evenly distributed along the scales (i.e., a continuum covering low, middle, and high scores for each personality trait). All of them were students from different faculties of the University of Jyväskylä and all except two were of Finnish nationality. Six participants had received formal music education and four had received formal dance education. For the rating of perceived emotions, a different group of 34 participants took part (17 females, 17 males, ages from 21 to 47, $M = 25.7$, $SD = 5.9$), all musicology students of the University of Jyväskylä, familiar with research of music and emotions, and of Finnish nationality. The inter-rater agreement of perceived emotions was high and the ratings correlated significantly with movement features of the participants that danced spontaneously (see Burger, Saarikallio, et al., 2013). The university granted approval for non-invasive and non-inductive experiments involving human participants. All participants gave verbal consent to the procedures after they were explained to them. No records were kept linking participants' identity and experimental data.

Stimuli

The stimuli used for the spontaneous dancing task were 30 audio excerpts of different popular music genres, chosen to have a variety of rhythmic complexity and tempo. All excerpts were 28-s long, solely instrumental, and had a binary meter. Further information can be found in the Appendix. At the beginning of each stimulus, one extra second of audio was added, composed by a sine tone at 300Hz lasting 0.5 s followed by silence. The same excerpts were used for the rating of perceived emotions, although they were trimmed to 15 s by removing the first and last 6.5 s. This was done to shorten the duration of the data collection session thus reducing risk of fatigue in the participants.

Apparatus

For the task involving spontaneous dancing, bodily posture was recorded with a Qualisys Pro Reflex optical motion-capture system composed of eight infrared video cameras. This system

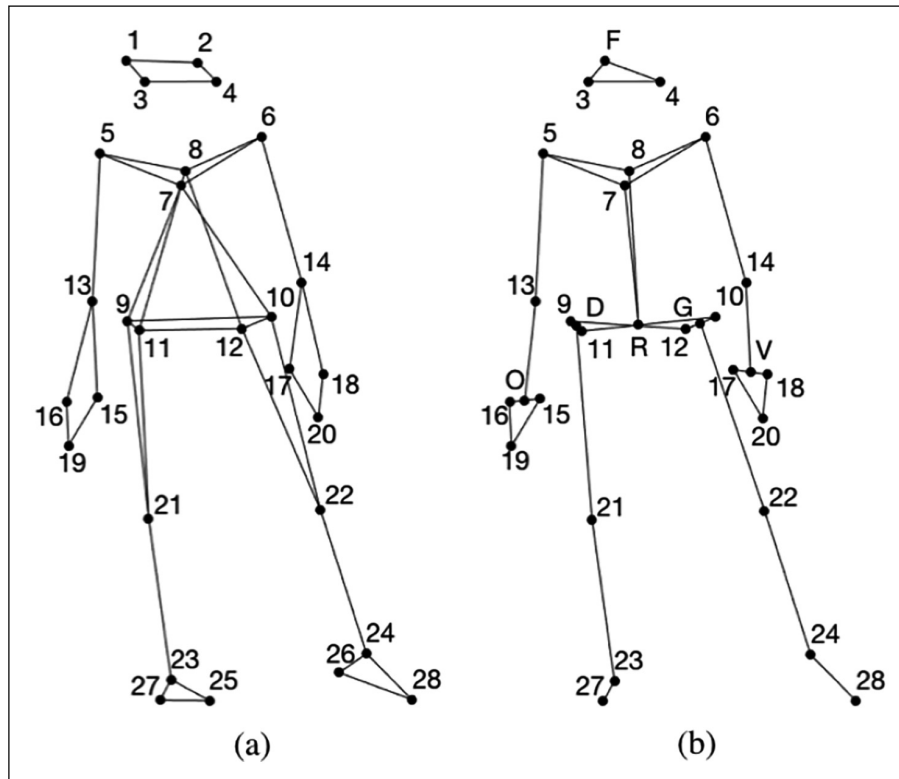


Figure 1. (a) Original Markers and (b) Retained Original Markers (Numbered) and Virtual Markers (Letters).

tracked at 120 frames per second the three-dimensional position of 28 reflective markers attached to the body. The ratings of perceived emotions were written on paper. For these two tasks, stimuli were played on studio monitor loudspeakers and presented in random order.

Procedure

For the spontaneous dancing task, each participant was recorded in a separate session in which they were asked to move to the music “in a way that feels natural.” In the session, a motion-capture recording was made for each stimulus. For the rating of perceived emotions, data collection took place in two sessions, each comprising half of the participants. They were asked to rate perceived emotions in music on seven-point scales for dimensional affect in terms of Arousal and Valence, and for discrete emotions Happiness, Anger, Sadness, and Tenderness (see Eerola & Vuoskoski, 2011). The random order of stimuli was different for each session.

Preprocessing of motion-capture data

Reflective markers were visualized as skeletons (Figure 1(a)) and rendered as video for visual inspection. Missing or corrupted data did not exceed 3 s, and were reconstructed with an automatic procedure (Tits et al., 2018) whose parameters were adjusted manually using video of the

Table 1. Kinematic (K) and Non-Kinematic (NK) Bodily Features.

Type	Description	Feature markers	Reference markers
K	All markers	All markers	–
K	Head	8, 3, 4, F	8, 7, R
K	Shoulders	8, 5, 6	8, 7, R
K	Arms	Right: 5, 13, O Left: 6, 14, V	Right: 5, 7, 8 Left: 6, 7, 8
K	Hands	Right: O, 15, 19, 16 Left: V, 17, 20, 18	Right: O, 13, 5 Left: V, 14, 6
K	Legs	Right: D, 21, 23, 27 Left: G, 22, 24, 28	Right: D, 10, 12 Left: G, 9, 11
K	Root horizontal plane	R	–
K	Root vertical axis	R	–
NK	Torso tilt	R, 8, G	R, G
NK	Torso rotation	R, G	R, 7, 8
NK	Hands distance	O, V	–
NK	Elbows distance	13, 14	–
NK	Feet distance	23, 24	–
NK	Area	R	–

Note. K: kinematic; NK: non-kinematic.

reconstructed skeletons. A new set of markers was derived by retaining some of the originals and producing additional virtual markers by averaging some combinations of the original (Figure 1(b)). The new configuration of markers was designed so that there would be enough points for a reference plane to translate and rotate body parts to their own local coordinate system, as appropriate to each bodily feature (explained below). Furthermore, markers at the heels were removed as they did not provide further information than the markers at the ankles and tip of the feet. All motion-capture recordings were trimmed to the duration of the musical excerpts.

Kinematic and non-kinematic bodily features were computed using the motion-capture data. They were crafted to represent a variety of aspects of bodily motion and posture (see Table 1). Features that represent movement of individual bodily parts use subsets of markers locked to a local coordinate system defined by a reference plane. This reduces collinearity among features, which is desirable when they are used as regressors in linear models (see below). Collinearity arises because parts of the body will move when another part moves. For example, an arm will move as the torso moves, and the torso will move along with the whole body. By locking the arm to a local coordinate system, the only motion remaining is that of the arm alone. To wit, features representing bodily parts locked to a local coordinate system are more intuitively interpreted, as they are related more to muscle activation than to mere displacement. Kinematic features were speed, acceleration (acc.), jerk (jrk.), and the square of speed (speed²) of markers as detailed in Table 1, resulting in 32 features. The square of speed was included as a supplemental measure for kinetic energy. Kinetic energy is half the mass multiplied by the squared velocity. As the mass is constant, it can be omitted from the equation. The Euclidean norm was computed for each feature, resulting in a single value for data corresponding to each motion-capture recording. For the six non-kinematic features, the median was computed to obtain a single value for each recording. The median was used as a magnitude measure as it is less sensitive to outliers compared with other average measures such as mean or mode. An exception is the feature “Torso rotation,” for which the standard deviation was computed. The feature “Area” is defined as the smallest rectangular area of a marker projected to the horizontal plane in a moving window of

4 s with a hop of one frame. The result was 38 bodily-feature values for each motion-capture recording.

Analyses

The bodily-feature data for each participant were composed of 38 features, each having 30 values (one for each motion-capture recording and musical stimulus). Each feature was correlated with each of the six ratings of perceived emotions, resulting in 228 values of embodied emotions for each participant. Rank correlation was used because an inspection of the correlated data revealed that they were not consistently normally distributed or linear. Kendall's *tau-b* rank correlation was preferred as its interpretation is straightforward. Then, two analyses were performed to assess the relations between embodied emotions and personality traits.

Analysis 1 comprised the rank correlation between personality traits and embodied emotions. This was achieved by computing Kendall's *tau-b* between scores of each personality trait and each of the 228 embodied emotions for all participants. Additionally, personality traits were correlated with six aggregated embodied emotions, one for each perceived emotion. These were computed as the sum of absolute values of embodied emotions corresponding to the same perceived emotion.

Analysis 2 consisted in ordinary least-squares linear regression models that predict each personality trait. Each model had regressors consisting of a combination of embodied emotions, but corresponding to only one bodily feature. This way, it was possible to examine the effect of each bodily part separately, though at the cost of reduced prediction power compared with using combinations of bodily features. The following equation describes a regression model for one personality trait:

$$P = \beta_0 + \beta_1 R_{valence} + \beta_2 R_{arousal} + \beta_3 R_{happiness} + \beta_4 R_{sadness} + \beta_5 R_{anger} + \beta_6 R_{tenderness} + E$$

where P is scores of a personality trait, β are regression coefficients with β_0 being constant, R are embodied emotions (i.e., regressors), and E is error. In contrast to the aggregation measure of the first analysis, this is a weighted linear combination. All 63 possible combinations of perceived emotions for 38 embodied emotions (one for each bodily feature) resulted in 2,394 models for each personality trait.

Instead of selecting models by their statistical significance, *relevance* was assessed empirically by comparing the cross-validated error of a data model and the error of a null model. This has the advantage that there is no need to arbitrarily set a significance threshold (typically p -value less than .05). For each model, three-fold cross-validated Root Mean Squared Error ($RMSE_{CV}$) with 10^5 Monte Carlo realizations, and relevance measure

$$\Delta RMSE = RMSE_{null} - RMSE_{CV}$$

were computed, where $RMSE_{null}$ is the error of a null model for each personality trait. A positive value for $\Delta RMSE$ indicates that the model is relevant, as it performs better than the null model and vice versa.

Results

The results of the first analysis reveal potential though weak relationships between each Big Five personality trait and the embodiment of each rated emotion by each bodily feature. Table 2 contains the rank correlations for all relationships with a p -value below .05, out of the 1,140 produced values. This threshold is used only to tabulate a subset of the results. The p -values were

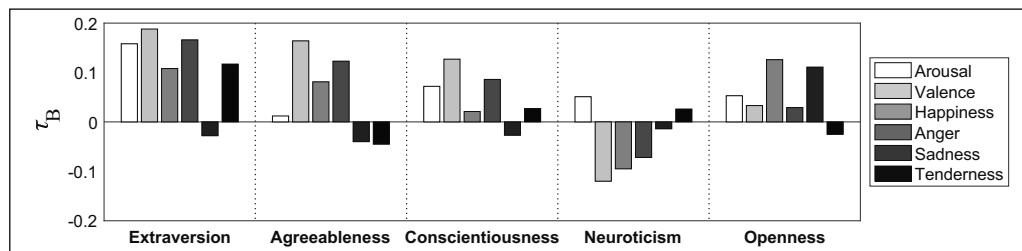


Figure 2. Kendall's Rank Correlation (τ_B) Between Personality Traits and Aggregated Embodied Emotions.

not adjusted to control for multiple comparisons as, due to the large number of values, any adjustment procedure would reduce statistical power to the extent that no results would remain significant. The interested reader may find further information about the appropriateness of adjustments for multiple comparisons in the work of Rothman (1990) and Althouse (2016). The p -values of rank correlations between personality traits and aggregated embodied emotions are shown in Table 3. These values are relatively high even though they were not adjusted. Figure 2 shows the corresponding correlation scores, which show tendencies that are weak and, as seen in Table 3, have poor statistical significance.

The second analysis produced 11,970 linear regression models. Figure 3(a) shows the relevance measure Δ RMSE clustered by personality trait, for all models performing better than the null model. A quick visual inspection reveals that the highest values are for Openness, followed by Agreeableness and then Extraversion. The greater number of relevant models are for Agreeableness, followed by Openness and Extraversion. Conscientiousness and Neuroticism have both the weakest values and smallest number of models. Figure 3(b) shows only models whose regressors are correlations between a bodily feature and any combination of ratings for only dimensional affect. Notably, none of these models for Openness perform better than the null model, and most models for Agreeableness perform better than models for the other traits. Figure 3(c) shows only models whose regressors are correlations between a bodily feature and any combination of ratings for only discrete emotions. In this case, the pattern is similar to when all regressors are allowed, but the best performing models for Agreeableness are not as strong as for regressors considering only dimensional affect or for all models. This is consistent with Vuoskoski and Eerola's (2011a) finding that, regarding music-induced emotions, discrete emotions have stronger relationships to individual differences than dimensional affect. When any combination of regressors for discrete emotions and dimensional affect is allowed, then the maximum Δ RMSE for Extraversion is significantly higher than when either only dimensional affect or discrete emotions are considered.

Tables 4 to 8 contain information about the models with positive and highest Δ RMSE for each bodily feature. The R^2 metric shows the performance of models fitted to the full data and the F -test p -value indicates the statistical significance of the fit. It is important to bear in mind firstly that the R^2 metric is not used here as a measure of prediction power. Instead, it is used as an intuitive way of understanding the closeness of the fit to the observed values, as the metric has a unit maximum and is zero when it matches a null model. In fact, in the tables, it is possible to see that there is no perfect rank correspondence between R^2 and RMSE_{CV} . This difference is due to the high variability of the errors from the cross-validation folds. Therefore, models for which that difference is higher may be less representative of the underlying phenomenon. Note that because models have been sorted by Δ RMSE (and therefore also by RMSE_{CV}), the rank correspondence between R^2 and *adjusted* R^2 is exact, therefore the latter was not tabulated. As for the first analysis, p -values were not adjusted for multiple comparisons.

Table 2. Kendall's Rank Correlation (τ_B) Between Personality Traits and Embodiment of Emotions.

Personality trait	Bodily feature	Emotion rating	τ_B	p
Extraversion	Feet distance	Sad.	.239	.009*
	Hands, speed ²	Sad.	.215	.019
	Area	Val.	.208	.023
	Hands, speed	Sad.	.204	.025
	Arms, acc.	Aro.	.200	.028
	Legs, acc.	Aro.	.194	.034
	Arms, jrk.	Ten.	-.192	.035
	Head, jrk.	Ten.	-.191	.036
	Legs, jrk.	Ten.	-.191	.037
	Head, acc.	Ten.	-.190	.037
	Root horiz., speed	Ang.	-.188	.039
	Root horiz., speed ²	Val.	.189	.039
	Area	Ten.	.188	.039
	Legs, jrk.	Ang.	.187	.040
	Root horiz., jrk.	Ten.	-.186	.042
	Root horiz., speed	Val.	.185	.043
	Root vertic., jrk.	Ten.	-.185	.043
	Shoulders, jrk.	Ten.	-.185	.043
	Arms, acc.	Ten.	-.184	.044
	Root horiz., speed	Ten.	.181	.048
Agreeableness	All mk., speed ²	Ten.	.246	.007*
	Hands distance	Hap.	.240	.009*
	Torso rotation	Ten.	.218	.017
	Hands distance	Val.	.219	.017
	All mk., speed ²	Val.	.194	.034
	Torso rotation	Ang.	-.192	.035
	Torso rotation	Val.	.192	.036
	Hands distance	Ang.	-.191	.037
	Root horiz., speed ²	Val.	.184	.044
	Root horiz., speed ²	Ten.	.184	.044
Conscientiousness	Elbows distance	Ang.	-.192	.034
	Hands distance	Hap.	.190	.037
	Root horiz., speed	Ten.	.184	.042
	All mk., speed ²	Sad.	.182	.045
	Legs, jrk.	Aro.	.182	.045
	Root horiz., jrk.	Aro.	.179	.048
Neuroticism	Hands distance	Val.	-.203	.026
	Root vertic., jrk.	Sad.	-.237	.009*
Openness	Root vertic., acc.	Sad.	-.207	.022
	Root vertic., jrk.	Hap.	.196	.032
	Shoulders, speed ²	Val.	.187	.040
	Root vertic., speed	Sad.	-.186	.041
	Head, jrk.	Sad.	-.185	.042
	Shoulders, speed	Val.	.184	.043
	Area	Sad.	-.182	.045
	Area	Val.	.182	.046
	Head, acc.	Sad.	-.180	.047
	Torso rotation	Sad.	-.181	.047

Note. Only correlations with non-adjusted $p < .05$ are tabulated.

* $p < .01$.

Table 3. Non-Adjusted p -Values for the Rank Correlation Between Personality Traits and Aggregated Embodied Emotions.

Emotion rating	Extraversion	Agreeableness	Conscientiousness	Neuroticism	Openness
Arousal	.083	.903	.428	.574	.565
Valence	.039*	.072	.160	.188	.720
Happiness	.236	.374	.818	.294	.163
Anger	.068	.177	.344	.428	.749
Sadness	.764	.668	.769	.878	.220
Tenderness	.198	.622	.769	.778	.788

* $p < .05$.

It is worthwhile to note that the embodiments of Valence and Tenderness by bodily feature “Root vertical, acceleration” are regressors for a model that is the most relevant for Extraversion and also one of the most relevance for Openness. A closer inspection of this model reveals that the coefficients for the regression fit are very similar for both personality traits. The coefficients for Extraversion are 21.857, 41.409, -48.245 and the coefficients for Openness are 34.011, 48.321, -47.252 , where the first values are the constant and the remaining are the regressors’ coefficients. However, both the fit and the prediction power of this model are greater for Extraversion, as revealed by its R^2 and $\Delta RMSE$ values.

The model selection method presented earlier is focused on the prediction performance of models, allowing the best combinations of regressors for each model, with the sole constraint of having regressors for only one bodily feature for each model. However, this means that regressors are removed from a model only to improve its prediction power. Even when the models have been cross-validated, it is possible that regressors remain in the model because of their noise instead of their true explanatory power. Therefore, it is convenient to also examine only those models that have all regressors for each type of emotional rating and also those models that have all emotional ratings. Table 9 shows all relevant models that have regressors considering all emotional ratings, all dimensional affect ratings or all discrete emotions ratings. In these conditions, no relevant models are found for Extraversion or Neuroticism. Additionally, all except the following bodily features appear in regressors for at least one relevant model: All markers’ speed, All markers’ jerk, Shoulders’ acceleration, and Head’s squared speed. These bodily features do not appear in Table 2. Hence, these features may be irrelevant.

Discussion

We explored relationships between the Big Five personality traits and musical emotions embodied in spontaneous movement to music. Embodied emotions were evaluated as the rank correlation between characteristics of spontaneous movement to music and perceived musical emotions in the music moved to. Two analyses were carried out. Analysis 1 consisted of rank-correlating personality and embodied emotions. Analysis 2 involved creating multiple linear models that predicted personality traits with the weighted scores of embodied emotions. The purpose of these analyses was to evaluate and highlight relationships that might be empirically tested in future studies.

Analysis 1 revealed moderately weak monotonic relations between bodily features and perceived emotions for all personality traits. Conscientiousness and Neuroticism had the weakest of such relations when considering the rank-correlation values and the number of bodily features involved. The relations between emotions embodied by aggregated bodily features and personality traits were rather weak (see Figure 2), and statistical analysis provided limited evidence that such

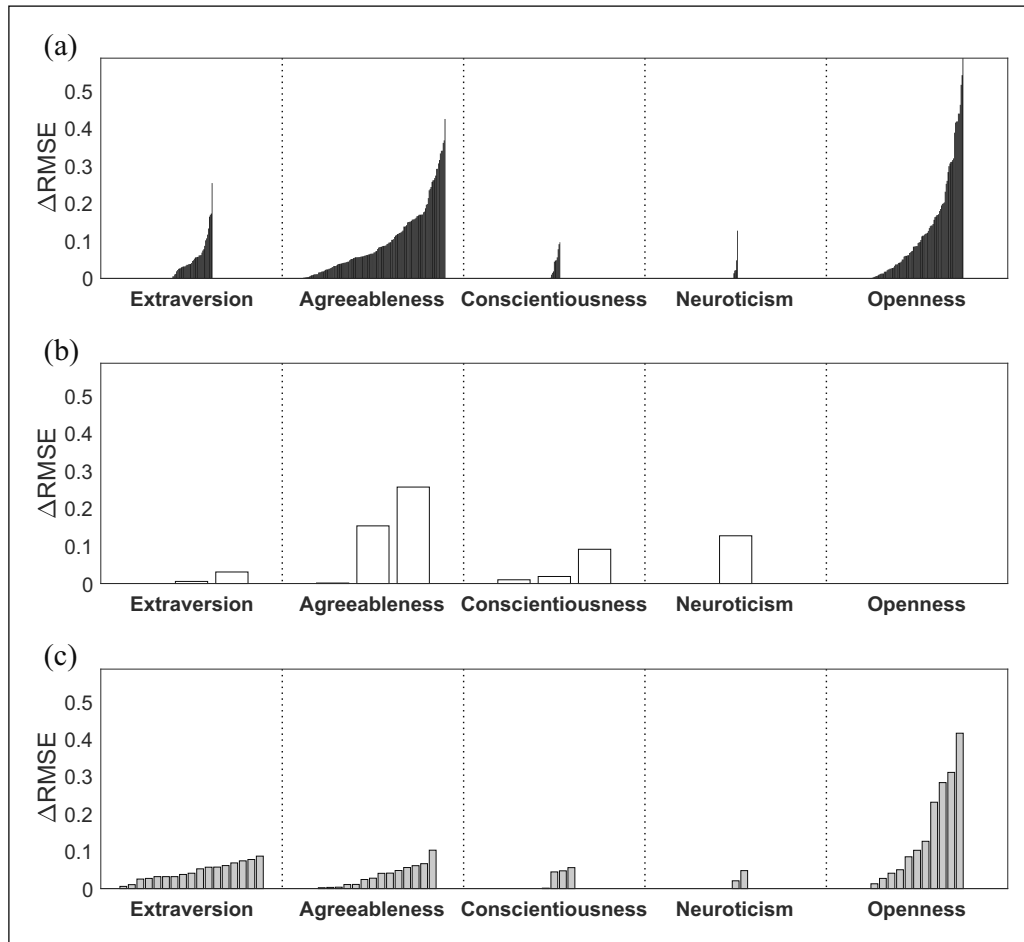


Figure 3. Relevant Models Sorted by Relevance, With Any Combination of Regressors Considering Correlation With All Emotion Ratings (a), Only Dimensional Affect (b), and Only Discrete Emotions (c).

relationships exist (see Table 3). The p -values shown in Table 3 may be useful in future research, for example, to re-test the highest correlation obtained or to discard the aggregation method.

The results of Analysis 2, however, revealed that predictions using linear regression models that are better than the null model are possible for all personality traits, albeit they range from weak to moderate. Regarding the performance of all regression models, the strongest predictions were found for Openness, followed by Agreeableness and Extraversion. The predictions for Conscientiousness and Neuroticism were the weakest and, as in Analysis 1, this general assessment considers the goodness-of-fit of models and the number of models involved. Regarding models that have regressors for either discrete emotions, dimensional affect or both (Table 9), the strongest predictions were still found for Openness, followed by Agreeableness and Conscientiousness. In this case, no relevant models were produced for traits Extraversion or Neuroticism.

In both analyses, distinct bodily features were found to embody musical emotions correlating with or predicting personality traits. No single bodily feature embodying a musical emotion was a high rank correlate of all personality traits. Likewise, no single combination of bodily

Table 4. Best Relevant Models for Each Bodily Feature, Predicting Extraversion.

Model		R^2	p	RMSE _{CV}	Δ RMSE
Bodily feature	Emotion ratings				
Root vertic., acc.	Val. Ten.	.173	.004	6.030	0.254
Root horiz., jrk.	Val. Sad. Ten.	.171	.014	6.119	0.164
Root vertic., jrk.	Val. Ten.	.144	.012	6.151	0.133
Arms, acc.	Ang. Ten.	.125	.022	6.197	0.087
Feet distance	Sad.	.092	.019	6.209	0.074
Hands, speed ²	Sad.	.087	.022	6.215	0.068
Head, acc.	Ang. Ten.	.114	.031	6.222	0.061
Hands, speed	Sad.	.082	.027	6.227	0.057
Arms, speed ²	Aro. Ang. Sad. Ten.	.195	.016	6.227	0.056
Arms, speed	Aro. Hap. Ang. Ten.	.194	.017	6.235	0.048
All mk., speed ²	Aro. Ang. Sad.	.186	.009	6.237	0.046
Legs, jrk.	Ang.	.071	.039	6.242	0.041
Hands distance	Aro. Val. Sad.	.156	.022	6.245	0.039
Legs, acc.	Aro.	.075	.034	6.253	0.031
Head, jrk.	Ang. Ten.	.106	.041	6.258	0.025

Note. RMSE_{CV}: cross-validated root mean squared error; RMSE: root mean squared error.

Table 5. Best Relevant Models for Each Bodily Feature, Predicting Agreeableness.

Model		R^2	p	RMSE _{CV}	Δ RMSE
Bodily feature	Emotion ratings				
Root horiz., speed ²	Aro. Hap. Ten.	.281	<.001	4.431	0.425
Hands distance	Aro. Ten.	.212	.001	4.564	0.292
All mk., speed ²	Aro. Sad. Ten.	.242	.001	4.594	0.262
Root horiz., speed	Aro. Hap. Ten.	.233	.002	4.595	0.261
Area	Aro. Val. Sad.	.222	.003	4.678	0.178
Root horiz., jrk.	Aro. Val. Sad.	.198	.006	4.706	0.150
Legs, jrk.	Aro. Val. Sad.	.182	.010	4.731	0.125
Legs, speed ²	Aro. Val. Sad.	.169	.015	4.741	0.115
Hands, acc.	Val. Hap.	.119	.027	4.770	0.086
Hands, jrk.	Val. Hap.	.121	.026	4.770	0.086
Feet distance	Hap. Sad.	.122	.025	4.790	0.066
Root horiz., acc.	Val. Ang.	.131	.018	4.800	0.056
Shoulders, jrk.	Val. Sad. Ten.	.149	.028	4.823	0.033
Torso tilt	Val. Ang.	.118	.028	4.836	0.02
Root vertic., speed ²	Aro. Val. Sad.	.142	.034	4.852	0.003
Torso rotation	Ten.	.090	.020	4.853	0.003
All mk., acc.	Val. Sad.	.118	.028	4.854	0.002

Note. RMSE_{CV}: cross-validated root mean squared error; RMSE: root mean squared error.

features embodying any combination of perceived emotions predicted all personality traits. However, some bodily features embodying an emotion did highly correlate with, or combine linearly to predict, more than one personality trait. For both analyses, the most relevant bodily

Table 6. Best Relevant Models for Each Bodily Feature, Predicting Conscientiousness.

Model		R^2	p	RMSE _{CV}	Δ RMSE
Bodily feature	Emotion ratings				
Head, speed	Aro. Ang.	.122	.025	6.292	0.096
Area	Aro. Val. Hap.	.167	.016	6.311	0.078
Elbows distance	Ang. Sad. Ten.	.158	.021	6.333	0.056
Root horiz., jrk.	Aro.	.064	.050	6.370	0.019
Legs, jrk.	Aro.	.062	.054	6.379	0.01
Root horiz., speed	Ten.	.056	.070	6.387	0.001

Note. RMSE_{CV}: cross-validated root mean squared error; RMSE: root mean squared error.

Table 7. Best Relevant Models for Each Bodily Feature, Predicting Neuroticism.

Model		R^2	p	RMSE _{CV}	Δ RMSE
Bodily feature	Emotion ratings				
Hands distance	Val.	.089	.021	5.983	0.127
Elbows distance	Hap.	.095	.017	6.062	0.048

Note. RMSE_{CV}: cross-validated root mean squared error; RMSE: root mean squared error.

Table 8. Best Relevant Models for Each Bodily Feature, Predicting Openness.

Model		R^2	p	RMSE _{CV}	Δ RMSE
Bodily feature	Emotion ratings				
Area	Aro. Val. Hap. Sad.	.308	<.001	5.942	0.587
Shoulders, speed	Val. Ten.	.176	.004	6.277	0.252
Head, jrk.	Val. Ten.	.166	.006	6.334	0.194
Root horiz., speed ²	Aro. Val. Hap. Sad.	.216	.009	6.349	0.179
Root vertic., acc.	Val. Ten.	.157	.008	6.357	0.171
Legs, speed	Val. Ang.	.145	.012	6.360	0.169
Arms, jrk.	Val. Ten.	.156	.008	6.384	0.145
Shoulders, speed ²	Val. Ten.	.149	.010	6.387	0.141
Root vertic., jrk.	Sad.	.095	.017	6.426	0.102
Root vertic., speed ²	Val. Ten.	.134	.017	6.434	0.094
Shoulders, jrk.	Val. Ten.	.138	.014	6.444	0.084
Legs, acc.	Aro. Val. Hap. Sad.	.177	.028	6.461	0.068
Root horiz., speed	Val. Ang.	.108	.039	6.486	0.043
Torso rotation	Sad.	.076	.033	6.501	0.027
Root vertic., speed	Val. Ten.	.109	.037	6.526	0.002

Note. RMSE_{CV}: cross-validated root mean squared error; RMSE: root mean squared error.

features involved in the tested relationships were identified. These results may serve as hypotheses in further investigation.

The tabulated rank correlations (Table 2) and models with all regressors allowed (Tables 4 to 8) exhibit an overall distribution considering the number of tabulated bodily features and the strength

Table 9. Relevant Models That Have Regressors Considering All Emotional Ratings (AVHAST), All Dimensional Affect Ratings (AV), or All Discrete Emotions Ratings (HAST).

Personality trait	Bodily feature	Emotion ratings	R^2	p	RMSE _{CV}	Δ RMSE
Agreeableness	Hands distance	AV	.191	.002	4.599	0.257
Agreeableness	Root horiz., speed ²	AVHAST	.307	.003	4.687	0.169
Conscientiousness	Head, speed	AV	.118	.028	6.298	0.091
Openness	Area	HAST	.245	.003	6.218	0.310
Openness	Area	AVHAST	.316	.002	6.213	0.315

Note. RMSE_{CV}: cross-validated Root Mean Squared Error; RMSE: root mean squared error; AVHAST: Arousal, Valence, Happiness, Anger, Sadness, and Tenderness; AV: Arousal, Valence; HAST: Happiness, Anger, Sadness, Tenderness.

of tested relationships. This distribution may be summarized in two clusters of personality traits. The first cluster is composed by Openness, Agreeableness, and Extraversion, whereas the second cluster consists of Conscientiousness and Neuroticism. However, this two-cluster pattern does not hold for rank correlation between personality traits and aggregated embodied emotions (Table 3). It also does not hold when regressors are forcibly embodiments of either dimensional affect, discrete emotions, or both (Table 9), and when the number of correct predictions is evaluated using the threshold method (Table 9). A special case is trait Openness, for which all assessments of prediction by regression models are the strongest (Tables 4 to 9). Also, trait Neuroticism is a special case, as it is related to the lowest number of bodily features (Tables 2 and 4 to 9).

The two-cluster pattern with the special case for Neuroticism is remarkably consistent with the results obtained by the meta-analysis conducted by Barańczuk (2019). That study found that lower levels of Neuroticism and higher levels of all other traits were associated with greater typically adaptive emotion regulation strategies (reappraisal, problem solving, and mindfulness) and lower typically maladaptive emotion regulation strategies (avoidance and suppression). In particular, the relationship between suppression of expression of emotions was found to be non-significant for Neuroticism and inverse for all other traits, Conscientiousness and Neuroticism being the weakest, and Extraversion being the strongest. While Extraversion does not appear in this study as a special case of strong direct relationships with embodied emotions, the relations observed for all other traits suggest that the embodiment of emotions may be related to the suppression of emotion. The relationships for Extraversion might have been affected by unobserved factors.

When the special case for Openness and the special case for Neuroticism are integrated to the two-cluster pattern, it is possible to observe that similar results were obtained by previous studies that have investigated the strength of correlation between personality traits and music preference or liking, across a variety of music genres and cultural backgrounds (Brown, 2012; Delsing et al., 2008; Dobrota & Reić Ercegovac, 2015; Dunn et al., 2011; Ercegovac et al., 2015; Fricke & Herzberg, 2017; Nave et al., 2018; Rentfrow & Gosling, 2003; Schäfer & Mehlhorn, 2017; Upadhyay et al., 2017; Vuoskoski & Eerola, 2011b; Zweigenhaft, 2008). These studies have found Openness to have the strongest correlations with music liking, followed by Agreeableness and Extraversion. Conscientiousness and Neuroticism were found to have the weakest correlations with music liking. Carlson et al. (2017) reported similar results, with the difference that correlation strength for trait Extraversion was much lower, closer to Neuroticism and Conscientiousness. Other studies measuring correlation between Big Five personality traits and preference for music have found distinct stronger correlations for Openness, and the other traits having weaker correlations (Cleridou & Furnham, 2014; Langmeyer et al., 2012; Upadhyay et al., 2017). Additionally, these observations are consistent with previous research that has found evidence that the preference for music is related to the emotional content of

music (Hunter et al., 2011; Ladinig & Schellenberg, 2012; Naser & Saha, 2021; Schäfer & Sedlmeier, 2011) or that has hypothesized it based on the relation between preference and bodily features of spontaneous dance (Luck et al., 2014). Likewise, Openness, Agreeableness, and Extraversion have been found to be associated with positive correlations between music preference and the strength of emotional response to music, Openness having the strongest association (Liljeström et al., 2012; Nusbaum & Silvia, 2011).

The highest association between liking for music and perceived emotions, being for trait Openness, is consistent with results obtained by previous studies that have investigated a variety of related phenomena. Openness has been found to correlate positively with chills as an effect of listening to music (McCrae, 2007), awe for music (Silvia et al., 2015), and also with the direct relation between liking for sad music and emotions elicited by sad music (Vuoskoski et al., 2012). Trait Openness has consistently been thought to be related to the experience of complex and strong emotions as a result of sensitivity to aesthetic experiences (Reisenzein & Weber, 2009; Terracciano et al., 2003). These observations about Openness may explain the results of this study showing stronger relations to embodied emotions compared with other traits when a number of embodied emotions are combined as regressors in a linear model, in contrast to embodied emotions as correlators.

Other patterns similar to the ones found in this study may be found in previous studies on the relations between Big Five personality traits and trait Empathy. The special case of trait Neuroticism may be related to trait Empathy as Melchers et al. (2016) and Bamford and Davidson (2019) have observed direct correspondence between Empathy Quotient and all Big Five traits, except Neuroticism that had inversely weak and insignificant correspondence, respectively. Those studies and the work by DeYoung et al. (2010) have found Agreeableness to be strongly and directly related with trait Empathy, which might contribute to explain the high-rank correlations and linear fits found for Agreeableness in this study. Also Conscientiousness exhibiting weak relations with embodied emotions may be explained by this trait being the only Big Five trait not related to emotional dispositions (Reisenzein & Weber, 2009).

The comparison made of results of this study with previous studies, show that for each Big Five personality trait there may be underlying, moderating, or mediating factors of the embodiment of emotions. It may be worthwhile to test each of these as separate hypotheses in future research. While the regression models in this study predict personality traits separately, the relationship also holds, at least theoretically, in the opposite direction. This means linear models with personality traits as regressors, predicting embodied emotions.

Regarding limitations of this study, the first and most evident is the sample size and composition. The statistical power of this study is substantially limited by the amount of collected data and the generalizability is limited by demographics. A straightforward solution to increase statistical power is to replicate the experiment using the same stimuli and ratings of perceived emotions. However, such replication will require that participants have different characteristics than this study, like nationality and distribution of gender. This may not be an easy study to conduct, as data collection is costly. Apart from expensive laboratory equipment, substantial time is spent in motion-capture recordings and responding questionnaires. It is also challenging to find a homogeneous sample of Big Five traits and then having respondents to participate in motion-capture sessions. Therefore, each replication may not have by itself considerable statistical power and it may take several replications, by different laboratories, to achieve robust conclusions. The second limitation of this study is that ratings of perceived emotions were done by a separate group of participants. There is an advantage of this, however, as these ratings are a controlled variable, meaning that the same measure is used for all participants and could be used in replications of the experiment as a standard. Nonetheless, it may be worthwhile to explore the possibility of improving predictions of personality by using ratings made by the dancing participants. Also ratings of felt emotions may give further insights to the relation between personality and embodied

emotions, as it has been observed that induced emotions affect dancing characteristics even if the music to which is being danced is emotionally neutral (Van Dyck et al., 2013). A further modification could be the use of self-chosen stimuli as this has been observed eliciting more intense emotional responses to music (Liljeström et al., 2012). The third limitation of this study is that the usability of linear regression models with few regressors is limited to observation or explanation of phenomena as they yield predictions with limited power. The predictive power of the best models found in this study may be not suitable, for example, to make clinical diagnosis or other kind of prediction that requires a very high degree of accuracy.

To conclude, this exploratory study provides empirical and quantitative evidence tentatively supporting the hypothesis that the emotional content of music expressed by spontaneous dance has distinct relationships with the Big Five personality traits. If one assumes that the observed characteristics of spontaneous dance are a result of the emotional content of music, and in light of previous research, it is possible to conjecture that the underlying causes of the embodiment of emotions are emotional dispositions, including empathy, as well as liking of the music being danced to. This study provides a foundation upon which future research could be built. Specifically, such work could empirically test the relationships and statistical models described herein, helping to further advance our understanding of the complex interrelationships between personality, movement, and emotion.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Juan Ignacio Mendoza Garay  <https://orcid.org/0000-0003-3996-7537>

References

- Agrawal, Y., Jain, S., Carlson, E., Toiviainen, P., & Alluri, V. (2020). Towards multimodal MIR: Predicting individual differences from music-induced movement. *arXiv preprint arXiv 2007.10695*. <https://arxiv.org/pdf/2007.10695.pdf>
- Althouse, A. D. (2016). Adjust for multiple comparisons? It's not that simple. *The Annals of Thoracic Surgery*, *101*(5), 1644–1645.
- Bamford, J. M. S., & Davidson, J. W. (2019). Trait Empathy associated with agreeableness and rhythmic entrainment in a spontaneous movement to music task: Preliminary exploratory investigations. *Musicae Scientiae*, *23*(1), 5–24.
- Barańczuk, U. (2019). The five factor model of personality and emotion regulation: A meta-analysis. *Personality and Individual Differences*, *139*, 217–227.
- Brown, R. A. (2012). Music preferences and personality among Japanese university students. *International Journal of Psychology*, *47*(4), 259–268.
- Burger, B., London, J., Thompson, M. R., & Toiviainen, P. (2018). Synchronization to metrical levels in music depends on low-frequency spectral components and tempo. *Psychological Research*, *82*(6), 1195–1211.
- Burger, B., Polet, J., Luck, G., Thompson, M. R., Saarikallio, S., & Toiviainen, P. (2013, June 11–15). *Investigating relationships between music, emotions, personality, and music-induced movement* [Conference session]. The 3rd International Conference on Music & Emotion, Jyväskylä, Finland.
- Burger, B., Saarikallio, S., Luck, G., Thompson, M. R., & Toiviainen, P. (2013). Relationships between perceived emotions in music and music-induced movement. *Music Perception: An Interdisciplinary Journal*, *30*(5), 517–533.
- 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.

- Burger, B., Thompson, M. R., Luck, G., Saarikallio, S. H., & Toiviainen, P. (2014). Hunting for the beat in the body: On period and phase locking in music-induced movement. *Frontiers in Human Neuroscience*, 8, Article 903.
- Burger, B., & Toiviainen, P. (2020a). Embodiment in electronic dance music: Effects of musical content and structure on body movement. *Musicae Scientiae*, 24(2), 186–205. <https://doi.org/10.1177/1029864918792594>
- Burger, B., & Toiviainen, P. (2020b). See how it feels to move: Relationships between movement characteristics and perception of emotions in dance. *Human Technology*, 16(3), 233–256. <https://doi.org/10.17011/ht/urn.202011256764>
- Carlson, E., Burger, B., London, J., Thompson, M. R., & Toiviainen, P. (2016). Conscientiousness and extraversion relate to responsiveness to tempo in dance. *Human Movement Science*, 49, 315–325.
- Carlson, E., Saari, P., Burger, B., & Toiviainen, P. (2017). Personality and musical preference using social-tagging in excerpt-selection. *Psychomusicology: Music, Mind, and Brain*, 27(3), 203–212.
- Cleridou, K., & Furnham, A. (2014). Personality correlates of aesthetic preferences for art, architecture, and music. *Empirical Studies of the Arts*, 32(2), 231–255.
- Cloninger, C. R. (1987). A systematic method for clinical description and classification of personality variants: A proposal. *Archives of General Psychiatry*, 44(6), 573–588.
- Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI): Professional manual*. Psychological Assessment Resources.
- Delsing, M. J., Ter Bogt, T. F., Engels, R. C., & Meeus, W. H. (2008). Adolescents' music preferences and personality characteristics. *European Journal of Personality*, 22(2), 109–130.
- DeYoung, C. G., Hirsh, J. B., Shane, M. S., Papademetris, X., Rajeevan, N., & Gray, J. R. (2010). Testing predictions from personality neuroscience: Brain structure and the big five. *Psychological Science*, 21(6), 820–828.
- DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: 10 aspects of the Big Five. *Journal of Personality and Social Psychology*, 93(5), 880–896.
- Dobrota, S., & Reić Ercegovic, I. (2015). The relationship between music preferences of different mode and tempo and personality traits—implications for music pedagogy. *Music Education Research*, 17(2), 234–247.
- Dunn, P. G., de Ruyter, B., & Bouwhuis, D. G. (2011). Toward a better understanding of the relation between music preference, listening behavior, and personality. *Psychology of Music*, 40(4), 411–428.
- Eerola, T., & Vuoskoski, J. K. (2011). A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music*, 39(1), 18–49.
- Ercegovic, I. R., Dobrota, S., & Kušević, D. (2015). Relationship between music and visual art preferences and some personality traits. *Empirical Studies of the Arts*, 33(2), 207–227.
- Fricke, K. R., & Herzberg, P. Y. (2017). Personality and self-reported preference for music genres and attributes in a German-speaking sample. *Journal of Research in Personality*, 68, 114–123.
- Gerra, G., Zaimovic, A., Franchini, D., Palladino, M., Giucastro, G., Reali, N., . . . Brambilla, F. (1998). Neuroendocrine responses of healthy volunteers to techno-music: Relationships with personality traits and emotional state. *International Journal of Psychophysiology*, 28(1), 99–111.
- Hunter, P. G., Schellenberg, E. G., & Stalinski, S. M. (2011). Liking and identifying emotionally expressive music: Age and gender differences. *Journal of Experimental Child Psychology*, 110(1), 80–93.
- John, O. P., & Srivastava, S. (1999). *The Big-Five trait taxonomy: History, measurement, and theoretical perspectives* (Vol. 2, pp. 102–138). University of California.
- Ladinig, O., & Schellenberg, E. G. (2012). Liking unfamiliar music: Effects of felt emotion and individual differences. *Psychology of Aesthetics, Creativity, and the Arts*, 6(2), 146–154.
- Langmeyer, A., Guglhör-Rudan, A., & Tarnai, C. (2012). What do music preferences reveal about personality? *Journal of Individual Differences*, 33(2), 119–130.
- Liljeström, S., Juslin, P. N., & Västjäll, D. (2012). Experimental evidence of the roles of music choice, social context, and listener personality in emotional reactions to music. *Psychology of Music*, 41(5), 579–599.
- Luck, G., Saarikallio, S., Burger, B., Thompson, M., & Toiviainen, P. (2010). Effects of the Big Five and musical genre on music-induced movement. *Journal of Research in Personality*, 44(6), 714–720.

- Luck, G., Saarikallio, S., Burger, B., Thompson, M., & Toiviainen, P. (2014). Emotion-driven encoding of music preference and personality in dance. *Musicae Scientiae*, 18(3), 307–323.
- McCrae, R. R. (2007). Aesthetic chills as a universal marker of openness to experience. *Motivation and Emotion*, 31(1), 5–11.
- McCrae, R. R., & Costa P. T., Jr. (2004). A contemplated revision of the NEO Five-Factor Inventory. *Personality and Individual Differences*, 36(3), 587–596.
- Melchers, M. C., Li, M., Haas, B. W., Reuter, M., Bischoff, L., & Montag, C. (2016). Similar personality patterns are associated with empathy in four different countries. *Frontiers in Psychology*, 7, Article 290.
- Naser, D. S., & Saha, G. (2021). Influence of music liking on EEG based emotion recognition. *Biomedical Signal Processing and Control*, 64, 102251. <https://doi.org/10.1016/j.bspc.2020.102251>
- Nave, G., Minxha, J., Greenberg, D. M., Kosinski, M., Stillwell, D., & Rentfrow, J. (2018). Musical preferences predict personality: Evidence from active listening and Facebook likes. *Psychological Science*, 29(7), 1145–1158.
- Nusbaum, E. C., & Silvia, P. J. (2011). Shivers and timbres: Personality and the experience of chills from music. *Social Psychological and Personality Science*, 2(2), 199–204.
- Park, M., Hennig-Fast, K., Bao, Y., Carl, P., Pöppel, E., Welker, L., . . . Gutyrchik, E. (2013). Personality traits modulate neural responses to emotions expressed in music. *Brain Research*, 1523, 68–76.
- Reisenzein, R., & Weber, H. (2009). Personality and emotion. In P. J. Corr & G. Matthews (Eds.), *The Cambridge handbook of personality psychology* (pp. 54–71). Cambridge University Press.
- Rentfrow, P. J., & Gosling, S. D. (2003). The do re mi's of everyday life: The structure and personality correlates of music preferences. *Journal of Personality and Social Psychology*, 84(6), 1236–1256.
- Rothman, K. J. (1990). No adjustments are needed for multiple comparisons. *Epidemiology*, 1(1), 43–46.
- Saarikallio, S., Luck, G., Burger, B., Thompson, M. R., & Toiviainen, P. (2013). Dance moves reflect current affective state illustrative of approach-avoidance motivation. *Psychology of Aesthetics, Creativity, and the Arts*, 7(3), 296–305. <https://doi.org/10.1037/a0032589>
- Schäfer, T., & Mehlhorn, C. (2017). Can personality traits predict musical style preferences? A meta-analysis. *Personality and Individual Differences*, 116, 265–273.
- Schäfer, T., & Sedlmeier, P. (2011). Does the body move the soul? The impact of arousal on music preference. *Music Perception*, 29(1), 37–50.
- Silvia, P. J., Fayn, K., Nusbaum, E. C., & Beaty, R. E. (2015). Openness to experience and awe in response to nature and music: Personality and profound aesthetic experiences. *Psychology of Aesthetics, Creativity, and the Arts*, 9(4), 376–384.
- Terracciano, A., McCrae, R. R., Hagemann, D., & Costa, P. T., Jr. (2003). Individual difference variables, affective differentiation, and the structures of affect. *Journal of Personality*, 71(5), 669–704.
- Terry, P. C., Lane, A. M., & Fogarty, G. J. (2003). Construct validity of the profile of mood states—Adolescents for use with adults. *Psychology of Sport and Exercise*, 4(2), 125–139.
- Tits, M., Tilmann, J., & Dutoit, T. (2018). Robust and automatic motion-capture data recovery using soft skeleton constraints and model averaging. *PLOS ONE*, 13(7), Article e0199744. <https://doi.org/10.1371/journal.pone.0199744>
- Toiviainen, P., Luck, G., & Thompson, M. R. (2010). Embodied meter: Hierarchical eigenmodes in music-induced movement. *Music Perception*, 28(1), 59–70.
- Upadhyay, D., Shukla, R., & Chakraborty, A. (2017). Factor structure of music preference scale and its relation to personality. *Journal of Indian Academy of Applied Psychology*, 43(1), 104–113.
- Upadhyay, D. K., Shukla, R., Tripathi, V. N., & Agrawal, M. (2017). Exploring the nature of music engagement and its relation to personality among young adults. *International Journal of Adolescence and Youth*, 22(4), 484–496.
- Van Dyck, E., Burger, B., & Orlandatou, K. (2017). The communication of emotions in dance. In M. Lesaffre, P.-J. Maes, & M. Leman (Eds.), *The Routledge companion to embodied music interaction* (pp. 122–130). Routledge.
- Van Dyck, E., Maes, P. J., Hargreaves, J., Lesaffre, M., & Leman, M. (2013). Expressing induced emotions through free dance movement. *Journal of Nonverbal Behavior*, 37(3), 175–190.
- Vuoskoski, J. K., & Eerola, T. (2011a). Measuring music-induced emotion: A comparison of emotion models, personality biases, and intensity of experiences. *Musicae Scientiae*, 15(2), 159–173.

- Vuoskoski, J. K., & Eerola, T. (2011b). The role of mood and personality in the perception of emotions represented by music. *Cortex*, 47(9), 1099–1106.
- Vuoskoski, J. K., Thompson, W. F., McIlwain, D., & Eerola, T. (2012). Who enjoys listening to sad music and why? *Music Perception*, 29(3), 311–317.
- Wakabayashi, A., Baron-Cohen, S., Wheelwright, S., Goldenfeld, N., Delaney, J., Fine, D., . . . Weil, L. (2006). Development of short forms of the Empathy Quotient (EQ-Short) and the Systemizing Quotient (SQ-Short). *Personality and Individual Differences*, 41(5), 929–940.
- Zweigenhaft, R. L. (2008). A do re mi encore: A closer look at the personality correlates of music preferences. *Journal of Individual Differences*, 29(1), 45–55.

Appendix. Musical Audio Excerpts Used as Stimuli.

	Artist	Song (album)	Start time	BPM*
1)	Alice Deejay	Better Off Alone (Who Needs Guitars Anyway?)	2:40 ^a	137
2)	Andre Visior	Speed Up	1:15	140
3)	Antibalas	Who is this America Dem Speak of Today? (Who Is This America?)	1:00	121
4)	Arturo Sandoval	A Mis Abuelos (Danzon)	1:53	108
5)	Baden Powell	Deixa (Personalidade)	1:11	100
6)	Brad Mehldau	Wave/Mother Nature's Son (Largo)	0:00	143
7)	Clifford Brown & %Max Roach	The Blues walk (Verve Jazz Masters, Vol. 44: Clifford Brown & Max Roach)	2:01	133
8)	Conjunto Imagen	Medley-Esencia de Guaguanco/ Sonero (Ayer, Hoy y Manana)	2:18	87
9)	Dave Hillyard & The Rocksteady 7	Hillyard Street (Playtime)	0:15	135
10)	Dave Weckl	Mercy, Mercy, Mercy (Burning for Buddy)	0:10	105
11)	Dave Weckl	Tower of Inspiration (Master Plan)	0:00	125
12)	DJ Shadow	Napalm Brain/Scatter Brain (Endroducing. . .)	3:29	73
13)	Gangster Politics	Gangster Politics (Guns & Chicks)	1:00	192
14)	Gigi D'Agostino	Blablabla (L'Amour Toujours)	0:00	133
15)	Herbie Hancock	Watermelon man (Cantaloupe Island)	0:00	132
16)	Horace Silver	The Natives Are Restless	0:00	139
17)	In Flames	Scream (Come Clarity)	0:00	100
18)	Jean Roch	Can You Feel it (Club Sounds Vol. 35)	0:33	126
19)	Johanna Kurkela	Hetki hiljaa (Hetki hiljaa)	3:22	122
20)	Juana Molina	Tres cosas (Tres Cosas)	0:00	110
21)	Kings of Leon	Closer (Only by the Night)	3:17	83
22)	Lenny Kravitz	Live (5)	3:02	84
23)	Martha & The Vandellas	Heat Wave (Heat Wave)	1:40	82
24)	Maynard Ferguson	Fireshaker (Live From San Francisco)	0:00	91
25)	MIA	20 Dollar (Kala)	0:17	120
26)	Nick Beat	Techno Disco	2:26	138
27)	Panjabi MC	Mundian To Bach Ke (Legalized)	0:47 ^b	98
28)	Patrick Watson	Beijing (Wooden Arms)	2:30	154
29)	The Rippingtons	Weekend in Monaco (Weekend in Monaco)	1:13	113
30)	Yuri Buenaventura	Salsa (Salsa Movie Soundtrack)	2:17	102

*Beats-per-minute

^a14s. repeated.

^b19s. repeated.



VI

**THE REARRANGER BALL: DELAYED GESTURAL CONTROL
OF MUSICAL SOUND USING ONLINE UNSUPERVISED
TEMPORAL SEGMENTATION**

by

Juan Ignacio Mendoza, 2023

To appear in
Proceedings of the Conference on
New Interfaces for Musical Expression

The Rearranger Ball: Delayed Gestural Control of Musical Sound using Online Unsupervised Temporal Segmentation

Juan Ignacio Mendoza
University of Jyväskylä
Finland
juigmend@student.jyu.fi

ABSTRACT

The state-of-the-art recognition of continuous gestures for control of musical sound by means of machine learning has two notable constraints. The first is that the system needs to be trained with individual example gestures, the starting and ending points of which need to be well defined. The second constraint is time required for the system to recognise that a gesture has occurred, which may prevent the quick action that musical performance typically requires. This article describes how a method for unsupervised segmentation of gestures, may be used for delayed gestural control of a musical system. The system allows a user to perform without explicitly indicating the starting and ending of gestures in order to train the machine learning algorithm. To demonstrate the feasibility of the system, an apparatus for control of musical sound was devised incorporating the time required by the process into the interaction paradigm. The unsupervised automatic segmentation method and the concept of delayed control are further proposed to be exploited in the design and implementation of systems that facilitate seamless human-machine musical interaction without the need for quick response time, for example when using broad motion of the human body.

Author Keywords

unsupervised, segmentation, music, gesture, controller

CCS Concepts

•Human → centered computing; •Computing methodologies → Machine learning; •Information systems → Music retrieval; •Applied computing → Performing arts;

1. INTRODUCTION

Musical instruments are usually designed to be controlled with fine movements of hands and fingers, as they afford precision and speed. These qualities are often described as the foundations of responsiveness, believed to be indispensable for musical expression. The instrument thus becomes an extension of the human body.

These ideas have permeated into the design of digital musical instruments (DMI) [15], and a response time approaching zero has become a standard goal [23, 9, 10]. The challenge extends to the design of DMI that recognise gestures “in the air”, using machine learning techniques. For example, a musician wears, holds or stands in front of, a device that may sense position (i.e., static gestures) or motion (i.e., continuous gestures). The musician makes a gesture in free space: describes a circle with the head, wiggles a hand, or stands in a particular pose. The DMI learns these gestures in a process called “training”, and it recognises them when they are performed. The recognition of a gesture can be mapped to a musical action, such as triggering a sound, activating an effect, etc. (e.g., [8]).

Two algorithms and variations of them have been extensively used to recognise continuous gestures, regardless of the sensing technology: Dynamic Time Warping (DTW) [7] and Hidden Markov Models (HMM) [1]. Both estimate the likelihood that a gesture being performed corresponds to a gesture that has been learned in the training. However, this likelihood may change while the gesture is executed, therefore recognition is only reliable after the gesture has been completed. This adds time to the recognition, arguably reducing responsiveness. In addition, training requires the beginning and ending of gestures to be explicit.

Given a stream of data from a sensor, individual gestures may be extracted by a process called “segmentation”, in which the start and ending points of gestures are identified. For example, when training the algorithm the user presses a button (e.g., [14]) or makes pauses between gestures (e.g., [16]). While this constraint has not prevented the use of the algorithms mentioned above in DMI, the ability of a machine to recognise and learn gestures without explicit training would open new avenues for human-machine musical interaction. Furthermore, the time required for the recognition of continuous gestures might not be a disadvantage if when designing a DMI we don’t hold the same standards of responsiveness as for the human voice or other non-electronic instruments. Consider that digital technologies have greatly expanded our possibilities for control of sound, far beyond what is possible with the human voice or with non-electronic devices. Why should we hold ourselves from exploring forms of gestural control that are not quick and precise, but instead slow and imprecise (i.e., delayed detection, perception, action, by the user and the automatic system) such as broad motion of the human body?

This article describes a system that was devised as a proof of concept towards exploring the feasibility of unsupervised learning of patterns in a continuous input signal, in a musical application that doesn’t require quick responsiveness. The system is conceptually a musical instrument in a broad sense, for it essentially allows a user to control sound.



Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). Copyright remains with the author(s).

NIME’23, 31 May–2 June, 2023, Mexico City, Mexico.

2. ONLINE UNSUPERVISED TEMPORAL SEGMENTATION

A signal may be segmented using the algorithm described by Foote [5], which has seen application in segmentation of musical audio and video [6, 21], dancing motion captured by an accelerometer [13], and daily activity recorded by wearable accelerometers [12, 17]. Its meta-parameters can be adjusted to detect boundaries of segments at different timescales. The cited sources described the use of the algorithm on recorded data. Conversely, Schätti [18] described an online version of the algorithm, that detects boundaries of audio data while the data is being produced. Later Mendoza [11] reported a study in which the algorithm’s segmentation of dancing motion captured by a hand-held accelerometer, was compared to manual segmentation of video recordings of the dancing. The meta-parameters were optimised for each accelerometry recording. The music used for dancing and the person doing the manual segmentation were the main factors affecting the quality of computed segmentation. These results suggest that the algorithm is suitable for gestural control of a DMI, albeit its meta-parameters might need contextual adjustment. Figure 1 succinctly illustrates the online segmentation procedure. It uses the same principle of buffering and computation of a local distance matrix, as described by Schätti [18] and Mendoza [11].

3. PROOF OF CONCEPT

3.1 Hardware

A polystyrene ball having 12 cm. of diameter was cut in half and the interior was carved to fit a Myo armband controller (Figure 2). The Myo was originally designed by Thalmic Labs to be worn on the forearm. It has several sensors, of which only its triaxial accelerometer was used in the system described here. The two halves of the ball are put together restoring the spherical shape, but it can be easily disassembled to recharge the battery of the Myo. The data from the sensors is broadcast in real time using the Bluetooth Low-Energy (BLE) specification. The BLE signal is captured by a computer nearby, and a piece of software written by Rodrigo Schramm¹ outputs the data in Open Sound Control (OSC) format to a User Datagram Protocol (UDP) port, where it can be accessed by other software. This controller was used for its convenience, as it was available to the researcher along with the software to get the data in real time.

3.2 Software

The segmentation procedure described in section 2 can detect in real-time boundaries between gestures performed with the hand-held controller continuously, without indicating their start or end. The effect of its meta-parameters are as follows: n sets the timescale of gestures to detect, n_{fit} sets the smoothness of the novelty score, θ is a factor of the maximum novelty score and sets a threshold below which novelty peaks are rejected (e.g., noise). A further meta-parameter was incorporated to prevent detection of segments of less than a given length n_{min} , such as transitions between gestures. The segmentation procedure, as

¹See [22]. Software available: <https://github.com/federicoVesi/KineToolbox/blob/master/input%20BML/DaemonMYO>

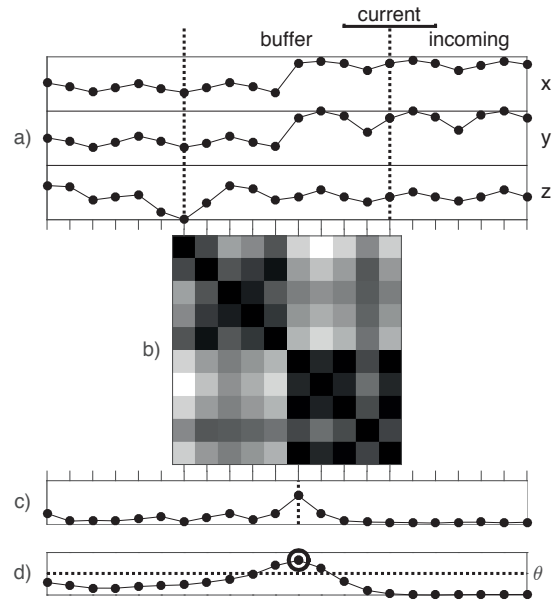


Figure 1: Online temporal segmentation. Horizontal axes represent time. (a) is accelerometer data composed of triaxial frames. (b) is a distance matrix of the data in the buffer having a length of n frames. Lighter shades represent more distance. (c) is a novelty score resulting from the correlation of the distance matrix with a gaussian-tapered checkerboard kernel. The vertical dotted line indicates the current result. (d) is the novelty score after smoothed by a gaussian filter of length n_{fit} , where θ is a threshold and the point in a circle is the selected peak indicating a boundary. Note that this visualisation shows (c) and (d) aligned in time, but in practice there will be a lag because of the filter. The total lag of the process is $(n + n_{fit})/2$ frames plus 3 frames for peak detection.



Figure 2: Left – Carved open polystyrene ball with the Myo armband in it. Right – Closed ball.

well as the musical application and its graphical user interface, were implemented in the Pure Data programming environment, which receives the accelerometry data using OSC as described in the previous subsection. The software is free and available (see Appendix).

The detected segments, each being a gesture, may be fed to a machine-learning process for training (i.e., gesture learning) and classification (i.e., gesture recognition). The DTW algorithm was chosen for this purpose, as it is available in the easy-to-use software Wekinator [4, 3], which communicates with Pure Data using OSC over a UDP port. However, another algorithm could be used (e.g., HMM). As with segmentation, the result of the recognition has lag due to buffering and latency due to logical processing.

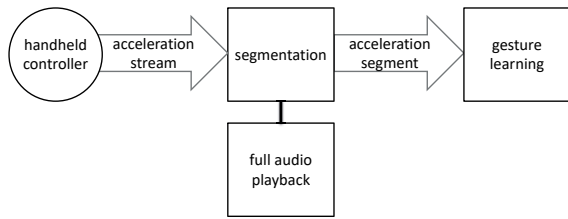


Figure 3: Cut stage

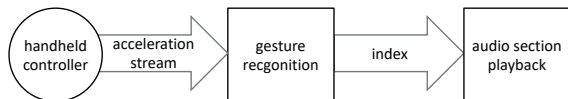


Figure 4: Perform stage

The segmentation and machine-learning processes are incorporated into a system that allows the user to reorder sections of an audio file. The use of the system has two stages: Cut and Perform. In the Cut stage (Figure 3) the audio file is played in its entirety while the user performs distinct gestures. The boundaries between gestures are detected in real time by the segmentation process and their time location is stored and labelled with a sequential index. The segments are fed as individual training examples to the gesture learning process. Also, in the graphical user interface a green vertical line is placed over a plot of the accelerometry signal, to indicate a successfully segmented gesture (Figure 5).

In the Perform stage (Figure 4) the gesture recognition process is continuously comparing the incoming accelerometry signal, to all the segments previously stored in the Cut stage. The segment that is closest to a stored one is deemed a match and its corresponding audio section plays in a loop. If a gesture different than the current is recognised, then the corresponding audio section will be played once the current audio section reaches its end.

3.3 Testing

During the implementation of the system, the author of this article conducted iterative testing using an upbeat electronic dance music piece, as it has been observed that this kind of music stimulates bodily motion [2]. Static gestures achieved by only changing the ball’s orientation, and gestures involving repetitive motion, were well segmented and recognised. Figure 6 shows a sequence of gestures that worked well with the following setting of meta-parameters, which was kept throughout the testing: $n = 80$, $n_{min} = 28$, and $n_{filt} = 24$, at a sampling rate of 20 frames per second yielding $lag = 55$ frames (0.4 seconds, not including logical processing latency), and $\theta = 0.03$. Parameters of the DTW process were also adjusted, but are not discussed as that algorithm is well documented [7, 3]. Since the ball is fully symmetrical, letters (A to F) were put on the orthogonal points to aid visually in manipulation. Later a small arrow was put next to each letter pointing to the next one (Figure 2, Right).

Additionally, extraction of features (e.g., amplitude, zero-crossings) from the triaxial accelerometry signal and its magnitude, was implemented. They did not improve segmentation but, because of being windowed processes, they did increase lag (i.e., frames needed for computation) and computation cost (i.e., logical processing). Therefore, devel-

opment and testing continued using only raw acceleration, to demonstrate what is possible without using extracted features.

When a functional version was completed, researchers and students of Musicology, Music Therapy and Music Education at the University of Jyväskylä were invited to evaluate the functionality of the system. With this group the following protocol was developed:

1. The researcher demonstrates the task comprising Cut and Perform stages, using the upbeat electronic dance music piece and the tested gestures sequence. The enclosed rectangle shown in Figure 6 is displayed on a paper.

2. The participant is invited to do the task. If in the Cut stage not all gestures were segmented successfully, the participant is invited to repeat the Cut, as many times as they want. Then, they are invited to try the Perform stage.

3. The participant is invited to freely improvise and/or to use another piece of music.

4. The participant is invited and encouraged to express their opinion on the experience. The researcher shall take observational notes such as number of gestures correctly segmented in a trial, comments and ideas expressed by and discussed with the participant, and if a new gesture is discovered.

The protocol described above was incorporated to a 7-hour presentation in an outreach event at the University of Jyväskylä. The following data was collected of 23 participants: age, gender, number of gestures successfully segmented consecutively from the first, and observations. Further notes were taken of more visitors. All participants used the upbeat electronic dance music, except one discarded for homogeneity. 17 participants (10 female, 7 male) performed the task as intended. Only six tried a second time, improving segmentation (see Figure 7). The medians of correctly segmented gestures was 4 for first time, 6 for second time and 5 for maxima. No correlation between number of correct segments and age or gender was observed. Most participants under 10 years old could not correctly perform all gestures, albeit they could successfully use the system by only changing the orientation of the ball.

3.4 Overall Assessment

Any set of orientations being different enough will work, but the 6 orthogonal orientations work flawlessly. Also, any combination and variation of repeated movements along the 3 orthogonal axes of the ball will work well. Sudden and energetic movements work best, as they are better measured by the accelerometer. Smooth movements are less likely to be detected by the system. Participants discovered a variety of gestures beyond those in the task. One of them is the “baby rocking”, consisting in holding the ball with two hands and moving it describing an upwards concave curve. Other semi-circular and circular motions, and “8” figures were successfully detected, inasmuch as the speed, and therefore radial acceleration, was powerful enough to produce a novelty score above the set threshold (θ).

If the transition from one gesture to the next is slow enough to have a duration equal or greater than n_{min} (minimum duration for gestures to be detected), the transition will be identified as a segment. In the Perform stage the system might get stuck looping these very short segments, due to the characteristics of the DTW algorithm (i.e., computation time is proportional to the length of the segment, parameter sensitivity). However, interestingly, two participants mentioned that they liked the result. One of them referred to it as “a DJ effect”. Another participant explored

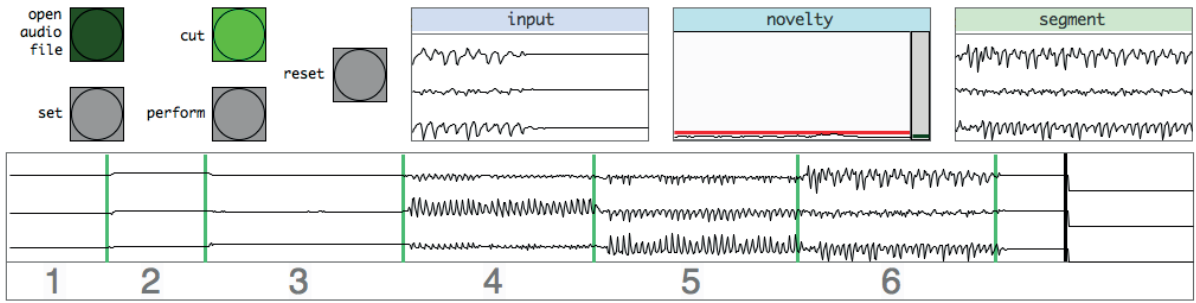


Figure 5: Graphical user interface

order	orientation	gesture	description
1	A	.	do nothing
		↶	rotate left
2	B	.	do nothing
		↑	rotate forward
3	C	.	do nothing
		↶	rotate left
4	D	↑↓↑↓...	move up-down
		↑	rotate forward
5	E	→	hit right
		↶	rotate left
6	F	↔↔	hit twice to each side
		.	do nothing
7	F	.	do nothing

Figure 6: Segmentation task

the possibility of not having to look at the ball when manipulating it. A discussion ensued leading to conclude that, since the ball is fully symmetric, it is not possible to be aware of its orientation without looking at it.

The task was challenging to different extents. Some participants wanted to try again to improve the number of correctly segmented gestures. All participants showed engagement and enjoyment. However, it is to expect that researchers and students have interest as the experience is related to their profession and studies. Likewise, visitors at the outreach event most probably attended because of curiosity.

4. DISCUSSION AND FUTURE WORK

The system described in this article demonstrates the feasibility of unsupervised learning of patterns in a continuous input signal, for gestural control, within a musical application. The process ineluctably produces a lagged response and therefore it is not suitable for the execution of fast notes

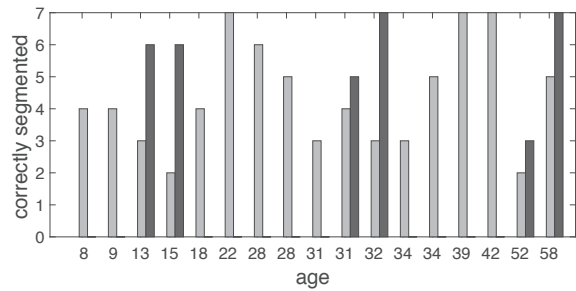


Figure 7: Data collected at the outreach event. Second trials are shown in darker shade.

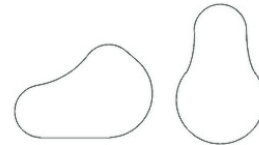


Figure 8: "Boot" form with rotational asymmetry. Left - lateral view. Right - zenithal view.

or rhythmic patterns. Nonetheless, the proposed musical application conforms to this constraint, supporting the concept of delayed control of musical sound. Participants of the assessment tended to regard the task as a challenge, which in combination with the discovery of new meaningful gestures, and the sense-making of the constraints, turned the experience into a ludic one. The system appears promising, offering opportunities for further research:

I. The reported assessment used recorded music, but any audio file may be used, and the meta-parameters may be tweaked for further exploration that may lead to unexpected yet interesting results.

II. The hand-held device will benefit from having rotational asymmetry, such that there is no need of looking at it for manipulation (Figure 8).

III. Using the raw accelerometry signal has established a baseline. Future research could evaluate the impact of features extracted from the raw signal. The computation of such features will impact the overall latency (lag plus logical processing), and the detection of novelty (and therefore the setting of meta-parameters) because of the information that the features carry.

IV. Incorporation of more sensors or sensing technologies other than accelerometry. Besides, several sensors may be used by more than one person simultaneously, as a group activity (e.g., [19, 20]).

V. Implementation of online multigranular segmentation, meaning the detection of gestural boundaries at different timescales.

VI. Current limitations to achieve **III**, **IV**, and **V**, are algorithmic complexity, processing power and software efficiency. Solutions may include low-level programming (possibly embedded software) and faster hardware (possibly parallel computing of several features and timescales).

VII. The setting of meta-parameters generalised well, which is unexpected as perceptual evaluations have suggested the adjustment of meta-parameters for each user [11]. A different setting might be needed when using other configurations of hardware, software, music, user, etc. Future research may assess the effects of meta-parameters on segmentation and user experience.

VIII. The methods described in this article have potential beyond the described application, in which the online segmentation procedure only contributes to display on the screen an indication when a gesture has been successfully segmented in the Cut stage. This allows the user, for example, to stop the Cut and restart if a gesture change was not detected. While this might be an advantage to the user, the online segmentation capability and its further possibilities for near-real-time interaction could be exploited more. For example, a musical system (e.g., a DMI, a sonic installation, a sonification) may learn gestures as they occur. This may be incorporated to interactive systems where both the user and the system discover and learn gestures at the same time, leading to a seamless process of human-machine musical interaction.

5. ETHICAL STANDARDS

All participants gave verbal informed consent for the use of their anonymous collected data, following the research ethics guidelines by the University of Jyväskylä.

6. REFERENCES

- [1] F. Bevilacqua, B. Zamborlin, A. Sypniewski, N. Schnell, F. Guédy, and N. Rasamimanana. Continuous realtime gesture following and recognition. In *Gesture in Embodied Communication and Human-Computer Interaction*, pages 73–84. Springer, 2010.
- [2] B. Burger and P. Toiviainen. Embodiment in electronic dance music: Effects of musical content and structure on body movement. *Musicae Scientiae*, 24(2):186–205, 2020.
- [3] R. Fiebrink. <http://www.wekinator.org/>
- [4] R. Fiebrink, D. Trueman, and P. R. Cook. A meta-instrument for interactive, on-the-fly machine learning. In *Proceedings of NIME*, 2009.
- [5] J. Foote. Automatic audio segmentation using a measure of audio novelty. In *Proceedings of ICME2000*, vol. 1, pages 452–455, 2000.
- [6] J. Foote and M. L. Cooper. Media segmentation using self-similarity decomposition. In *Proc. SPIE, Storage and Retrieval for Media Databases*, vol. 5021, pages 167–175., 2003.
- [7] N. Gillian, B. Knapp, and S. O’modhrain. Recognition of multivariate temporal musical gestures using n-dimensional dynamic time warping. In *Proceedings of NIME*, 2011.
- [8] N. E. Gillian. *Gesture recognition for musician computer interaction*. PhD thesis, Queen’s University Belfast, 2011.
- [9] R. H. Jack, A. Mehrabi, T. Stockman, and A. McPherson. Action-sound latency and the perceived quality of digital musical instruments: Comparing professional percussionists and amateur musicians. *Music Perception*, 36(1):109–128, 2018.
- [10] A. McPherson, R. Jack, and G. Moro. Action-sound latency: Are our tools fast enough? In *Proceedings of NIME*, 2016.
- [11] J. I. Mendoza. Segmentation boundaries in accelerometer data of arm motion induced by music: Online computation and perceptual assessment. *Human Technology*, 18(3):250–266, 2022.
- [12] J. I. Mendoza, A. Danso, G. Luck, T. Rantalainen, L. Palmberg, and S. Chastin. Musification of accelerometry data towards raising awareness of physical activity. In *Proceedings of SoniHED*, 2022.
- [13] J. I. Mendoza and M. R. Thompson. Modelling perceived segmentation of bodily gestures induced by music. In *Proceedings of ESCOM*, pages 128–133, 2017.
- [14] D. J. Merrill and J. A. Paradiso. Personalization, expressivity, and learnability of an implicit mapping strategy for physical interfaces. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 2152–2161, 2005.
- [15] F. R. Moore. The dysfunctions of midi. *Computer music journal*, 12(1):19–28, 1988.
- [16] D. Murad, F. Ye, M. Barone, and Y. Wang. Motion initiated music ensemble with sensors for motor rehabilitation. In *2017 international conference on orange technologies*, pages 87–90. IEEE, 2017.
- [17] J. Rodrigues, P. Probst, and H. Gamboa. Tssummarize: A visual strategy to summarize biosignals. In *International conference on Bio Signals, Images, and Instrumentation*, pages 1–6, 2021.
- [18] G. Schätti. Real-time audio feature analysis for decklight3, 2007.
- [19] E. Staudt, Pascal; Sarigöl, M. Lussana, M. Rizzonelli, and J. Hyun Kim. Automatic classification of interactive gestures for inter-body proximity sonification. In *Proceedings of SoniHED*, 2022.
- [20] K. Tahiroğlu, N. N. Correia, and M. Espada. Pesi extended system: In space, on body, with 3 musicians. In *Proceedings of NIME*, 2013.
- [21] D. Tardieu, R. Chessini, J. Dubois, S. Dupont, S. Hidot, B. Mazzarino, A. Moinet, X. Siebert, G. Varni, and A. Visentin. Video navigation tool: Application to browsing a database of dancers’ performances. In *5th International Summer Workshop on Multimodal Interfaces*, pages 35 – 40, 2009.
- [22] F. Visi. *Methods and Technologies for the Analysis and Interactive Use of Body Movements in Instrumental Music Performance*. PhD thesis, Plymouth University, 2017.
- [23] D. Wessel and M. Wright. Problems and prospects for intimate musical control of computers. *Computer music journal*, 26(3):11–22, 2002.

APPENDIX

Software and documentation: https://gitlab.jyu.fi/juigmend/temporal_segmentation_gestural_control



VII

MUSIFICATION OF ACCELEROMETRY DATA TOWARDS RAISING AWARENESS OF PHYSICAL ACTIVITY

by

Juan Ignacio Mendoza, Andrew Danso, Geoff Luck,
Timo Rantalainen, Lotta Palmberg, & Sebastien Chastin, 2022

In Proceedings of the Conference on
Sonification of Health and Environmental Data

<https://doi.org/10.5281/zenodo.7243875>

Musification of Accelerometry Data Towards Raising Awareness of Physical Activity

Juan Ignacio Mendoza

University of Jyväskylä,
Department of Music, Art, and Culture Studies
juigmend@student.jyu.fi

Geoff Luck

University of Jyväskylä,
Department of Music, Art, and Culture Studies,
Centre of Excellence in Music, Mind, Body and Brain
geoff.luck@jyu.fi

Lotta Palmberg

University of Jyväskylä,
Faculty of Sport and Health Sciences,
Gerontology Research Center
lotta.m.palmberg@jyu.fi

Andrew Danso

University of Jyväskylä,
Department of Music, Art, and Culture Studies,
Centre of Excellence in Music, Mind, Body and Brain
andrew.a.dansoadu@jyu.fi

Timo Rantalainen

University of Jyväskylä,
Faculty of Sport and Health Sciences,
Gerontology Research Center
timo.rantalainen@jyu.fi

Sebastien Chastin

Glasgow Caledonian University,
School of Health and Life Sciences;
Ghent University,
Department of Movement and Sports Sciences
sebastien.chastin@gcu.ac.uk

ABSTRACT

Previous research has shown that the temporal dynamics of human activity recorded by accelerometers share a similar structure with music. This opens the possibility to use musical sonification of accelerometry data to raise awareness of daily physical activity. In this study a method was developed for quantifying the daily structure of human activity using multigranular temporal segmentation, and applying it to produce musical sonifications. Two accelerometry recordings of physical activity were selected from a dataset, such that one shows more physical activity than the other. These data were segmented in different time-scales so that segmentation boundaries at a given time-scale have a corresponding boundary at a finer time-scale, occurring at the same point in time. This produced a hierarchical structure of daily events embedded in larger events, which is akin to musical structure. The segmented daily data of each subject was mapped to musical sounds, resulting in two short musical pieces. A survey measured the extent to which people would identify the piece corresponding to the most active subject, resulting in a majority of correct answers. We propose that this method has potential to be a valuable and innovative technique for behavioural change towards reducing sedentary behaviour and increasing physical activity.

Copyright: © 2022 Juan Ignacio Mendoza et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 Unported License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

1. INTRODUCTION

Miniature sensors, wearable devices and mobile technologies can track daily activity of people, both in extent (i.e., amount of movement) and type (e.g., walking, sitting). This capability has been utilised as a behavioural change technique [1] in interventions to promote a healthier lifestyle, increase physical activity and reduce sedentary behaviour¹ [3,4]. These technologies may be effective aids in interventions to increase physical activity and reduce sedentary behaviour [5], but only in the short-term. Long-term adherence is still a major challenge [6–10]. Recent reviews suggest that more engaging methods are needed to effectively produce a change in behaviour [11].

Sonification is a potential strategy to increase long term engagement and adherence, especially since it has been shown that the temporal dynamics of human motion and activity share similarity with that of music [12, 13]. Several studies have explored the use of real-time sonification of movement to aid sports performance and rehabilitation [14]. For example, Ley-Flores et al. [15] found that sonification of exercise with metaphorical sounds affect body perception, causing people to feel strong and thus increase their amount of physical activity. Other studies investigated presenting activity patterns as musical sound to raise awareness about behaviour. For example, Krasnoskulov [16] developed a system in which data measured by an accelerometer and optical heart-rate sensor were mapped to musical parameters such as pitch, timbre, tempo, space and loudness. This form of musical sonification is rather direct and may not result in a clear representation of events. Consequently, some

¹ A short article by O’Keeffe, Scheid and West [2] explains the differences and similarities between physical activity and sedentary behaviour.

studies have considered segmentation of data, so that the resulting sonification is structured in blocks that preserve the temporal relations of events. Last and Usyskin [17] developed a sonification paradigm that segments data into a user-defined number of segments, which was successful to convey the desired information. Vickers and Höldrich [18] progressed this to produce segments using zero-crossing of a one-dimensional data-stream. Then the segments were mapped to sound. These studies show that sonification and musical sonification are feasible ways to convey activity data. Temporal segmentation may be a relevant part of the process, as it allows for mappings between data and sound that have a clear correspondence. However, the temporal segmentation methods used by the mentioned studies have important limitations, as they are based on threshold, zero-crossings or clustering. These methods require careful calibration of input parameters and do not generalise well when patterns in data are multidimensional.

The present study has focused on the development of a system to produce musical sonification (also referred to as *musification*) of daily activity data recorded by wearable devices. The method employs a novel approach to multi-granular temporal segmentation, that results in a clear correspondence between daily events and sound. Additionally, the system does not require the final user to do any fine-tuning of segmentation parameters. We propose this system as an aid in behavioural change, by raising awareness of people’s own daily physical activity in an engaging way.

2. METHODS

2.1 Accelerometry Data

We used two multiple-day recordings of accelerometry from 75-year-old adults. These were chosen from the AGNES database [19, 20] so that one corresponds to a low-activity sedentary subject while the other corresponds to a high-activity non-sedentary subject. The data was obtained by two tri-axial accelerometers, one chest-worn and the other thigh-worn. These data were pre-processed to obtain features for successive non-overlapping epochs of 5 seconds. One feature is the Mean Absolute Deviation (MAD) of the square norm [21], from the thigh-worn accelerometer (Fig. 2a). The other features are the activities identified from the orientation of the accelerometers: lying, sitting, upright posture and walking [22] (Fig. 2b).

2.2 Segmentation

The segmentation procedure is shown in Fig. 1. After MAD is computed and activities are identified, numerosity is reduced by integrating in windows of 120 data points of 5 seconds each (10 minutes) with an overlap of half the length of the window. For MAD the integration is

$$A_i = \log_b \left(1 + \sum_{j=1}^n w_j \right), i = \{1 \dots N\},$$

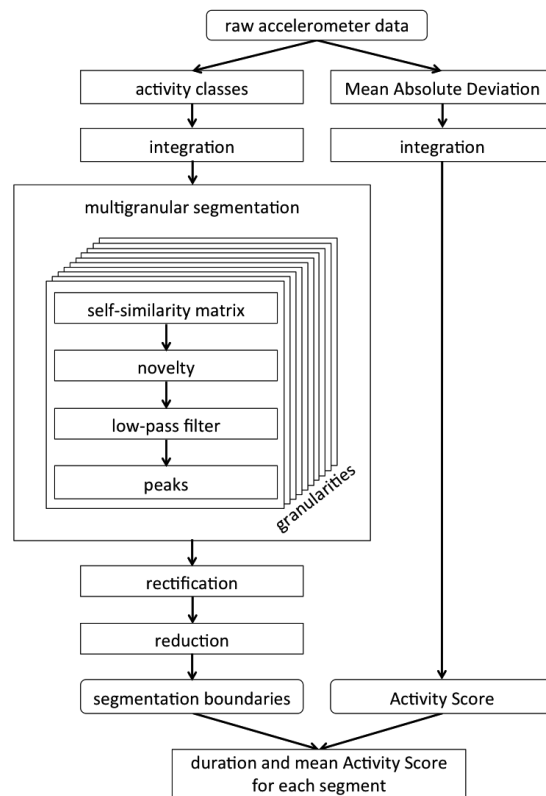


Figure 1. Multigranular segmentation of daily activity.

where vector A of length N is the Activity Score, N is the number of windows, the logarithmic base b is a free parameter to rescale A , and w is one window of length n . The logarithm preserves the data distribution, as the relation between time of inactivity and activity follows a power-law distribution [12]. For the examples reported in this article the logarithm has a base $b = 3$. Each activity is a binary vector, where an activity is represented by a one, otherwise a zero (Fig. 2b). The integration of each activity vector is the sum of the window, with the same length and overlap as for MAD. Additionally, integration acts as a low-pass filter removing unnecessary detail.

The next step is segmentation of the integrated data using the algorithm described by Foote [23]. That algorithm has been used for segmentation of musical audio and video. It can detect boundaries of segments at different *granularities* (i.e., time-scales). It has also been tested for segmentation of accelerometry data of dance [24] and daily activities [25].

The segmentation algorithm first computes a self-similarity matrix of the integrated activities (Fig. 3a). Then, a checkerboard kernel (i.e., a small matrix of four sections where the diagonal is negative units and the anti-diagonal is positive units) tapered by a normal distribution, is correlated along the diagonal of the self-similarity matrix. This was done several times, each with a checkerboard kernel of minimally different size. The size of the kernel corresponds to the granularity of the

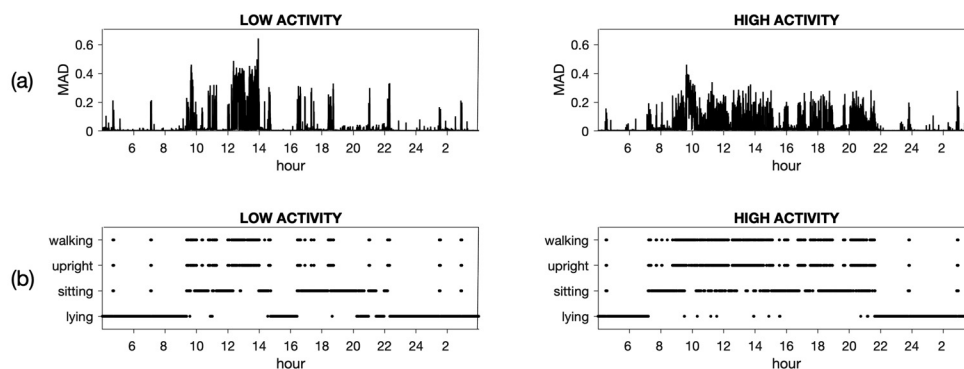


Figure 2. a) Mean Absolute Deviation every 5 seconds of accelerometer data; b) Classification of daily activities.

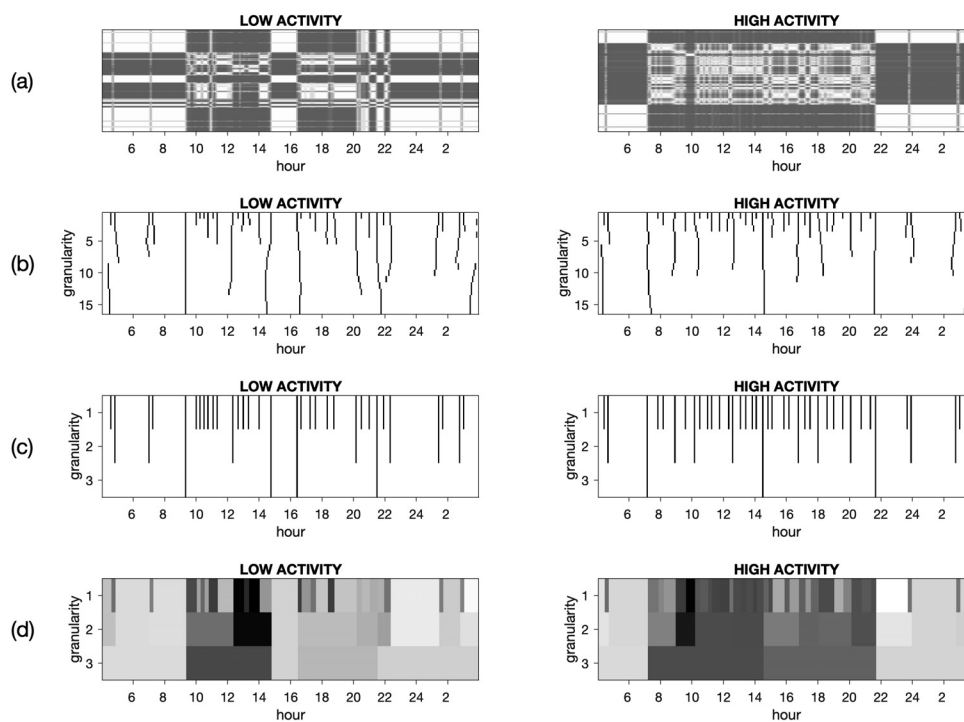


Figure 3. a) Self-similarity of activities; b) multigranular raw segmentation boundaries; c) rectified and reduced multigranular segmentation boundaries; d) segmented Activity Score, darker shades show greater average activity for a segment and vice versa.

segmentation. A smaller kernel detects finer granularity segments and vice versa. The size of the kernels was specified as the standard deviation σ_k of their normal distribution tapering. For the examples shown here, $\sigma_k = \{2, 4, \dots, 32\}$ windows. This resulted in several novelty scores, one for each granularity, each of which was then smoothed with a normal-distribution (i.e., Gaussian) low-pass filter to remove irrelevant peaks. The peaks of

each novelty score represent segmentation boundaries (Fig. 3b). The size of each filtering vector was set to each corresponding value in σ_k , while the standard deviation was set to 0.4 for all of them.

The segmentation boundaries at different granularities are not perfectly aligned in time (Fig. 3b) because, as the checkerboard kernel gets larger, it incorporates more information causing the novelty peak to move slightly in

Table 1. Segmented Activity Score

fine		medium		coarse	
duration	mean Activity	duration	mean Activity	duration	mean Activity
5	0.45	8	0.45	38	0.48
3	1.11	-	-	-	-
30	0.48	30	0.48	-	-
8	1.03	21	0.92	88	1.48

First 4 lines of “high activity” segmented (corresponding to Fig. 3c and d). Duration is windows of integrated data. The headers of this table are not part of the actual list.

either direction. However, because the granularities were set with minimal difference ($\Delta\sigma_k = 2$ windows), it is safe to assume that they correspond to the same segment. Following this logic, every coarser granularity boundary has an origin in a finer granularity boundary, except for those at the borders. The temporal structure is hierarchical, where segments are embedded in larger segments. This reflects the structure of human daily activity. For example, a large portion of the day such as the morning, may contain activities like waking-up and getting ready, breakfast, commuting, and so forth. This hierarchical structure is also analogous to musical structure. For example, a song has sections like introduction, verse and chorus, each of which have sub-sections, such as melodic lines. However, in music the boundaries of each section exactly match in time, unlike the structure resulting from the procedure described above. If that multigranular structure were to be used as musical structure for sonification, it would result in a seemingly unnatural performance. For example, each granularity level may be assigned to a different musical instrument. If so, then instruments would begin and change sections of the song at different times.

Therefore, the segmentation boundaries were aligned to the finest-granularity boundary. Also the boundaries at the borders were removed. This resulted in sequences at different granularities being identical or slightly different. Thus, the finest and coarsest granularity sequences were kept, as well as the sequences that provide greatest variety in number of boundaries. For the examples given here, the reduction resulted in sequences at 3 levels of granularity: fine, medium and coarse (Fig. 3c). Finally, the median Activity Score was computed for each segment at each granularity level (Fig. 3d).

2.3 Musical Sonification

The result of the segmentation procedure is a list of paired columns, where the paired values are segment duration, in windows, and the mean Activity Score for the segment. If a segment’s boundary doesn’t have a corresponding boundary at a coarser granularity level, the values are omitted. The first line assumes a boundary at all levels of granularity. Table 1 shows an example.

The list was formatted as a CSV file and has a header line composed by the number of windows, the sum and grand mean of Activity Scores (from the matrix depicted

by Fig. 3d), and the number of granularities. This file is the input to a separate sonification program consisting mainly of a sequencer and two synthesis modules (Fig. 4.). The sequencer loads the CSV file and immediately reads the header. The user specifies how long the performance will last and the program computes the duration of each window in real time units (e.g., milliseconds), using the first value in the header (number of windows). The second value of the header (sum of Activity Score) is used as the seed for all pseudo-random generators, to obtain a deterministic performance (i.e., the sonification of a CSV file will always be the same). This may help to perceive a strong connection between sonic material and actual daily activity information. The third value in the header (grand mean Activity Score) sets the tempo. The mean between the values of both subjects was mapped to 120 BPM (beats per minute) for crotchet notes (60 BPM for minim notes), as the typical healthy average heartbeat at rest is just over 60 BPM [26] and both preferred musical tempo and average walking steps have a period of about 120 BPM [27]. Hence, the sonification for the high-activity subject will have a slightly higher tempo than the sonification for the low-activity subject. The last element of the header (number of granularities) is used to compute the mean Activity for each combined segment. For example, for the first row in Table 1, all mean Activity values will be added and divided by 3. For the second row, the only value is for the finest granularity and will be divided by 3.

The user inputs a duration in seconds and clicks a button to start the performance. Then, the first line in the CSV file (i.e., the first row of Table 1) will be read and it will wait the duration given by the leftmost value multiplied by the duration of each window, then it will read the next line and so on. When each line is read, the values are sent to the synthesis modules as described below. This process continues until the final line is reached or until the user interrupts it by the click of a button.

Synthesis module 1 is composed by three synthesizers that produce bell-like sounds, whose pitches are pseudo-randomly produced according to a distribution that smoothly transitions from chromatic (i.e., all 12 tones allowed) to a user-selected scale. For this study a pentatonic scale was used. The transition is given by the mean Activity of all segments at the start of a finest-granularity segment. The higher this value is, the closer the distribution will be to the selected scale. For example, when the program begins playing the list in Table 1, it will compute the mean of the “mean Activity” values of the first row, which will determine the distribution of the pseudo-randomly produced notes. Given this distribution, each synthesiser produces a note at the start of each segment and the duration of the note is the duration of the segment. Each synthesiser has been set to play only at a distinct octave, with the synthesiser allocated to the coarsest granularity playing the lowest octave and vice versa.

The resulting sounds are somehow dissonant when activity is low and consonant within the user-defined scale, when the activity is moderately energetic. This defines

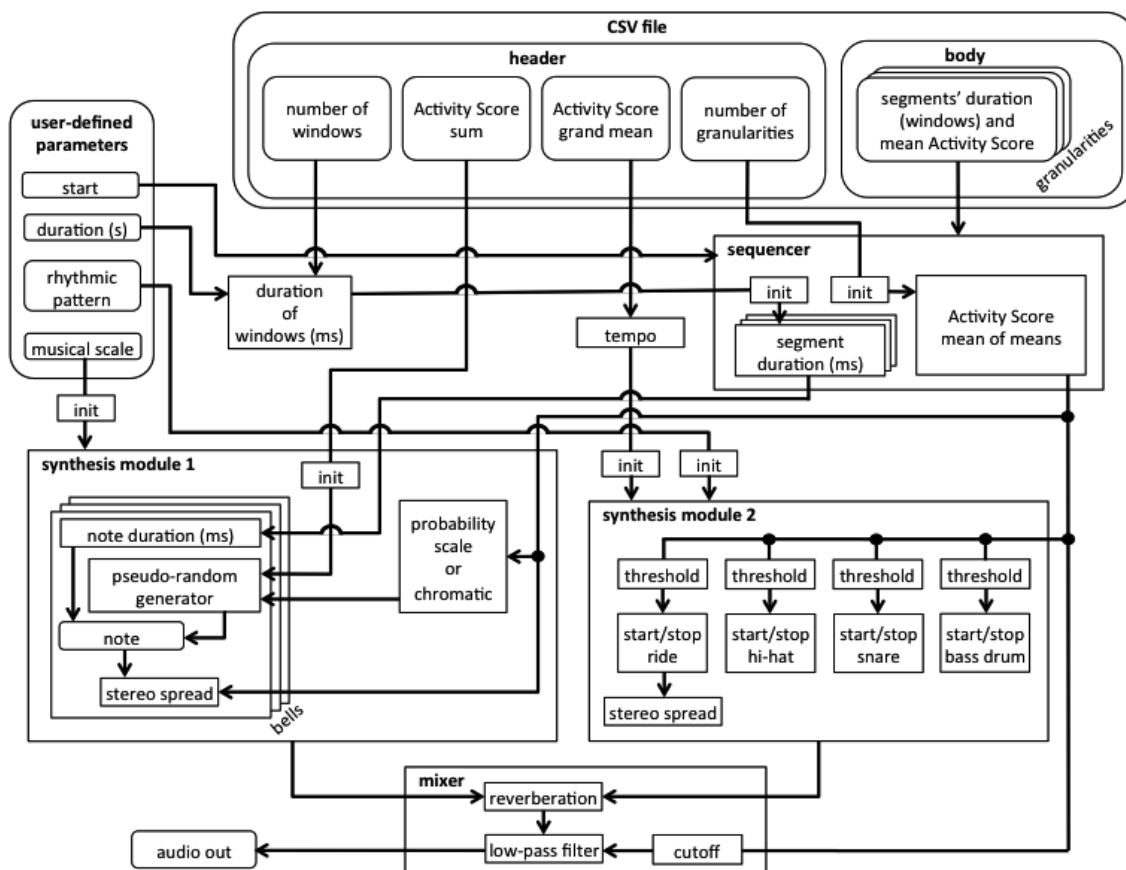


Figure 4. Sonification program.

high amount of activity as consonant and low amount of activity as dissonant. Also each synthesiser has a “stereo spread” capability that has been mapped so that the higher the mean Activity, the wider the stereo allocation of the notes, meaning the pseudo-random balance between their output to the two main output audio channels.

Synthesis module 2 is a drum-machine with 16 steps (quaver notes) and 5 voices produced by frequency modulation: ride cymbal, open hi-hat cymbal, closed hi-hat, snare drum and bass drum. The rhythmic pattern can be programmed by the user. The tempo is given by the grand mean Activity Score as explained previously and each instrument is activated when a level of mean Activity of the current segment exceeds a defined threshold. For this study the bass drum was set to be permanently active, while the ride cymbal was set to fade in when activity changes from very low to moderately low. Also the ride cymbal was set to permanently have a full stereo spread, resulting in a subtle and surrounding rhythmical noise. The open hi-hat was set to a medium threshold and the closed hi-hat was not used for this study. The snare drum was set to a moderately high threshold. The full drum set is active when activity is energetic.

In the examples (see Fig. 2 and Fig. 3), the full drum set and only notes within the defined scale play between

about 12:30 and 15:00 for the low-activity subject and from about 9:00 to 10:00 for the high-activity subject. The output from the bell synthesisers and drum-machine is mixed and subtle reverberation is added to blend the sounds. Finally, a low-pass filter is applied to the final mix and its cut-off frequency is controlled directly by the mean activity of the current segment, so that the resulting sound is slightly brighter as there is more energetic activity and vice versa.

2.4 Perceptual Assessment

Two audio files were produced with the method described above, using excerpts from 6:00 to 23:00 of the data presented in the figures. These audio files were used as stimuli for a perceptual assessment. Data for this assessment were collected during 31 days by means of a short survey using QuestionPro, a service to make and publish questionnaires which can be answered with an internet browser, and Twitter. Participants were recruited via Twitter and Facebook using both free and paid adverts, the latter targeting Finland and major English-speaking countries, and via authors' direct contact within their acquaintance networks. In the survey, participants were asked to listen to each audio file, and indicate which of them represented the more active person. The order of presentation was randomised. The survey included the researchers' con-

tact information, notified participants that no personal information would be collected, and that data collection complied with the General Data Protection Regulation of the European Union. The stimuli can be listened on Twitter: <https://twitter.com/listeningsurvey> and Facebook: <https://www.facebook.com/ListeningsurveyJYU>.

3. RESULTS AND DISCUSSION

The methods described in this report are firstly, a system that processes accelerometry data of daily activity, resulting in multigranular hierarchical segmentation akin to musical structure. The second method is a program devised as a proof of concept, to demonstrate a possible musical sonification of the daily activity data utilising the segmentation obtained. The resulting sonification has, by design, one main property, which is that there is a clear association between sonic events and daily activity. The perceptual assessment of two example sonifications produced with the system described measured the extent to which a person would correctly identify the sonification for high activity data, when presented along the sonification for low activity. A total of 1847 responses were collected by a survey on the internet, of which 1225 (66.3%) correctly identified the sonification corresponding to high activity. A one-proportion z-test was performed to evaluate the statistical significance of the results, yielding $z = 14.03$, with a p -value $< 1 \times 10^{-5}$. This may be sufficient to reject the null hypothesis, suggesting that the proportion of correct responses is significant.

The described musical sonification system may be useful in public health interventions towards increasing healthy physical activity or reducing sedentary behaviour, by making a person aware of their intraday activity in an engaging manner. In practice, the musical sonification system would be part of a portable system comprising hardware and software. Such a system would record daily activity, produce the musical sonification and possibly recommend actions to the user. The hardware may be composed of already existing technologies such as miniature accelerometers and mobile computing devices like a smartphone or smartwatch. Future research shall be carried out to implement the system and test it in ecologically valid conditions. Preliminary testing shall be carried out in order to explore the extent to which the musical sonification may work as an engagement strategy, and to identify the conditions in which it may be effective. These conditions may include personal characteristics of target users such as age, personality or income, as well as environmental conditions. Also it would be useful to compare the multigranular segmentation daily profiles of users with self-reports on their activities, to assess the extent of their correspondence.

While this report describes a method for multigranular segmentation and musical sonification of intraday activity of one subject, it is trivial to expand the method to work with different data. First, instead of using classified data for the segmentation, the Activity Score may be used alone. Also instead of using a single time period, like a day, an average of several days may be used, resulting in a rep-

resentation of a typical day. Furthermore, instead of using data for a single subject, a group of subjects may be used. A population may be pre-clustered in groups with homogeneous characteristics, such as age, gender, and so on. The resulting multigranular temporal segmentation may be useful to examine the typical intraday behaviour of the group. Its musical sonification will represent the group and this may open new and interesting doors for community music making. For example, daily data of a person may be uploaded to a server, where it would be combined with data of other people in their social circle. This would enable them to produce music as a group, instead of individually. This way of collaborative music-making may be a relevant avenue for exploration in further research, as it has been observed that social support through collaboration was the primary motivator for adults to maintain activity tracker use [28].

4. CONCLUSION

This study has developed a system to produce musical sonification of daily activity data recorded by wearable devices. The sonification may be used as a tool for raising awareness and behaviour change by conveying daily activity information to users in a clear and engaging way. This capability may be used in interventions to increase physical activity (i.e., total amount of bodily motion) and reduce sedentary behaviour (i.e., proportion of time sitting or lying down) in hard to reach populations such as older adults, teenagers or people with visual or learning difficulties. A key property of the musical sonification is that it shows clearly not only the overall physical activity over a period of time, but of the temporal structure within, such as commuting to work, or taking a lunch break. This property would allow someone to identify, by listening to the sonification, the times of the day they were more or less active and spent more or less time sitting. That was achieved by devising a novel multigranular temporal segmentation procedure that preserves the time relations between events.

Acknowledgments

Petri Toiviainen suggested using probability to generate musical notes.

This work was supported in part by the Finnish Cultural Foundation (Suomen Kulttuurirahasto).

The AGNES study was financially supported by the Advanced Grant from the European Research Council (grant 310526) and the Academy of Finland (grant 693045), both to Taina Rantanen. The funders had no role in the design of the study and data collection, analysis, and interpretation of data, and in writing the manuscript. The content of this article does not reflect the official opinion of the European Union. Responsibility for the information and views expressed in the article lies entirely with the authors.

5. REFERENCES

- [1] S. Michie, M. Richardson, M. Johnston, C. Abraham, J. Francis, W. Hardeman, M. P. Eccles, J. Cane, and C. E. Wood, "The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions," *Annals of behavioral medicine*, vol. 46, no. 1, pp. 81–95, 2013.
- [2] N. O’Keeffe, J. L. Scheid, and S. L. West, "Sedentary behavior and the use of wearable technology: An editorial," *International Journal of Environmental Research and Public Health* 17(12), p. 4181, 2020.
- [3] R. Daryabeygi-Khotbehsara, S. M. Shariful Islam, D. Dunstan, J. McVicar, M. Abdelrazek, and R. Maddison, "Smartphone-based interventions to reduce sedentary behavior and promote physical activity using integrated dynamic models: Systematic review," *J Med Internet Res*, vol. 23, no. 9, p. e26315, Sep 2021. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/34515637>
- [4] F. Monteiro-Guerra, O. Rivera-Romero, L. Fernandez-Luque, and B. Caulfield, "Personalization in real-time physical activity coaching using mobile applications: A scoping review," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 6, pp. 1738–1751, 2020.
- [5] R. T. Larsen, V. Wagner, C. B. Korffitsen, C. Keller, C. B. Juhl, H. Langberg, and J. Christensen, "Effectiveness of physical activity monitors in adults: systematic review and meta-analysis," *BMJ*, vol. 376, 2022. [Online]. Available: <https://www.bmj.com/content/376/bmj-2021-068047>
- [6] K.-J. Brickwood, G. Watson, J. O’Brien, and A. D. Williams, "Consumer-based wearable activity trackers increase physical activity participation: Systematic review and meta-analysis," *JMIR Mhealth Uhealth*, vol. 7, no. 4, p. e11819, Apr 2019. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/30977740>
- [7] S. A. Buckingham, A. J. Williams, K. Morrissey, L. Price, and J. Harrison, "Mobile health interventions to promote physical activity and reduce sedentary behaviour in the workplace: A systematic review," *Digital health*, vol. 5, 2019. [Online]. Available: <https://doi.org/10.1177/2055207619839883>
- [8] M. I. Cajita, C. E. Kline, L. E. Burke, E. G. Bigini, and C. C. Imes, "Feasible but not yet efficacious: a scoping review of wearable activity monitors in interventions targeting physical activity, sedentary behavior, and sleep," *Current Epidemiology Reports*, vol. 7, no. 1, pp. 25–38, 2020. [Online]. Available: <https://doi.org/10.1007/s40471-020-00229-2>
- [9] J. Y.-W. Liu, P. P.-K. Kor, C. P.-Y. Chan, R. Y.-C. Kwan, and D. S.-K. Cheung, "The effectiveness of a wearable activity tracker (wat)-based intervention to improve physical activity levels in sedentary older adults: A systematic review and meta-analysis," *Archives of Gerontology and Geriatrics*, vol. 91, p. 104211, 2020. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0167494320302053>
- [10] A. V. Creaser, S. A. Clemes, S. Costa, J. Hall, N. D. Ridgers, S. E. Barber, and D. D. Bingham, "The acceptability, feasibility, and effectiveness of wearable activity trackers for increasing physical activity in children and adolescents: A systematic review," *International Journal of Environmental Research and Public Health*, vol. 18, no. 12, 2021. [Online]. Available: <https://www.mdpi.com/1660-4601/18/12/6211>
- [11] W. Wang, J. Cheng, W. Song, and Y. Shen, "The effectiveness of wearable devices as physical activity interventions for preventing and treating obesity in children and adolescents: Systematic review and meta-analysis," *JMIR Mhealth Uhealth*, vol. 10, no. 4, p. e32435, Apr 2022. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/35394447>
- [12] S. Chastin and M. Granat, "Methods for objective measure, quantification and analysis of sedentary behaviour and inactivity," *Gait & Posture*, vol. 31, no. 1, pp. 82–86, 2010. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S096663620900602X>
- [13] D. J. Levitin, P. Chordia, and V. Menon, "Musical rhythm spectra from bach to joplin obey a 1/f power law," *Proceedings of the National Academy of Sciences*, vol. 109, no. 10, pp. 3716–3720, 2012. [Online]. Available: <https://www.pnas.org/doi/abs/10.1073/pnas.1113828109>
- [14] N. Schaffert, T. B. Janzen, K. Mattes, and M. H. Thaut, "A review on the relationship between sound and movement in sports and rehabilitation," *Frontiers in psychology*, vol. 10, p. 244, 2019.
- [15] J. Ley-Flores, L. T. Vidal, N. Berthouze, A. Singh, F. Bevilacqua, and A. Tajadura-Jiménez, "Soniband: Understanding the effects of metaphorical movement sonifications on body perception and physical activity," in *CHI '21 Conference, Yokohama, Japan, May 8-13, 2021*, Y. Kitamura, A. Quigley, K. Isbister, T. Igarashi, P. Bjørn, and S. M. Drucker, Eds. ACM, 2021, pp. 521:1–521:16. [Online]. Available: <https://doi.org/10.1145/3411764.3445558>

- [16] A. Krasnoskulov, "Family album: How does your daily activity sound?" in *ICAD Conference, Northumbria University*, April 2019. [Online]. Available: https://www.researchgate.net/profile/Alex-Krasnoskulov/publication/341071964_FAMILY_ALBUM_HOW_DOES_YOUR_DAILY_ACTIVITY_SOUND
- [17] M. Last and A. Usyskin (Gorelik), *Listen to the Sound of Data*. Cham: Springer International Publishing, 2015, pp. 419–446. [Online]. Available: https://doi.org/10.1007/978-3-319-14998-1_19
- [18] P. Vickers and R. Höldrich, "Direct segmented sonification of characteristic features of the data domain," *arXiv*. [Online]. Available: <https://arxiv.org/abs/1711.11368> 2017.
- [19] T. Rantanen, M. Saajanaho, L. Karavirta, S. Silta-nen, M. Rantakokko, A. Viljanen, T. Rantalainen, K. Pynnönen, A. Karvonen, I. Lisko, L. Palmberg, J. Eronen, E.-M. Palonen, T. Hinrichs, M. Kauppinen, K. Kokko, and E. Portegijs, "Active aging –resilience and external support as modifiers of the disablement outcome: Agnes cohort study protocol," *BMC Public Health*, vol. 18, no. 1, p. 565, 2018. [Online]. Available: <https://doi.org/10.1186/s12889-018-5487-5>
- [20] E. Portegijs, L. Karavirta, M. Saajanaho, T. Rantalainen, and T. Rantanen, "Assessing physical performance and physical activity in large population-based aging studies: home-based assessments or visits to the research center?" *BMC Public Health*, vol. 19, no. 1, p. 1570, 2019. [Online]. Available: <https://doi.org/10.1186/s12889-019-7869-8>
- [21] H. Vähä-Ypyä, T. Vasankari, P. Husu, A. Mänttari, T. Vuorimaa, J. Suni, and H. Sievänen, "Validation of cut-points for evaluating the intensity of physical activity with accelerometry-based mean amplitude deviation (mad)," *PloS one*, vol. 10, no. 8, p. e0134813, 2015.
- [22] T. Rantalainen, K. Koivunen, E. Portegijs, T. Rantanen, L. Palmberg, L. Karavirta, and S. Chastin, "Is complexity of daily activity associated with physical function and life space mobility among older adults?" *Medicine and science in sports and exercise*, February 2022. [Online]. Available: <https://doi.org/10.1249/MSS.0000000000002883>
- [23] J. Foote, "Automatic audio segmentation using a measure of audio novelty," in *2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No.00TH8532)*, vol. 1, 2000, pp. 452–455 vol.1.
- [24] J. I. Mendoza and M. R. Thompson, "Modelling perceived segmentation of bodily gestures induced by music," in *ESCOM 2017 Conference*, E. Van Dyck, Ed. Ghent University, 2017, pp. 128–133.
- [25] J. Rodrigues, P. Probst, and H. Gamboa, "Tssummarize: A visual strategy to summarize biosignals," in *2021 Seventh International conference on Bio Signals, Images, and Instrumentation (ICBSII)*, 2021, pp. 1–6.
- [26] D. Nanchen, M. J. Leening, I. Locatelli, J. Cornuz, J. A. Kors, J. Heeringa, J. W. Deckers, A. Hofman, O. H. Franco, B. H. C. Stricker *et al.*, "Resting heart rate and the risk of heart failure in healthy adults: the rotterdam study," *Circulation: Heart Failure*, vol. 6, no. 3, pp. 403–410, 2013.
- [27] B. Burger, M. R. Thompson, G. Luck, S. H. Saarikallio, and P. Toiviainen, "Hunting for the beat in the body: on period and phase locking in music-induced movement," *Frontiers in Human Neuroscience*, vol. 8, no. 903, 2014. [Online]. Available: <https://doi.org/10.3389/fnhum.2014.00903>
- [28] A. Kononova, L. Li, K. Kamp, M. Bowen, R. Rikard, S. Cotten, and W. Peng, "The use of wearable activity trackers among older adults: Focus group study of tracker perceptions, motivators, and barriers in the maintenance stage of behavior change," *JMIR Mhealth Uhealth*, vol. 7, no. 4, p. e9832, Apr 2019. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/30950807>