Ghent University
Faculty of Engineering and Architecture
Department of Electronics and Information Systems
Faculty of Arts and Philosophy
Department of Art, Music and Theater Studies

The analysis of bodily gestures in response to music.

Methods for embodied music cognition based on machine learning.

Denis Amelynck

Promotor: Prof. Dr. Marc Leman
Co-Promotor: Prof. Dr. Ir. Jean-Pierre Martens

Thesis submitted to fulfill the requirements for the degree of
Doctor of Engineering
Academic year 2013-2014

Ghent University
Faculty of Engineering and Architecture
Department of Electronics and Information Systems
Faculty of Arts and Philosophy
Department of Art, Music and Theater Studies

Promotor:	Prof. Dr. Marc Leman
	Prof. Dr. Ir. Jean-Pierre Martens
Reading committee:	Prof. Hans De Meyer
	Prof. Dr. Ir. Leon van Noorden
	Prof. Thierry Dutoit
	Prof. Giovanni De Poli
Examination committee:	Prof. Patrick De Baets
	Prof. Kris Demuynck

Ghent University
Faculty of Engineering and Architecture
Department of Electronics and Information Systems

Faculty of Arts and Philosophy
Department of Art, Music and Theater Studies

IPEM - Institute for Psychoacoustics and Electronic Music
Blandijnberg 2, B-9000 Ghent, Belgium

Tel.: +32-9-264.41.26
Fax.: +32-9-264.41.43

Thesis submitted to fulfill the requirements for the degree of
Doctor of Engineering
Academic year 2013-2014

This thesis was typeset in LaTeX, using a modified version of the Ghent University's template available on-line (http://latex.ugent.be/). The bibliography was managed using BibDesk and BibTeX with the apacite package.

I certify that this thesis, and the research to which it refers, are the product of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline. I acknowledge the helpful guidance and support of my supervisors, Prof. Dr. Marc Leman and Prof. Dr. Ir. Jean Pierre Martens.

This work is carried out in context of the EmcoMetecca Methusalem project at Ghent University, and funded by the Flemish Government.

ISBN 978-90-8578-700-6

Acknowledgments

I would like to express my deep gratitude to Professor Marc Leman and Professor Jean Pierre Martens, my research supervisors, for their patient guidance, enthusiastic encouragement and useful critiques of this research work.

I would also like to thank Professor Leon van Noorden, for his contribution in the form of novel ideas and solutions.

My grateful thanks are also extended to Dr. Pieter-Jan Maes designer and implementor of the experiments, to Mr. Ivan Schepers, technical coach and handyman and to Ms. Katrien De Bouck for freeing me from all administrative hassle.

Submitting a manuscript to an international peer-reviewed journal requires thoroughly checking its English language for spelling, vocabulary and grammar. I would like to thank Jim Wickenhiser and his wife Jenny for their assistance in checking my manuscripts.

Finally, I wish to thank my family and especially my wife for their support and encouragement throughout my study.

Waasmunster, februari 2014
Denis Amelynck

Table of Contents

Outlook - A glimpse of the future **6-1**

Appendix - Technical Background **A-1**

List of Acronyms

A

ACE	Adaptive Communication Environment With an explanation
ASAP	As Soon As Possible

B

BP	Beat Point
BPM	Beats Per Minute

C

CCA	Canonical Correlation Analysis
CRF	Conditional Random Field

D

DAG	Directed Acyclic Graph
DGM	Directed Graphical Model
DPM	Dirichlet Process Mixture
DPGMM	Dirichlet Process Gaussian Mixture Model
DPMMM	Dirichlet Process Multinomial Mixture Model
DSP	Digital Signal Processing
DTW	Dynamic Time Warping

E

EMC	Embodied Music Cognition

F

FFT	Fast Fourier Transform
FIR	Finite Impulse Response
FDA	Functional Data Analysis
fMRI	Functional Magnetic Resonance Imaging
FPCA	Functional Principal Component Analysis

G

GMM	Gaussian Mixture Model
GMR	Gaussian Mixture Regression
GUI	Graphical User Interface

H

HCI	Human Computer Interface
HMM	Hidden Markov Model
HSMM	Hidden semi-Markov Model

I

IPEM	Institute for Psychoacoustics and Electronic Music - Ghent University

L

LL	Log Likelihood
LMA	Laban Movement Analysis
LOOCV	Leave-One-Out-Cross Validation
LSE	Least Square Error

M

MDS	Multi Dimensional Scaling
MIDI	Musical Instrument Digital Interface
MCMC	Markov Chain Monte Carlo

N

NRMSE	Normalized Root Mean Square Error

P

PbD	Programming by Demonstration

R

RMSE	Root Mean Square Error

S

STFT	Short-Time Fourier Transform

U

UGM	Undirected Graphical Model

Nederlandse samenvatting –Summary in Dutch–

Mooi aan muziek is haar universeel karakter maar toch blijft het moeilijk om muzikale belevenissen op een objective manier te beschrijven. Meestal eindigt men ergens met subjectieve beschrijvingen zoals blijkt uit het veelvuldig gebruik van metaforen om muzikale ervaringen te beschrijven (vb. muziek is licht, muziek is zonnig, . . .).

De theorie van muzikale lijfelijkheid van Prof. Marc Leman [1] biedt hiervoor echter een praktische oplossing. Het begint met de vaststelling dat mensen wanneer ze muziek horen ook de neiging hebben om te bewegen met deze muziek. Het idee is nu om deze bewegingen te bestuderen en aan de hand van deze bevindingen een indirect maar objectief oordeel te kunnen vellen over hoe mensen muziek ervaren.

Hoewel deze benadering objectief is, is ze daarom niet altijd eenvoudig. Een groot probleem bij het bestuderen van deze muzikale bewegingen is namelijk de hoge dimensionaliteit. Die hoge dimensionaliteit ontstaat uit het opmeten van lichaamsdelen (zoals handen, benen, hoofd, romp, ...) in drie dimensies voor verschillende subjecten maar ook in het meten van gegevens die indirect met beweging te maken hebben en dan denken we in eerste instantie aan biometrische gegevens (hartslag, ademvolume, tot zelfs het meten van bloedstromen in de hersenen met behulp van fMRI). Al deze metingen leveren een gigantische hoeveelheid data op en we kunnen hier terecht verwijzen naar de tegenwoordig zeer populaire term *big data*. De uitdagingen van big data liggen niet alleen bij het verzamelen van gegevens of de opslag van gegevens maar vooral ook bij de analyses en methodes van visualizatie.

Voor methodes van analyse en visualizatie doen we hierbij vooral beroep op ideeën uit de wereld van het *Machinaal Leren (Machine Learning)*. Daar wordt er onderscheid gemaakt tussen drie types van methodes:

1. Gesuperviseerd Leren (Supervised Learning) : Hierbij wensen we scores toe te kennen aan een object en dit gebaseerd op een aantal meetwaarden. Deze scores kunnen nominale waarden zijn (zoals bijvoorbeeld ”type1”, ”type2”, ...) en dan spreken we over classificatie. Deze scores kunnen echter ook continue grootheden zijn (zoals bijvoorbeeld lengte) en dan spreken we over regressie. Fundamenteel is dat we een

model bouwen dat deze scores kan voorspellen op basis van metingen. Dit model wordt gebouwd aan de hand van trainingsvoorbeelden waar we een set hebben van metingen en een set van corresponderende scores. Methodes van Gesuperviseerd Leren worden voornamelijk gebruikt in hoofdstuk 2 (Towards E-Motion Based Music Retrieval) en hoofdstuk 4 (Beating-Time Gestures: Imitation Learning for Humanoid Robots).

2. Ongesuperviseerd Leren (Unsupervised Learning) : Meestal zijn er geen kant-en klare trainingsvoorbeelden met metingen én scores ter beschikking of is het opstellen daarvan een titanenwerk. In deze omstandigheden biedt Unsupervised Learning een alternatief. Unsupervised Learning is een set van technieken met als doel het brengen van structuur in data. Dit kan zijn door de data eenvoudiger voor te stellen, door bijvoorbeeld naar een lagere dimensie over te schakelen. Denk bijvoorbeeld aan twee dimensionale wegenkaarten die een reductie zijn van wat er in dree dimensies gebeurt op de wereldbol. Een andere manier van structuur aanbrengen, is clustering, het groeperen van data in groepen van gelijken. Deze methodes vormen de basis voor de analyses in hoofdstuk 3 (Expressive Body Movement Responses to Music are Coherent, Consistent, and Low Dimensional) en in hoofdstuk 5 (The Surprising Character of Music. A search for sparsity in music evoked body movements).

3. Ondersteund Leren (Reinforcement Learning) : Hierbij gaat een *robot* leren hoe hij, vertrekkende vanuit een aantal waarnemingen, te werk moet gaan om een doel (maximale winst) te bereiken door het uitvoeren van een aantal acties in een bepaalde volgorde. Deze tak van Machinaal Leren is vooral belangrijk voor de robotica wat niet direct tot het onderzoeksgebied van deze thesis behoort en daarom wordt er in deze thesis van ondersteund leren geen gebruik gemaakt. De ideeën zouden echter wel nuttig kunnen zijn voor muzikale opvoedingsspellen [2].

Om te kunnen werken met Machinaal Leren hebben we gegevens (data) nodig. Helaas zijn de gegevens niet altijd direct beschikbaar in de vorm die we wensen. Voor muziek is bijvoorbeeld periodiciteit heel belangrijk, maar periodiciteit is niet iets dat we zomaar kunnen meten. Het wordt meestal afgeleid uit andere (ruwe) gegevens, bijvoorbeeld met behulp van een Fourier analyse. Het proces van omzetten van ruwe gegevens naar grootheden die rechtstreeks bruikbaar zijn voor de gewenste analyse noemen we voorverwerking. Voorverwerking (preprocessing) van de gegevens speelt een belangrijke rol in deze thesis.

Met voorverwerking bedoelen we dus meer dan enkel het filteren van ruis uit de signalen, of het wegwerken van uitschieters, het betekent ook de gegevens omzetten in andere gegevens die dan weer nuttig zijn voor verdere analyses. Zo waren voor ons werk onder andere de grootheden volume, dimensionaliteit, nabijheid, en richting van beweging heel belangrijk, ook al omdat ze verband houden met de emotionele intentie van muziek. Deze grootheden werden afgeleid uit de plaatscoordinaten.

De vorige paragrafen benadrukten vooral de individuele belevenis van muziek, maar muziek is ook een sociaal gebeuren. Als we enkel naar de bewegende massa's kijken op popconcerten begrijpen we dat ook dit aspect zeer belangrijk is. Ook hier levert ons onderzoek een bijdrage. We hebben namelijk de begrippen coherentie en consistentie geïntroduceerd om deze fenomenen te beschrijven. Met coherentie bedoelen we dat een groep mensen gelijkmatig beweegt in één aaneensluitend tijdsinterval. Consistentie definiëren we als gelijkmatigheid over tijdsintervallen die uit elkaar liggen. Deze begrippen zijn dan weer sterk verankerd met expressiviteit. Expressiviteit is eigenlijk een verhaal van *min of meer* en kunnen we daardoor enkel meten door te vergelijken met een referentie, bijvoorbeeld het gemiddelde van een groep. Het onderzoek naar deze begrippen staat in hoofdstuk 3 (Expressive Body Movement Responses to Music are Coherent, Consistent, and Low Dimensional).

Alle methodes en technieken hebben één ding gemeen en dat is dat het gaat over het nemen van *beslissingen* gebaseerd op het maken van *vergelijkingen*. Beslissingen: wat is het beste model, de beste voorspelling, ...? en vergelijkingen: Hoe definiëren we *best*? Belangrijk hierbij, zijn definities voor gelijkheid en afstanden. Er bestaat een grote hoeveelheid aan definities en in onze thesis gebruiken we slechts een beperkte set van probabilistische en niet-probabilistische definities. Hier is een overzicht :

- niet-probabilistisch:
  - tussen twee verhoudingen (mixtures) : cosinus afstand
  - tussen twee data punten : euclidische afstand
- probabilistisch:
  - tussen twee data samples : verschil in probabiliteit volgens een probabilistisch model
  - tussen twee probabiliteitsdistributies : f-divergenties (waaronder Kullback-Leibler divergentie)

Ten slotte leggen we nog kort uit hoe deze thesis is gestructureerd : Hoofdstukken 2, 3, 4 and 5 zijn uitgebreide versies van artikels die opgestuurd zijn naar internationale collegiaal getoetste tijdschriften. Hoofdstuk 6 is dan weer een beetje apart omdat we daar vooruitkijken naar de toekomst en dit in het licht van de recente ontwikkelingen in de wereld van Machinaal Leren. We bespreken die ontwikkelingen in functie van hun impact naar onderzoek in het muzikale vakgebied.

Hierna volgt nog een bondige samenvatting van de inhoud van hoofdstukken 2, 3, 4 and 5 :

- Hoofdstuk 2 Towards E-Motion Based Music Retrieval.

  Dit hoofdstuk schetst een nieuw mechanisme om luistermuziek te selecteren. Het idee bestaat er uit om met armbewegingen de muziek te kiezen. Dit is nu praktisch haalbaar omdat nieuwe mobiele toestellen (zoals smartphones en MP3 spelers) sensoren bevatten die bewegingen kunnen detecteren. In onze set-up leiden we emotionaliteit af uit de armbewegingen en die emotionaliteit gebruiken we dan om een selectielijst van muziek samen te stellen. Het gebruik van emotionaliteit als tussenliggende stap laat grotere flexibiliteit toe, te vergelijken met het gebruik van een marshalling panel.

- Hoofdstuk 3 Expressive Body Movement Responses to Music are Coherent, Consistent, and Low Dimensional.

  De gevonden resultaten tonen aan dat met een laag-dimensionaal model de expressiviteit van een groep dansers kan beschreven worden. Het model is laag dimensionaal omwille van een grote coherentie en consistentie tussen de dansers. Een directe toepassing van dit model ligt bijvoorbeeld in het selecteren van een groep dansers uit een grote groep dansers met als criterium het hebben van dezelfde expressiviteit.

- Hoofdstuk 4 Beating-Time Gestures: Imitation Learning for Humanoid Robots.

  Dit hoofdstuk beschrijft hoe we uit een aantal dirigeerbewegingen (weliswaar beperkt tot de maat slaan op een metronoom), eerst een veralgemeende beweging kunnen destilleren, die dan verder kan gebruikt worden om de maat aan te geven op willekeurig gekozen muziek.

- Hoofdstuk 5 The Surprising Character of Music. A search for sparsity in music evoked body movements.

  Dit hoofdstuk bevat toepassingen van het type Ongesuperviseerd Leren (Unsupervised learning). Door het gebruik van clusteringstechnieken op zowel positionele data als op richtingsdata, kunnen we de resultaten van een dansuitvoering niet alleen beter beschrijven maar ook beter begrijpen in functie van de muziek. Dit wordt gevisualizeerd in een *directogram* dat kan worden gebruikt als een beschrijving van een muziekstuk. De gebruikte clusteringstechnieken kunnen voorts ook helpen om weinig voorkomende bewegingen te ontdekken, als mogelijke indicatoren voor verrassingselementen in de muziek.

Referenties

[1] M. Leman. *Embodied music cognition and mediation technology.* The MIT Press, 2008.

[2] Y. Björnsson, V. Hafsteinsson, A Jóhannsson, and E. Jónsson. *Efficient use of reinforcement learning in a computer game.* Proceedings of International Journal of Intelligent Games & Simulation, 2008.

English Summary

Across the globe, many people, being composers, performers, dancers or listeners, experience and enjoy music on a daily basis. Still, it is challenging to describe musical experiences in an objective way. Commonly, descriptions make use of metaphors: music is aggressive, music is sunny, . . . but these descriptions are subjective and are no good starting point for a systematic research of music.

The theory of musical embodiment from Leman [1] offers not only a practical but also an objective solution to this problem. A first thing he notices is that when listening to music people tend to move along with the music. His idea is then to study movement in order to understand how people experience music. This leaves the subjective path of appreciation and judgment and enters the world of exact science using measured data. This solution is only possible thanks to recent advances in sensor technology and increasing computing power.

Although the principle is very simple, the realization is not straightforward. One of the challenges of this method is the dimensionality of the data. The high dimensionality of the data finds root in measurements for several subjects, for several body parts (like hands, legs , heads, trunks, . . .) and also in the measurement of additional information indirectly related to movement. We think here for example about biometric data (heart rate, inhalation and expiration volume, and cerebral blood flow (measured by techniques like fMRI). All these measurements result in huge datasets, hence using the hyped term *big data* might be appropriate. The challenges for *big data* lay not only in collection and storage but especially in the analysis and of visualization methods.

For handling analysis and visualization, inspiration can be found in the realm of *Machine Learning*. Machine Learning groups methods in roughly three categories:

1. Supervised Learning: the techniques that resort under supervised learning are mainly Classification and Regression. These are techniques that identify data models starting from a training data base (with labeled data). The main purpose is to predict future values. These techniques are used in chapter 2 (Towards E-Motion Based Music Retrieval) and chapter 4 (Beating-Time Gestures: Imitation Learning for Humanoid Robots).

2. Unsupervised Learning: For many tasks there are no ready-made training data bases available or setting-up such a training database would require a huge effort. In these circumstances unsupervised learning offers an alternative. The main purpose of unsupervised learning is to discover structure in the data. This can be done by dimension reduction and so reducing the complexity of the data set. Think for example about road maps that are a two dimensional reduction of the three dimensional reality of the globe. Another way of adopting structure, is clustering, grouping of data in groups of resemblance. These methods are at the base of chapter 3 (Expressive Body Movement Responses to Music are Coherent, Consistent, and Low Dimensional) and chapter 5 (The Surprising Character of Music. A search for sparsity in music evoked body movements).

3. Reinforcement Learning: this part of Machine Learning was not employed in this thesis. It supports mainly applications in the world of robotics. Nevertheless we would like to mention that ideas from reinforcement learning could be beneficial to musical education games [2].

Machine Learning techniques require data and that is not as straightforward as one might wish. Readily available data (like positional data) is important but it does not stop there. For music, periodicity is very important but periodicity is not something we can just measure. Most of the time it is deducted from other data with the help of a Fourier analysis. This and other forms of preprocessing of the data play an important role in this thesis.

With preprocessing we mean more than just the filtering or outlier handling: data conversion to other (calculated) attributes like volume, dimensionality, nearness and direction of movement is as important. Most of these new attributes are of interest because of their direct link with the emotional content of music.

Previous paragraphs stressed the role of the individual experience of music, but music is not only an individual experience it is also a social happening. Just think about the thousands of people watching a live pop concert. Our research advances the state of the art on this topic as well.. We introduce the concepts of coherence and consistency to describe these phenomena. With coherence we mean a group of people moving in a similar way on music. With consistency we refer to similar movement in distant time intervals. These concepts have a strong relationship with expressiveness. Expressiveness is something that can not be measured in absolute terms, it is rather a story of *less and more* by comparing to a reference. An obvious choice for reference is the group average as described in our research in chapter 3 (Expressive Body Movement Responses to Music are Coherent, Consistent, and Low Dimensional).

All the techniques and methods mentioned so far have one thing in common: *It is about decisions and comparisons.* It is about decisions: “What is the best model, the best prediction, . . . ?" and about comparisons: "How do we quantify *best*?" Crucial hereby, are definitions for equality and/or distances. The reader should be informed that many, many definitions exist. Our thesis uses just a small subset of the entire set of probabilistic and non-probabilistic definitions for equality and here is an overview of what we used:

- non-probabilistic:
    - between two mixtures : cosine distance
    - between two data points : euclidean norm
- probabilistic models:
    - between two data samples : the difference in probability according to a probabilistic model
    - between two probability distributions : f-divergences (i.e. Kullback-Leibler divergence)

To conclude this section, we summarize how this thesis is structured. Chapters 2, 3, 4 and 5 are extended versions of articles submitted to international peer-reviewed magazines. Chapter 6 is a little bit special as it gives a glimpse of the future based upon recent developments in the world of Machine Learning.

Hereafter follows a short summary of what can be found in chapters 2, 3, 4 and 5:

- Chapter 2 Towards E-Motion Based Music Retrieval.

  In this chapter we explain a new mechanism to retrieve music from a music library. The idea is to use arm movement to produce a playlist of songs. Tracking arm movement is nowadays feasible as a lot of mobile devices have movement sensors built in. The link between arm movement and music is done via an intermediate step using the valence and arousal plane. This intermediate step acts like a marshalling panel and allows flexibility. Once movement is translated in terms of valence and arousal, it is compared with an annotated (in terms of valence and arousal) music library.

- Chapter 3 Expressive Body Movement Responses to Music are Coherent, Consistent, and Low Dimensional.

  The results here allow to describe the expressiveness of spontaneous dance movement in a low dimensional model. The low dimensionality is due to the large coherence and consistency observed amongst dancers. A direct application is the selection of a subgroup of dancers having the same expressiveness out of a large group of dancers.

- Chapter 4 Beating-Time Gestures: Imitation Learning for Humanoid Robots.

  In this chapter we deduct from a set of conducting gestures (on a metronome) a generalized movement that can be diversified to a series of conducting gestures suitable for any type of music.

- Chapter 5 The Surprising Character of Music. A search for sparsity in music evoked body movements.

  This chapter is the prototype of unsupervised learning. Clustering techniques on positional and directional data give us a better understanding of a dance performance. This is formalized in a *directogram* that can be used as a descriptor for a musical excerpt. The clustering techniques help to identify sparse movement that can be understood as an indicator for the surprising character of music.

References

[1] M. Leman. *Embodied music cognition and mediation technology.* The MIT Press, 2008.

[2] Y. Björnsson, V. Hafsteinsson, A Jóhannsson, and E. Jónsson. *Efficient use of reinforcement learning in a computer game.* Proceedings of International Journal of Intelligent Games & Simulation, 2008.

List of Publications

Publications in international journals

Amelynck, D., Grachten M., Van Noorden L., Leman M. (2012). Toward E-Motion-Based Music Retrieval. A Study of Affective Gesture Recognition. *IEEE Transactions on Affective Computing*, vol 3(2), p.250-259.

Amelynck, D., Maes, P.-J., Martens, J.P., Leman, M. (2014). Expressive Body Movement Responses to Music are Coherent, Consistent, and Low Dimensional. *IEEE Transactions on Cybernetics*, In Press

Amelynck, D., Maes, P.-J., Martens, J.P., Leman, M. (2014). Beating-Time gestures: Imitation Learning for Humanoid Robots.

Amelynck, D., Maes, P.-J,Leman, M. (2014). The Surprising Character of Music. A search for Sparsity in Music Evoked Body Movements. *European Conference on Data Analysis 2014*, Accepted

Maes, P.-J, Amelynck, D., and Leman, M. (2012). Dance-the-Music: An educational platform for the modeling, recognition and audiovisual monitoring of dance steps using spatiotemporal motion templates. *EURASIP Journal on Advances in Signal Processing*, 2012 p.1-16.

Maes, P.-J., Amelynck, D., Lesaffre, M., Arvind, D.K., and Leman, M. (2012). The 'Conducting Master': an interactive, real-time gesture monitoring system based on spatiotemporal motion templates. *International Journal of Human-Computer Interaction*,2013, 29(717) p.471-487.

Nijs, L., Coussement, P., Moens, B., Amelynck, D., Lesaffre, M., Leman, M. (2012). Interacting with the Music Paint Machine: relating the concepts of flow experience and presence. *Interacting with Computers*, vol 24(4), p.237-250.

Without A Theory The Facts Are Silent.

Friederich A. von Hayek, 1974

1 Introduction

The content of this thesis is perfectly summarized by its title, namely "The analysis of bodily gestures in response to music". Measurement and analysis of music evoked body movement is indeed the central theme. This might not be considered as a breakthrough but it gains importance if we place it in a theoretical perspective. The theoretical framework used, is referred to in the subtitle of this thesis, namely "Methods for embodied music cognition based on machine learning". So, prior to explaining experiments and results, let us start with a brief introduction on cognition.

Heylighen [1] defines cognitive science as the modern science of the mind: "Cognition derives from the Latin verb cognoscere, which means *get to know*. This means that cognition focuses on knowledge, albeit not as a static substance or *thing*, but as a process. More generally, when we speak about cognition we are focusing on the mind as an information processor, i.e. a system that acquires, uses and transforms information."

The traditional, sometimes called the naive, view of cognition is best understood by the theory of Descartes. Descartes understands cognition by proposing two independent realms, namely mind and matter. While matter follows the laws of mechanics, mind has a logic of its own that cannot be reduced to mechanical principles. This philosophy is known as dualism.

A recent approach to cognition is called the embodied cognition. The main argument is that cognitive systems need to have the equivalent of a body through which they can interact with their environment [2, 3]. Enactive cognitive science [4] is a closely related concept: thought or knowledge only becomes meaningful when it is implemented, “acted out”, or enacted via interaction with the environment. Even perception can be seen as perceptually guided action: We don’t perceive an apple but an apple-to-eat. This is what is called by Gibson an affordance [5].

For readers who want more information about cognitive science in general we refer to the vast amount of books on this topic. For detailed information on *music* cognition we refer to the work of Leman [6]. Still, because of its importance for this thesis, an overview of embodied music cognition is given in the next section (Section 1.1).

Figure 1.1 on the next page explains how this Introduction Chapter is structured. The Chapter starts with an overview of *embodied music cognition* (Section 1.1), followed by an explanation of some of its concepts, namely Gestural Descriptors, Key Points/Goal Points, and gestures in Social Context (Section 1.2). Prior to analyzing these concepts we have to face a number of data pre-processing challenges (Section 1.3). The analyses are eventually realized by using a variety of Machine Learning methods which are introduced in general in Section 1.4 and which are subsequently applied in Section 1.5. The final section 1.6 discusses the results of the experiments.

1.1 Fundamentals of Embodied Music Cognition

The power of music as non-verbal expressive communication system is widely recognized [6–9]. Yet, the mechanisms that support the encoding and decoding of musical expression are still poorly understood.

In recent work, it has been suggested that gestures play a central role in the encoding and decoding of musical expression [6, 10]. Gestures are believed to facilitate the non-verbal expression and communication of emotions, feelings, ideas and intentions [11], both in music playing and music listening. When playing music, the expressive patterns of a gesture are encoded into sound, typically through the use of a musical instrument. The structural features inherent to a musical composition (e.g., melodic lines, rhythm, etc.) combined with the expressive performance of a musician (e.g., timing, dynamics, etc.) create, what has been called “moving sonic forms” [12] that reflect expressive gestural characteristics. When listening to music, people can decode these expressive gestural characteristics through the moving sonic forms back into actual movement patterns [10, 13, 14]. By internal simulation and/or actual performance, these movement patterns can be further connected to other modes with which actions are typically associated, like emotions, situations, images. Correspondingly, music is experienced and understood as intentionally, expressively, and semantically

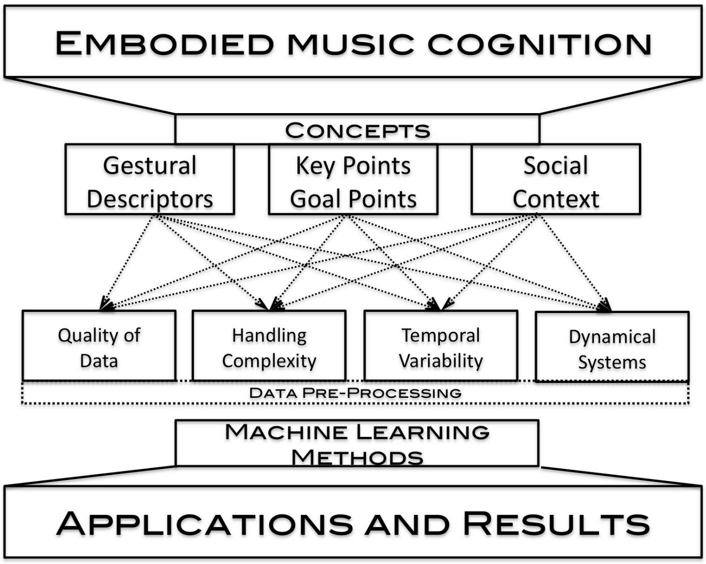


Figure 1.1: Overview

Figure 1.2: Brain in a vat

meaningful.

In this thesis, we focus exclusively on the decoding of musical expression, leaving aside the encoding aspect. The goal of this thesis is to provide a way of describing and modeling the way in which listeners decode this expression. In particular, we are interested in (a) the description of the fine-temporal structure of the decoding of musical expression through gestures and in (b) extending the study of individual embodied music cognition to social embodied music cognition, which implies a focus on what is common and different among subjects.

The background of our approach is contained in the viewpoint that human interaction with music is embodied [3] [4]. We see the human mind as the seat of our personal and musical experience and we adopt the idea that these experiences are expressed in the peculiarities of the human body that mediates information from and to the musical environment. In that sense the body is more than just an interface with sensors and actuators in the environment (as stated in the traditional cognitive science). Instead, the body is an expressive mediator of the interaction. This expressiveness is rooted in an action-perception engine that draws upon a repertoire that makes interaction effectively conveying and predictive. Owing to this engine, there is a tight connection between music perception and body movement in the sense that the perception of conveying properties of music can be straightforwardly expressed into accompanying movements, or gestures.

Note that the above viewpoint stands in sharp contract with most of traditional cognitive science, which, according to Heylighen [15], tends to see the mind as separated from the outside world. He says [1]: “Even when modern science admits that the mind cannot exist independently of the matter in the brain, the assumption is that the brain alone is sufficient to produce intelligent behavior. This leads us to envisage a theoretical disembodied intelligence as a “brain in a vat”, a brain artificially kept alive connected to some electrodes that stimulate it, but with no body attached to it (Fig.1.2). Critics claim that such a brain would not be able to exhibit intelligence, because intelligence evolved to pursuit interaction with the world.”

To understand the theory of embodied music cognition, and the concept of music gesture that forms a core part in the encoding and decoding of music, we draw upon the work of our promotor Prof. dr. Marc Leman, in particular [6]. According to Leman, interaction with music is embodied (i.e. mediated by the body), situated (i.e. embedded in an environment), and enacted (i.e. put into practice through action and gestures) [4] [16]. In The Power of Music [17], he provides a dynamic model inspired by control theory that serves us as a starting point for the modeling framework that will be further developed in this thesis. He considers the study of the dynamics of the action-perception coupling system a hot research topic in modern cognitive science and his work forms a rich source of inspiration for this thesis.

Figure 1.3 shows the basic scheme of embodied music cognition. The figure represents an agent that acts in the environment and that receives information from that environment. Agents for spontaneous dance movement require an additional entry in the basic scheme as their action is triggered by the environment, i.e. by enactive (musical) perception (0). Enactive perception is due to the role of sensory-motor engagement in musical experience. Music is namely perceived in terms of actions (affordances) [18].

When the agent decides to perform an action (e.g. the intended playing of a note on a music instrument) it will rely on the action repertoire (1) to launch an action pattern that gets executed (2). Along that pathway, a copy of the action pattern and its predicted outcome (3) is made and compared (4) with the actual action execution and the sensed outcome of the action (the instrument’s sound) (5) and (6). Based on the perceived outcome of the action the motor pattern can be adjusted. Leman makes a distinction between two mechanisms for adjustment of body movements, called the sensor-motor loop and the action-perception loop. The sensor-motor loop (7) is a low-level circuit where the motor activity is basically driven by sensory input from the environment. In contrast, the action-perception loop (8) is a high level-circuit that involves the action repertoire.

The model is linked with several concepts that play a role in the embodied music cognition paradigm. A core concept is the repertoire. According to Leman a repertoire of actions and action consequences is called an action-oriented ontology [17] (pg 22): "The action repertoire can be conceived as the reservoir of experiences, including experiences of expressiveness. This reservoir comprises connections between action commands, sensations of the external world through our senses (exterioception), but also sensations of our body movements (proprioception) and of our body state (interioception). Moreover, as corporeal articulations and actions are carried out in space and time, it seems natural to conceive the action repertoire as a container of spatial-temporal patterns."

The spatial-temporal patterns constitute the musical gestures and are the main topic of this research. The action repertoire is considered a component of a more complex mechanism that controls the interaction between environment and subjective experience. Leman [17] (pg 26): "This mechanism is called the action-perception coupling system, or the action-perception engine, and is responsible for prediction and for issues that involve musical intentions. It is the circular flow of information that takes place between the subject and the real world in the course of a sensory-guided sequence of behavior towards a goal. Each action causes changes in the environment that are perceived and that lead to the processing of further actions. The latter cause new changes that are analyzed and lead to new actions, and so the cycle continues."

The dynamic model is about action-perception-based interactions with the environment, which, from an observer's viewpoint are called body movements in response to stimuli. However, interactions with the environment may filter gradually to the human mind of the agent, which is then able to conceptualize these experiences and perhaps answer questions concerning the intended nature of the observed movements. In his book Musical Gestures - Sound, Movement and Meaning [19] Leman devotes particular attention to the description of gesture-based subjective experiences (pg 139): "With respect to the effect of gesture on experience, there are three types of personal experience that deserve some particular attention: namely the experience of flow, the experience of presence and the experience of cause-effect. The experience of flow [20] can be characterized as an experience in which the subject's skills are fully preoccupied with a task. Presence can be defined as the illusion of non-mediation [21]. This illusion may occur when your musical instrument is no longer considered as an obtrusive object but as an instrument that gives you a way of expressing yourself in music. The

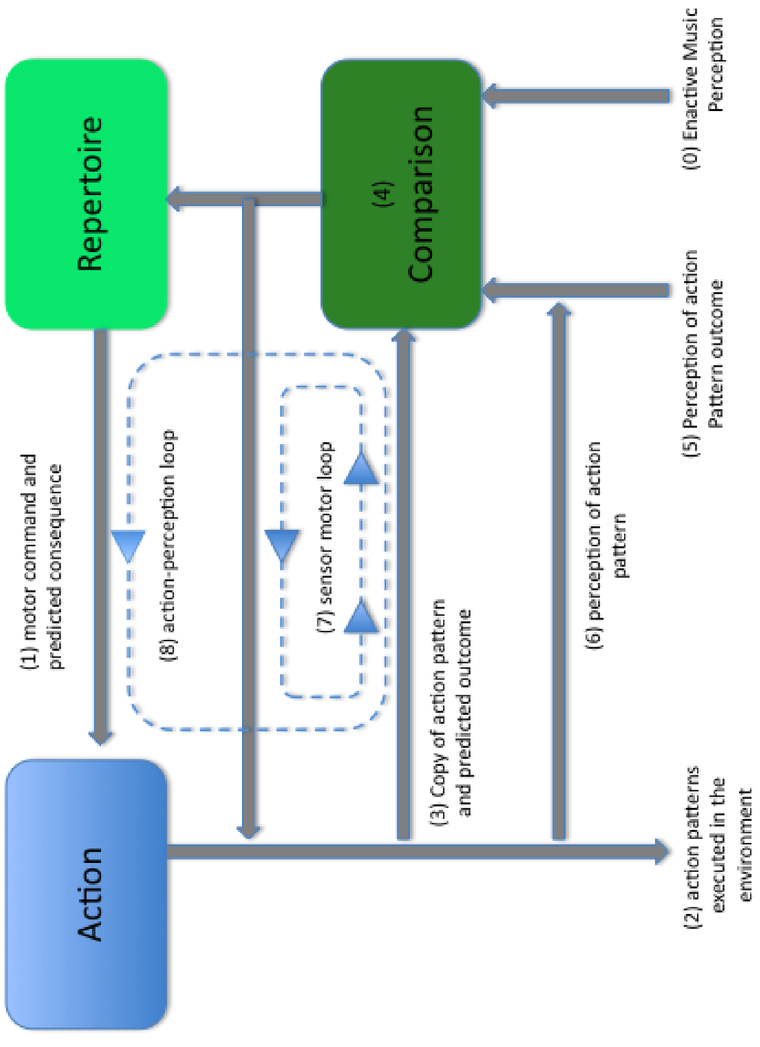

Repertoire
(4)
Comparison
Action
(0) Enactive Music Perception
(5) Perception of action Pattern outcome
(1) motor command and predicted consequence
(8) action-perception loop
(7) sensor motor loop
(3) Copy of action pattern and predicted outcome
(6) perception of action pattern
(2) action patterns executed in the environment

Figure 1.3: Action-Perception coupling system. See text for full explanation.

perception of a cause-effect relationship, in the domain of music perception and gesture, can be considered an experience of the cause of a sound from a gestural perspective, rather than a conceptual understanding of the causality relationship as such. One hears the sound of moving feet, rather than its acoustic properties."

A major appeal of the embodied music cognition theory is that it provides a consistent way of linking experiences with body movement, music perception and the musical environment. The theory is in agreement with the idea that body movements (called gestures, see below) express core aspects of the expressions and intentions of perceived music. These body movements can be measured in an objective way. To interpret these measurements we can link them with verbal reports of experiences. Our contribution in this thesis is to propose a way to describe how gestures evolve over time and how gestures can be dealt with as a characteristic of groups of people (rather than just an individual person).

1.2 Main Concepts for Musical Gestures Research

The above fundamentals of the embodied music cognition paradigm provide the background for understanding our approach. We now relate embodied music cognition to a set of concepts that play a role in our research. These concepts are subdivided into three categories called: (i) Laban descriptors for gestures, (ii) key points and goal points as landmarks for musical gestures, and (iii) descriptors for gestures in the social context.

Let us first define what we understand by gestures in general and by a musical gesture in particular. There exist many definitions for gestures and the definitions differ from the field where they are used. For our work we use the definition of gesture from Leman and Godøy [22] , namely, that "a gesture is a movement of part of the body with the goal to express an idea or meaning". Like Jensenius et al. state in [23]: "When speaking about the musical activity of musicians and dancers it is tempting to call the involved embodiment *gestures* rather than *movements.* The main reason for doing so is that the notion of gesture somehow blurs the distinction between movement and meaning. Movement denotes physical displacement of an object in space, whereas meaning denotes the mental activation of an experience". Integrating movement (as a display of matter) and meaning (as a display of mind) is exactly what is at the core of the embodied music cognition. That makes research on musical gestures the main topic of this thesis and descriptors for gestures can help here to give insight.

Consequently, we focus on concepts that facilitate the link (correlation) between the so-called second-person descriptions and third-person descriptions. The first-person description is about the subjective experience of intentions attributed to music, the second-person description about the corporeal articulation of these intentions and the third-person is about objective measurements, either measuring musical signal properties or measuring movement properties [6]. Main advantage of this focus is that it helps to understand the link between movement and musical intentionality.

1.2.1 Laban descriptors for gestures

In this section, we present a set of descriptors, relevant to our empirical research. The set is not exhaustive but it is rather a list of novel descriptors generated in concertation with musicologists. For the reader's convenience we use the categorization from Laban's Movement Analysis (LMA) [24], even if we do not consider all categories in this thesis. LMA distinguishes the following four categories:

- Body descriptors for structural and physical characteristics of the human body while moving, such as which body parts are moving, which body parts are connected.

- Effort descriptors that handle the dynamics of movement and are closely related to the energetic intention of the movement.

- Space descriptors that mark out the kinesphere (the area within which the body is moving), the spatial intention (the directions or points in space the mover is identifying or using) and the geometrical observations of where the movement happens.

- Shape descriptors that give an account of how the body changes shape during the movement. It comprises the description of static shapes as well as the description of the dynamics of these shapes.

To describe gestures in response to music, we focus in this thesis on two categories that are readily feasible and measurable, namely space and shape.

1.2.1.1 Descriptors for Space

The descriptors we discuss here are: concentration, volume, elevation and dimensionality.

1. Concentration

   Definition: Concentration describes the phenomenon that some locations and/or directions are more and longer frequented than others.

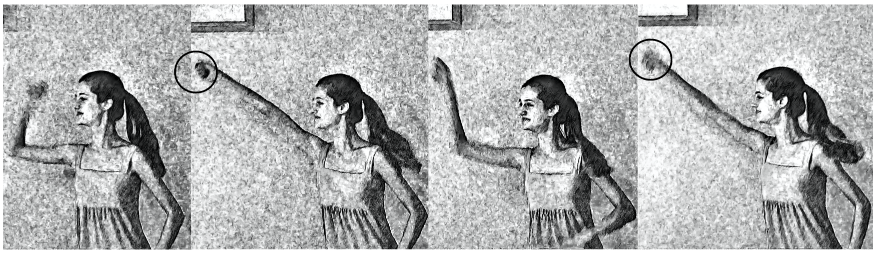


Figure 1.4: Illustration of the concept "concentration". Concentration is here visualized as a small dedicated area (indicated by a circle) that is longer frequented than others.

Musical Relevance: Concentration means that some positional areas are more important than others. A first motivation could be found in the gestural affordances of musical sound [18]. Godøy links musical sound features with shapes that may be gesturally rendered. Among examples of sound features he gives "accents and articulations", having very clear gestural requirements of energetic motions, and "cyclical patterns", i.e. grouping of sonic events, such as in meter, resulting in recurrent gestures. A second motivation is found in the theory of goal points [25] where gestures are considered as goal-directed. Musical movement is then considered as a succession of clear discontinuous postures and continuous motion between those postures. These postures are reference points for musical gestures what makes us suppose that they occur in areas that are more or longer frequented.

Visualization : Fig. 1.4

2. Volume.

   Definition: We define volume as the volume inside the convex hull or convex envelope of the movement trajectories deployed in space.

   Musical Relevance: The link between the volume of a gesture and particular musical characteristics is for example represented in Hodgins' model [26], which is a model that points to a number of choreomusical parallels between features in dance and music. The dynamic parallel as he calls it relates the size and volume of dance movements to dynamics in musical sound (such as intensity and loudness). Another link with music is found in LMA (Laban Movement Analysis). LMA considers volume as a measurement of the kinesphere or personal space, the area around the body within reaching capabilities of the limbs without changing place. Certain emotional affects can be made visible by studying the kinesphere as demonstrated in the

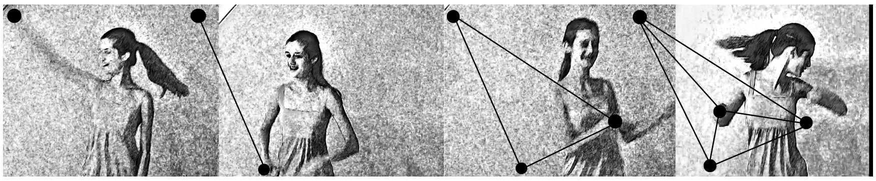


Figure 1.5: Illustration of the concept "volume". Volume is visualized here as the volume inside a convex hull made-up by consecutive right hand movement. In the shown example, it is so that the four right hand poses make up a convex hull, defining a volume (being a tetraeder in this case). All other intermediate poses (not shown) fall inside this volume and are discarded for volume calculation.

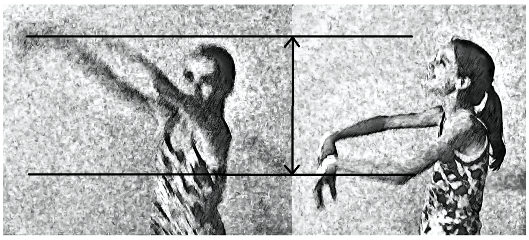


Figure 1.6: Illustration of the concept "elevation". It is here visualized as a height difference between two right hand poses but it can also be expressed as difference in height between any limbs, or as a difference between minima and maxima in height taken over some time intervals.

work from Camurri [27]. Or as Ruud [28] says :"Listening to music or playing an instrument seems to lead to an awareness of space within oneself which is totally distinct and not accessible to other people. Sometimes this is called the true-self which may be dramatically met by a sudden mood in the music, or by a voice, or an artist.

Visualization : Fig. 1.5

3. Elevation.

   Definition: Elevation refers to the spatial height of musical gestures.

   Musical Relevance: Elevation of musical gestures links with musical emotion and referring to literature (for example [29] [30]) it is found that happiness is associated with higher elevation in the movement. Beside happiness elevation correlates also with pitch. Pitch is classified in many languages by using terms that have a spatial connotation referring to elevation such as *high* and *low* (e.g. in Chinese, English, French, German, Italian, Polish and Spanish) [31].

   Visualization : Fig. 1.6

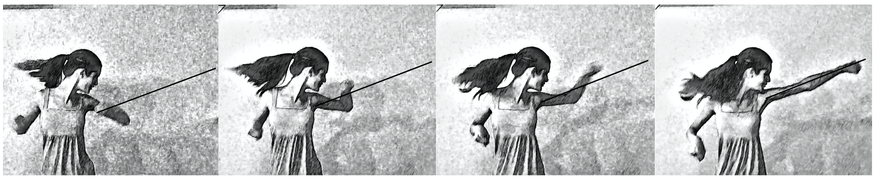


Figure 1.7: Illustration of the concept "dimensionality" for musical gestures. The musical gesture shown in the above figure nears a straight line and has consequently a dimensionality of one.

4. Dimensionality.

   Definition: With dimensionality of movement we have a measure that indicates if movement merely goes along a line (one dimensional), or in a plane (two dimensional) or if it covers the full three-dimensional space. Although this definition is valuable, it can be broadened to include movement in relation to other geometrical shapes, like circles or surfaces of ellipsoids. A possible measure for dimensionality is then to calculate the fractal dimension of the movement [32], which is a measure of the number of active variables required to model the dynamics.

   Musical Relevance: The musical relevance of dimensionality is backed up by previous research work [33] [30] [34] [35]. Results show for example that sad music is reflected in rather simple movements of low dimensionality.

   Visualization : Fig. 1.7

1.2.1.2 Descriptors for Shape

We discuss two descriptors of shape, namely proximity and direction.

1. Proximity

   Definition: Commonly, proximity is defined as nearness in space (http://dictionary.com) but here, as a shape parameter, it refers to all limbs, being folded on the body.

   Musical Relevance: The shape category has a subcategory called *shape-forms* that describes the static shapes a body takes, such as ball-like or wall-like. Proximity is then described by changes in *shape-forms*, e.g. when the body is currently opening (growing larger with more extension) or closing (shrinking with more flexion). These changes in shape-forms relate to musical intentions like intimacy. The link be-

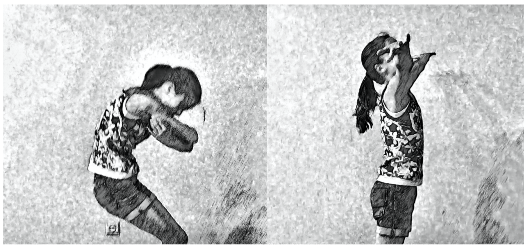


Figure 1.8: Illustration of the concept "proximity" for musical gestures. The two pictures show the two extremes for proximity: a high degree of proximity where every limb directs at the body-center and a low degree of proximity where all limbs point away.

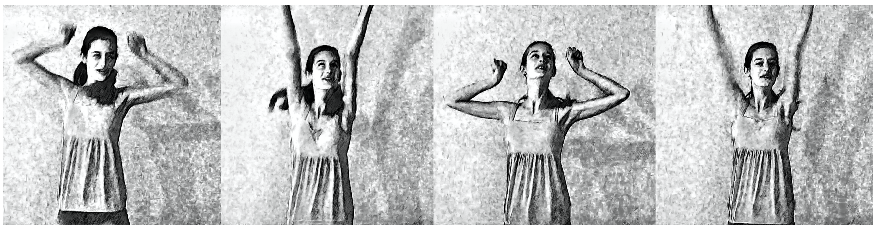


Figure 1.9: Illustration of the concept "direction" for musical gestures.The vertical movement shown here is the embodiment for power.

tween emotional response to music and intimacy can be found in for example work from Bicknell [36].

Visualization : Fig. 1.8

2. Direction

   Definition: Firstly, in absolute terms we can define direction in terms of vertical versus horizontal movement. Secondly, in relative terms we interpret direction as convergent-divergent movement (towards or away from one's own body) versus equidistant movement (at same distance of the body).

   Musical Relevance: Horizontal/vertical movement is linked to concepts of power [37] and musical style [38]. Convergent-divergent movement is classified, according to Laban's terminology (shape category), respectively as Spoke-Like and Arc-Like. These movements are found to be basic gestures of dance styles [39].

   Visualization : Fig. 1.9

1.2.2 Key Points and Goal Points, landmarks for musical gestures

A second category of descriptors that we use in this thesis relates to landmarks for musical gestures. A first motivation for this new set of descriptors is that gestures may reveal particular aspects of their intention, expression and/or meaning at particular points (landmarks) in their deployment. A second motivation lays in the ease of memorizing and comparing movement trajectories: landmarks reduce movement trajectories to a set of frames of reference (cfr [39] where this concept was introduced). In this sense landmarks are key contributors to the "repertoire"-node and the "comparison"-node in the action-perception coupling diagram (Fig. 1.3).

1.2.2.1 Spatial Landmarks - Key Points

Definition: Originally, a landmark literally meant a geographic feature used by explorers and others to find their way back or through an area (http://en.wikipedia.org). This definition comes very close to what we want to achieve with our concept of key points, namely a distinct set of samples that represents the essence of a musical gesture as it is deployed. These samples are based upon movement characteristics and we refer to them as key points or spatial landmarks. Extrema (minima and maxima) are straightforward examples of such movement characteristics but other more complex features like for example state changes (a combination of position and speed changes) can be used as well.

Musical Relevance : The concept of key points reduces a movement trajectory to its most simple form, namely a set of key points. This has as advantage that it is easy to memorize movement trajectories and that it helps comparing them, as they are reduced to a set of frames of reference (cfr [39] where this concept was introduced).

Visualization: See Fig. 1.10

1.2.2.2 Temporal Landmarks - Goal Points

Definition: A second and also common definition for a landmark (http://dictionary.com) associates a landmark with something used to mark the boundary of land. Elaborating on this definition, we define now temporal landmarks as indicators that chunk a gesture in time. Obviously, the chunks are not arbitrarily chosen but are defined by musical events. We define the time stamps of these events and the postures at these events as

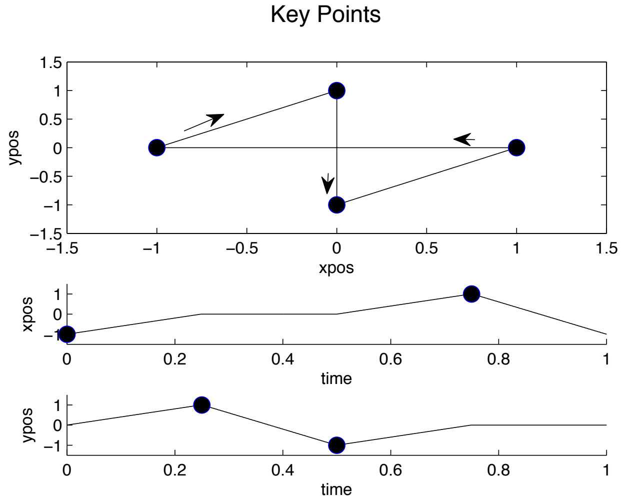


Figure 1.10: Illustration of the concept 'key points' for a gesture. A gesture's trajectory and its key points (big dots) are displayed in a two-dimensional plane at the top plot. The arrows indicate how the gesture progresses in time. The key points in this simple example are determined by the minima and maxima over time in the x- and y- coordinates (bottom plots). They reduce the gesture to a set of 4 samples

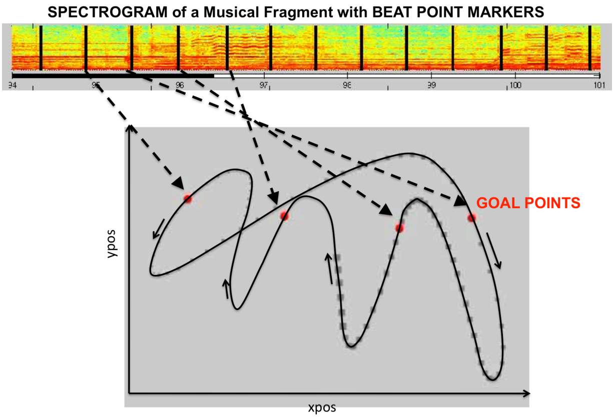


Figure 1.11: Illustration of the concept 'goal points' for a gesture. A gesture's trajectory and its goal points are displayed in a two-dimensional plane at the bottom plot. The arrows indicate how the gesture progresses in time. The goal points are determined by temporal landmarks identified by musical characteristics, being here the beat points of a musical fragment shown at the top.

goal-points or temporal landmarks.

Musical Relevance: Our definition of temporal landmarks links to the theory of goal-points from Godøy [25]. Godøy interprets movement as a combination of discontinuous postures and continuous motion between those postures [40]. He suggests to take downbeats or other accented points in the music as goal points. Naveda and Leman [39] adhere to this concept projecting beats onto a dance gesture and using beat times as temporal landmarks.

Visualization: See Fig. 1.11

1.2.3 Descriptors for gestures in social context

A third category of descriptors used in this thesis is related to the social context. For this category we define concepts like expressiveness and coherence/consistency.

1.2.3.1 Expressiveness

Definition: Traditionally, expressiveness is defined as a deviation from a regular or neutral performance [41]. The terms *deviation* and *regular* have

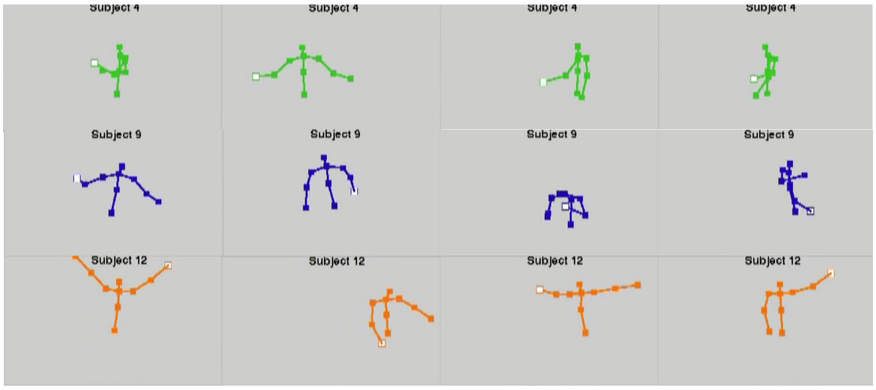


Figure 1.12: Illustration of expressiveness. We defined expressiveness as variation around a reference. To illustrate we present here four snapshots for in total three subjects taken from a dance performance. The snapshots are taken at identical time stamps. If we take subject 9 as the reference then we see that subject 4 is less expressive and subject 12 more expressive.

however negative connotations in the sense that they imply that regular is the norm and expressiveness is the deviation. We believe that it is just the other way around, namely, that expressiveness is the norm and therefore we use the terms *variation* and *reference* instead. Summarized, our definition for expressiveness is that it stands for variation around a reference.

Musical Relevance: The original definition for expressiveness comes from music psychology pioneer Carl Seashore [41] and he described *deviations from the regular* for sound properties such as loudness, tempo (rubato), articulation and intonation. This idea has been followed by many researchers since that time [42]. Davidson [43] added the idea that performances are embodied. She argued that each movement type (for instance, the wiggle) can be executed in a range of ways that give the potential for a range of expressiveness levels to be elicited. The findings from Seashore and Davidson concern music performances but similar results are found in work on dance performances. For example, Camurri et al. [44] studied expressive gestures as gestures superimposing expressive content (deviation) to normal gestures (regular).

Visualization: See Fig. 1.12

1.2.3.2 Coherence and Consistency

We use coherence and consistency to describe the behavior of a group in terms of levels of expressiveness. The description is based upon ordering group-subjects and discovering how well this ordering is kept over time. It allows to answer questions like: “Is the most expressive subject the most expressive subject over a whole time interval ?”. Note that for ease of understanding we used the term “ordering”, whereas the mathematical correct term is “correlating”.

Definition Coherence: Coherence stands for high correlations between levels of expressiveness in a group at every two distinct time stamps in a continuous time interval. We define then the performances as coherent in this continuous time interval. It implies (1) synchronicity between subjects within this continuous time interval and it requires (2) preserving the *ordering* of subjects in levels of expressiveness. For our research we define coherence and consistency in conjunction with expressiveness. Changes to the definition are obvious for research requiring other gestural characteristics .

Definition Consistency: Consistency stands for high correlations between time stamps in distinct coherent time intervals.

Musical Relevance: Music has an interesting relationship with social cognition. In fact, music has been compared with a virtual agent with whom the listener dynamically and socially interacts [45] [6]. The neural mechanism underlying this phenomenon (especially the action-perception coupling) has been related to the mirror neuron system [46](see Sevdalis, V., & Keller, P. E. [47] for an overview), whereas the behavioral study of the movement deployment has been related to synchrony [48]. In social cognition, as in music there is an important component of non-verbal communication that is based on synchrony, which can be understood in relation to the action-perception couplings that engage in the understanding of communicative signals and social adaptation behavior [49]. Synchronization is also a basic principle of musical entrainment, where two systems engaged in synchronization adapt to each other [50].

Visualization: See Fig. 1.13

Figure 1.13: Illustration of coherence for musical gestures. Coherence stands for high correlations between subject performances at every two distinct time stamps in a continuous time interval. It implies (1) synchronicity between subjects and it requires (2) preserving the ordering of subjects in levels of expressiveness. The most extrovert dancer is the most extrovert dancer over the whole time interval.

1.3 Challenges

1.3.1 Variability and Constraints

There exists some confusion in the literature between the use of the terms variability and variation. Here we follow the vision of Van Belle [51] describing variability and uncertainty as two different categories of variation, involving different sources and kinds of randomness. So, the term variability refers to natural variation in some quantity whereas uncertainty refers to the degree of precision with which a quantity is measured.

We do not discuss uncertainty in this thesis but it should be clear to the reader that if there is a choice between measurement equipment, precision (or less uncertainty) is an important decision maker. In this section we focus primarily on variability.

Applying best practices needs some work upfront, namely before the actual execution of the experiment. Important tasks are the identification and reduction of all sources of variability. To identify the sources of variability we refer to the variability model of Desmet [52]. It distinguishes four main sources and their interactions:

- Human variability: originates from neurological, skeletal, and muscular variability in people producing musical gestures.
- Device variability: this comes from devices used in the experiment. Think for example about the markers fetched to the body of a dancer for data collection: are they always fetched at the exact same spot in longitudinal studies?
- Sonic variability: variability in the sonic forms can come from the mediator (instrument or the intentions of a performer). So in a sense it is part of human and device variability but the reason why we

make a separate class for it is that here we focus on the intentionality. Sonic variability stands for variability by intentionality. For example, a single musical fragment can have a combination of fast and slow tempo's. This is variability intended by the composer.

- Environmental variability: with environmental variability we refer to items like temperature in the experimentation rooms, presence or absence of daylight, organizing morning or evening sessions. All these are factors that can have impact on the results of various experiments.

Interactions: all the above sources of variability are not necessarily independent but can be inter- and intra-correlated. For example human-human variability (intra-correlated human variability) is a factor to deal with in experiments with a group of people (we refer for example to [53]).

In section 1.2.3.1 we defined expressiveness as variation (variability) around a reference. It is clear that in research on expressiveness we do not want to reduce this source of variability but we want to keep it to its full extent. Constraints can hinder this process and have to be investigated or removed:

- Human constraints: wearing a motion-capture suit can hinder certain movements, hence reduce expressiveness.

- Device constraints: the equipment required for fMRI causes constraints for what experiments one can do (e.g. the use of metallic music instruments is impossible).

- Sonic constraints: cochlear implants are designed for the 0-4KHZ range (speech) but not for the high frequency range.

- Environmental constraints: the experimentation lab might have limited accessibility and availability.

- Data Analysis: some data models require lots of data, or need a minimum number of participants. Other restrictions can come from lack of computing power or even from the current state of science having no suitable data models for an envisioned experiment.

A major consequence of having these constraints is that it reduces the relevance of the results or that it does not allow to generalize to an ecological setting.

1.3.2 Need for Data Pre-Processing

After an experiment we are confronted with a set of challenges that need to be handled before an analysis can take place. A first challenge lays in the quality of the measured data. The original signal can be blurred by noise, measurement errors, ... and these problems have to be adequately handled. A second challenge concerns methods to reduce complexity. The problem is that an experiment gathers large amounts of data collected over considerable time intervals from several sensors attached to different body parts of several subjects. A third challenge is rather specific to our research of musical gestures. Musical gestures experience two main sources of variation, namely spatial and temporal variation and that makes it difficult to analyze. The third challenge lays then in dealing with these sources of variation.

1.3.2.1 Quality of Data

The first major challenge concerns the representation and quality-checking of the collected data. This includes data cleaning, normalization, transformation, feature extraction and feature selection, etc. [54]. Commonly used techniques here, are outlier detection and handling and noise filtering. We refer to the literature for more information on these topics.

1.3.2.2 Handling Complexity

The second challenge relates to the complexity of our process. A proven method in analyzing complex problems with an abundant number of variables collected over considerable time intervals is breaking down the problem into smaller, more manageable parts. For breaking down, two techniques are used here, namely segmentation and decomposition. They can either be applied separately or even jointly:

Complexity reduction by segmentation

Segmentation means time chunking of musical gestures in segments. This leads to the concepts of elementary gestures, segment boundaries and transitions that all can be studied individually and that eventually can be brought together to understand the full picture.

Elementary Gestures

A musical gesture can be understood as a concatenation of gestural components. The set of gestural components is not endless as some of the components are recurring. These distinct components are called elementary gestures and constitute a gesture dictionary.

Segment Boundaries

The result of the segmentation process is a set of distinct elementary gestures. However, not only the elementary gestures are of interest but also the segment boundaries can provide us with insights. Segment boundaries can be linked to the previously discussed concepts of key points and goal points.

Transitions

Transitions stand for smoothness at the segment boundaries. This is the case where a subject adapts the end of an elementary gesture to prepare for the beginning of the next elementary gesture, comparable to a cross-fade effect used by DJ's to move to a new song. Note that some authors will refer to this effect as coarticulation and this because of the gestural context (http://en.wikipedia.org/wiki/Coarticulation). From the viewpoint of analysis, transitions blur the notion of segment boundaries. It is therefore difficult to identify the exact location of segment boundaries.

All the above describes the results of a segmentation process (elementary gestures, segment boundaries, transitions). The process itself and more precisely how the chunking is done is also worthwhile explaining. We distinguish basically two methods depending on the input used for segmentation: either data external to movement data can be used as input for segmentation or either the movement data itself can be used as input.

Segmentation using external data

Segmentation of musical gestures is in essence time chunking of the gestures. An easy and obvious way to time chunk is by just dividing time into small intervals of equal length. This results in the investigation of gestural components that are of equal time duration. In a way this is analog to a short-time Fourier transform (STFT) where also fixed time intervals are analyzed.

A more advanced way starts with noticing that the timing of musical gestures is highly determined by the music. This leads to using the musical signal as another external source for segmentation. A straightforward solution is to use the beat time stamps from the musical signal as segmentation input. An illustrative (one dimensional) example of this method of segmentation is displayed in Fig.1.14.

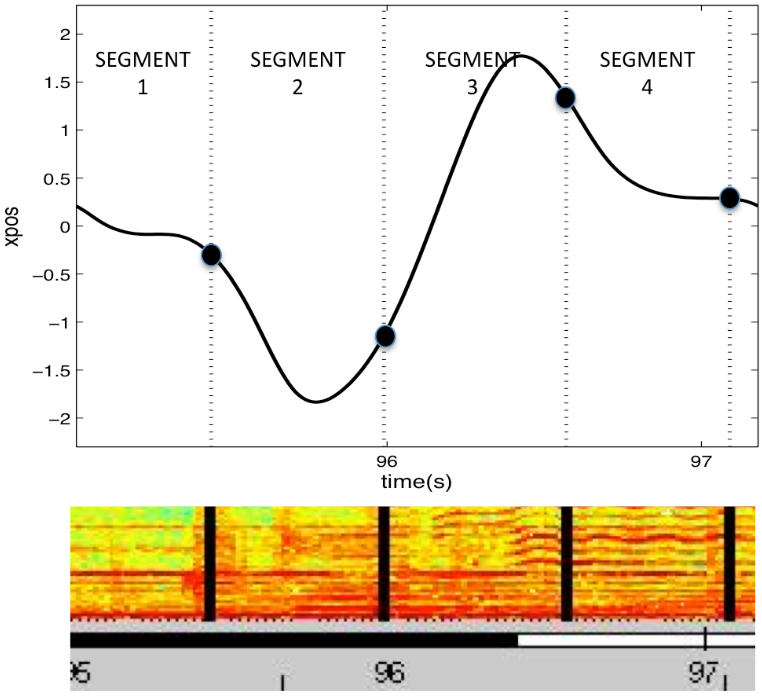


Figure 1.14: Example of segmentation of a one dimensional movement signal by beat time stamps taken from a musical fragment. The bottom picture displays the position of the time stamps of the beats in the musical fragment. The beat time stamps are displayed on the spectrogram of the musical fragment and this for the convenience of the reader. The top picture illustrates the segmentation process using these beat time stamps. (Extension from one dimensional to multi dimensional movement data is straightforward).

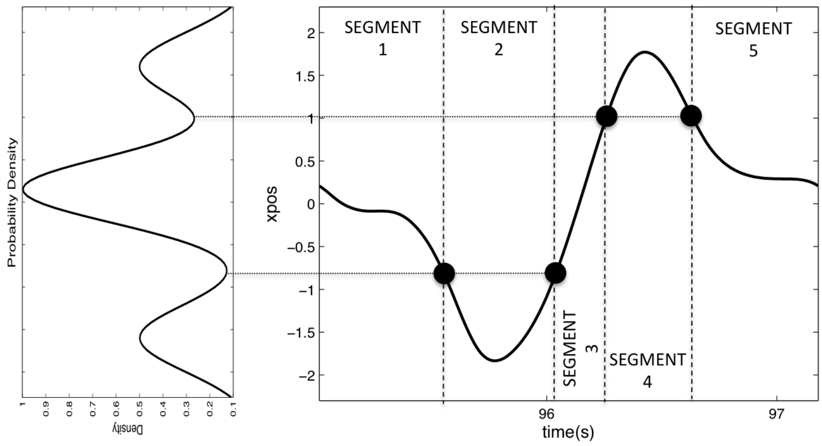


Figure 1.15: Example of segmentation of a one dimensional movement signal by taking information from the movement data itself. In the left picture the density of the movement data is displayed (rotated counterclockwise over 90°). This information is used in the right picture to segment the movement data. The segmentation boundaries correspond with areas of low density. (Extension from one dimensional to multi dimensional movement data is straightforward).

Segmentation using movement data

This type of segmentation uses the movement data itself to segment the gestures. An example is to use spatial landmarks to define segment boundaries. This can be achieved for example by means of a heuristic algorithm identifying local minima in the velocity signal or locating extreme amplitudes. More advanced methods look at minima and maxima into the density curves of the movement variables. An illustrative example of this method is displayed in Fig.1.15.

Musical Relevance of segmentation

Segmentation is a well known technique by musicologists. Segmentation lays for example at the basis of the Labanotation (Kinetography) and it is also strongly related to the concept of basic gestures introduced by Van Noorden [55]. Understanding musical gestures as concatenations of elementary gestural components is for example the starting hypothesis in a study on guqin performance [56] and in a study on a clarinetist's performance [57]. Other examples are [44] [58]. These studies reveal that sonic movement (identified in the music) reflects sound-producing movement (hand movements), and that this movement can be understood as concatenations of elementary gestural components.
Using music (i.e. time) to segment gestures can be linked to Godøy's theory about goal points. Godøy states that both sound-producing and sound-accompanying movements are centered around certain salient events in the music such as downbeats, or various accent types, or melodic peaks [25] and which he calls goal points. In music performance, these goal-points are reflected in the positions and shapes of the performers' effectors (fingers, hands, arms, torso, etc.) at certain moments in time, similar to what is known as keyframes in animation. This idea is also found in the approach followed by Leman and Naveda [39]. In their Samba dance study they segmented movement signals using beat points as boundaries for segmentation intervals. The method resulted in a spatiotemporal model, an important aspect for the repertoire in the action-perception coupling system (Fig.1.3).

Complexity reduction by decomposition

Just like segmentation, decomposition reduces the complexity of an analysis. It does so not by time chunking but by *decomposing* a signal (gesture) $x(t)$ into a linear combination of other signals $\phi_k(t)$ (1.1). This approach works for the entire gesture time interval but can also be applied to the individual time chunks obtained after segmentation.

$$x(t) = \sum_{k=1}^{K} c_k \phi_k(t) \tag{1.1}$$

The signals $\phi_k(t)$ are called basis functions. To make analyses tractable the number of basis functions must be limited. The number of basis functions is usually calculated in a validation step. Further, there exists a wide variety of basis signals and we will not discuss them all. Here, we limit the discussion to a set of basis functions that are of interest to our research:

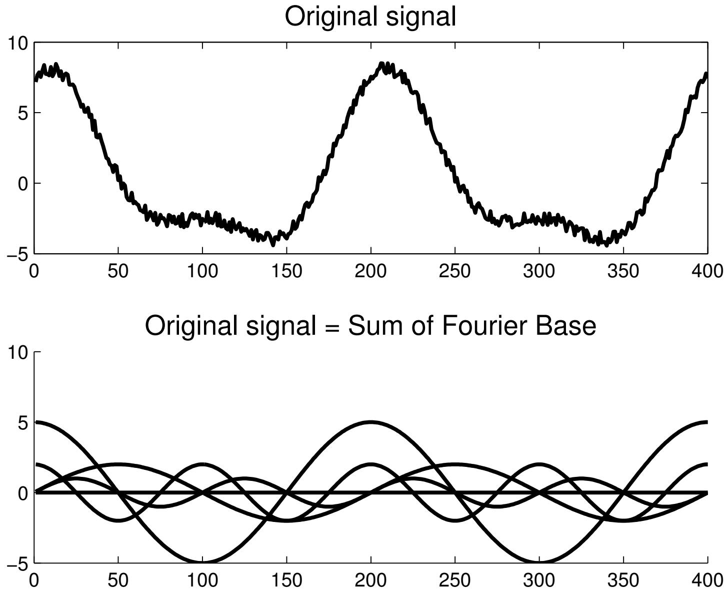


Figure 1.16: Example of a decomposition of a signal in Fourier basis components. The signal at the top plot can be approached by a linear combination of the 4 sine/cosine signals at the bottom plot.

Fourier Basis System

The basis functions are $\phi_0(t) = 1,\ \phi_{2k-1}(t) = sink\omega t,\ and\ \phi_{2k}(t) = cosk\omega t.$ They are periodic and smooth, as illustrated in Fig. 1.16. That makes the Fourier basis decomposition useful for smooth functions that exhibit some periodicity but inappropriate for functions with discontinuities or a-periodic behavior.

B-Spline Basis System

A spline is a polynomial specified over an interval and of order m. The order m of the polynomial determines the number of parameters and is one more than its degree, its highest power. Any function can be approximated by splitting the function in intervals and approximating the function in each interval by a spline. A way to improve this fit is to increase the number of time intervals (breakpoints) or the order of the splines. The location of breakpoints can be equally spaced (equal time intervals) or it can depend on the complexity of the curve. For equally spaced intervals the term *uniform splines* is used. An additional requirement can be that the fitted curve and its derivatives must be continuous at the breakpoints and this will reduce the degrees of freedom for the fit.

Now we discuss how this can be done in practice. In our research we used the B-spline basis system developed by de Boor [59] to implement B-spline decomposition. The system used by de Boor does not work with the individual splines but with basis functions. In his approach every basis function is a linear combination of splines and is positive over no more than m adjacent intervals. A basis function has also order m as it is a linear combination of m order splines. B-spline decomposition means then approximating a function by a linear combination of B-spline basis functions. An example of the 13 B-spline basis functions for an order 4 spline with 9 breakpoints is shown in Fig. 1.17. Spline functions are very common to approximate non-periodic functional data.

Empirical Basis System

The recommendation of using Fourier bases for periodic data and B-splines for non-periodic data underlines how important it is that basis systems match the characteristics of the data. A logical question is then whether we can build a basis system starting from the data. The answer is affirmative and usually this is constructed from a function principal component analysis (FPCA) as illustrated in Fig. 1.18. The basis functions are then eigenfunctions which usually lead to a compact function representation that models the variance observed in the considered time interval. Because this technique is a variance based technique it should be clear to the reader that it requires multiple realizations of a function (either coming from many subjects or either many realizations coming from one subject).

In practice the empirical basis system is combined with either a Fourier Basis System or a B-Spline Basis System. That means that the original function and also its eigenfunctions are expressed in terms of either a Fourier basis system or either a B-spline basis system from which the empirical bases are calculated. This has as immediate consequence that also the eigenfunctions are smooth functions.

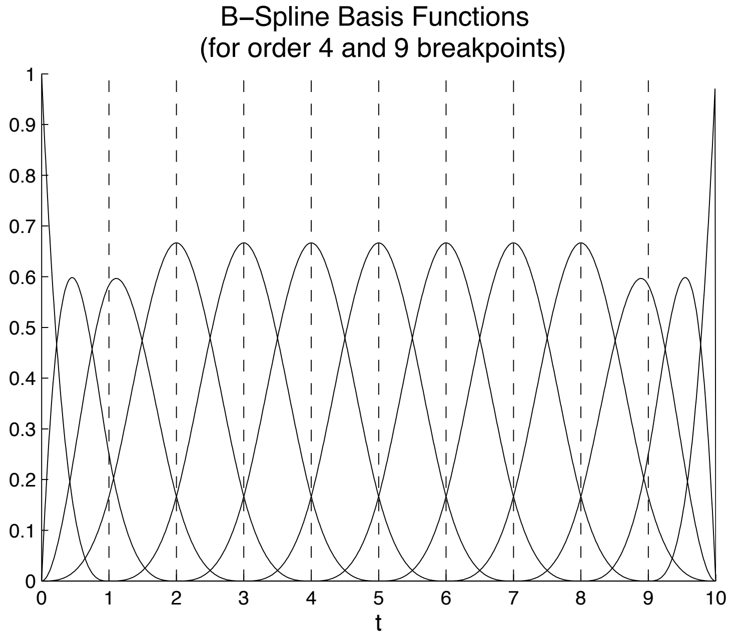


Figure 1.17: As an example we show here 13 B-spline basis functions for an order 4 spline with 9 breakpoints. The breakpoints are displayed as vertical dashed lines. These basis functions can be used in a linear combination to approximate any function, given the values at the breakpoints.

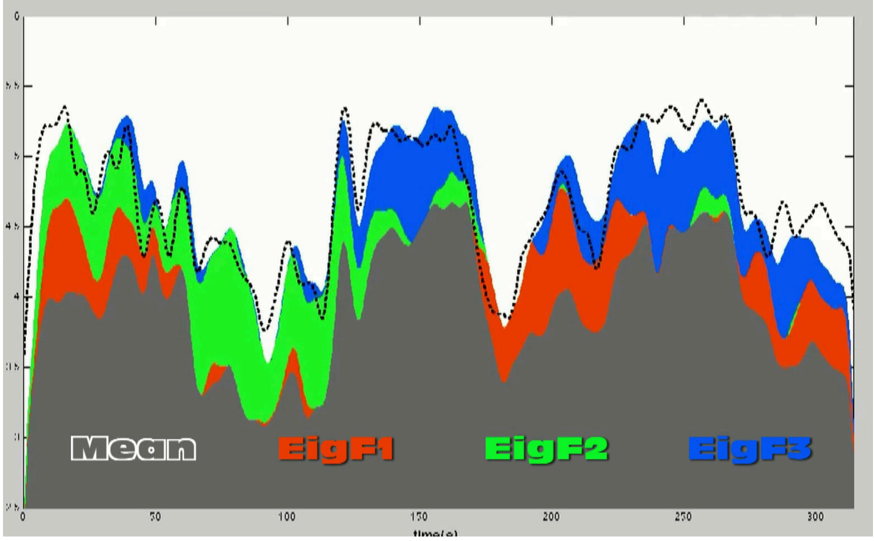


Figure 1.18: Example of decomposition in eigenfunctions. An individual signal, here indicated by the dotted line, can be decomposed as a linear combination of an average signal (gray area on plot) and a number of eigenfunctions. The average signal and the eigenfunctions are calculated by FPCA from the complete set of signals. Here we show a decomposition with three eigenfunctions (orange, green, and blue areas) covering over 70% of the variance of the set of signals.
Note that this figure is a symbolic representation. For visualization purposes the negative contributions of the eigenfunctions are not shown and are absorbed in the positive contributions of other eigenfunctions.

Other Basis Systems

We briefly mention here a number of other basis systems (for a detailed overview we refer to [60]). There exists a wavelet basis system that combines the periodic capabilities of the Fourier basis system with the time-localized features of splines. Disadvantage is however that a wavelet basis system is less handy when derivatives are required.
Exponential bases ($e^{\lambda_1 t}$, $e^{\lambda_2 t}, ..., e^{\lambda_i t}, ...$) are useful because solutions of linear differential equations with constant coefficients are sums of exponentials.
In some cases there is no need to work with a sophisticated basis system. A simple basis system like a polygonal basis system, a step-function basis system or even a constant basis system can do the work. These simple basis systems can also be considered as special cases of a B-spline basis system.
Noteworthy to mention as another example of gesture decomposition is the so-called 'Periodicity Transforms' [61], which is a technique to decompose a data sequence into a sum of simple periodic sequences by projecting onto a set of periodic subspaces, leaving residuals whose periodicities have been removed.

Musical Relevance of decomposition

Musical signals show periodicity (think about beats per minute) and this periodicity appears also in musical gestures. A decomposition in a Fourier basis system therefore makes sense. Additionally, a decomposition in an empirical basis system is also easily motivated by the definition of expressiveness (cfr section 1.2.3.1). Expressiveness is defined as a variation around a reference. Empirical basis decomposition takes the group average as reference and the variation is then explained by the eigenfunctions.
Decomposition (except for Fourier basis system) is not so popular as segmentation and less research work is done in this area. We refer here to work of Vines [62] for an illustration of the empirical basis decomposition and Leman and Naveda [39] for using the periodicity transformation technique.

1.3.2.3 Handling Temporal Variability

Musical gestures experience variation in both amplitude and time (phase) and confounding these two sources of variation may lead to problems. Handling this is our third challenge. Ramsay [63] illustrates this problem with an example taken from another discipline, namely the acceleration in children's height. Results show that an estimate of the average acceleration does not resemble any of the observed curves. See Fig. 1.19 for an equivalent illustration of this problem.

A solution lays in what some call registration of the data, involving transformations of the argument t rather than the amplitude x(t). An

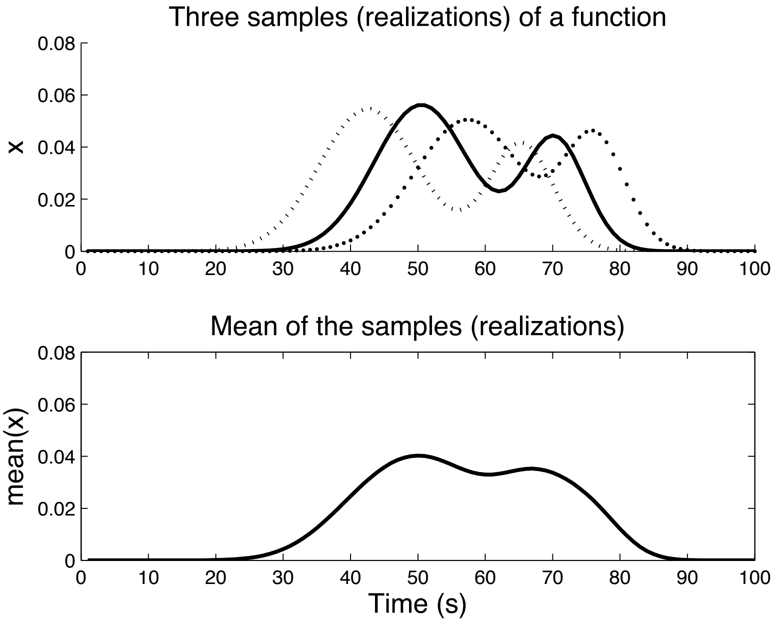


Figure 1.19: Illustration of temporal variability. In the top figure we see three realizations of a function that do very much resemble each other except that the peaks occur at different time stamps. When averaging (bottom plot) we see that much of the information (like alternation of a large and a small peak) is lost. This illustrates the need for handling temporal variability prior to analysis. See also Fig. 1.20)

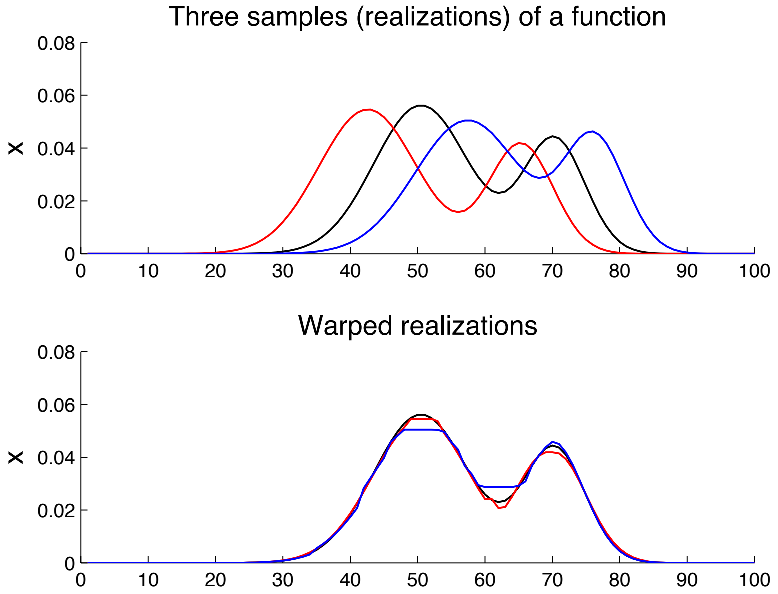


Figure 1.20: Illustration of handling temporal variability by dynamic time warping (DTW). In the top plot we see the same three realizations from Fig. 1.19. The bottom plot shows these realizations again but now time warped with reference to the first one. We see that much of the information (like alternation of a large and a small peak) is now preserved. This illustrates the need for handling temporal variability prior to analysis.

easy and straightforward method to achieve this is by making use of landmarks. In this case two musical gestures are aligned using the extrema of the amplitudes of their signals. The intervals between the extrema are then interpolated (i.e. linearly stretched or shrinked in time).

A more advanced method makes use of Dynamic Time Warping (DTW) (Fig. 1.20). DTW is not limited to landmarks but uses every value of the signal. In general, DTW is a method that allows an algorithm to find an optimal match between two given sequences (e.g. musical gestures) with certain restrictions. The sequences are “warped” non-linearly in the time dimension by maximizing a measure of similarity independent of certain non-linear variations in the time dimension. Further analyses can take the warped signals as input.

Musical Relevance of handling temporal variability

Temporal variability is not something that we find “by accident” in music and musical gestures. No, think about Tempo Rubato (or Italian for stolen time) [64] that refers to the rhythmic freedom by slightly speeding up or slowing down the tempo of a musical piece. This is a form of expressiveness (deviation from the regular) sometimes applied by a soloist or a conductor.

1.3.2.4 Dynamical Systems

We believe that a dynamical system approach for gestural data modeling is in general part of best practices. A human subject, its environment, and their interaction, is best modeled as a dynamical system that is determined by a set of quantitative variables changing simultaneously and interdependently over time. A dynamical system is a model describing the temporal evolution of a system. The evolution starts from an initial state and is usually formulated as a differential equation in the continuous domain (1.2) or as a difference equation in the discrete domain (1.3). A dynamical system relates present values with first and higher order derivatives. Derivatives have to be calculated in advance, before actual analysis and that is the reason why this section falls under data preprocessing.

$$\frac{\delta^n x(t)}{\delta t} = f(t, x(t), \frac{\delta x(t)}{\delta t}, \frac{\delta^2 x(t)}{\delta t}, .., \frac{\delta^{n-1} x(t)}{\delta t}) \tag{1.2}$$

$$x_{n+1} = f(x_n, x_{n-1}, ..., x_0) \tag{1.3}$$

In our research we propose three proven methods of bringing the dynamical systems approach into practice :

1. *Augmented Feature Space Construction*

The first proven method is the construction of an augmented feature space. This method combines positional data and its derivatives (e.g. velocity, acceleration,..) in a what is then called an augmented feature space. This solution is congruent with a multivariate data approach, where we observe and analyze the outcome of more than one variable. The augmented feature space can then be used as input for many multivariate analyses (e.g. Principal Components Analysis (PCA), Factor Analysis (FA), Canonical Correlation Analysis (CCA), Cluster Analysis). Take caution however as augmenting the feature space means that it becomes now a collection of variables with very different units (m, m/s, m/s^2...). Therefore it is recommended to perform a normalization step prior to analysis. Normalization puts all features on an even footing meaning that relative changes in variables can be compared even if their original units are different.

2. Sequential Data

Handling data as sequential data is a second proven method to bring a dynamical systems approach in practice. This is conform equation (1.3) where we notice that the value of a sample at time stamp t_n depends on the values of samples before time stamp t_n. What this says is that the values at different time stamps are not independent but dependent. In other words an independent and identically distributed (i.i.d.) model does not hold for our type of data. We need models for sequential data instead (for example Markov models, Recurrent Neural Networks (RNNs), ...).

3. Functional Data

All movement data that we collected is data sampled at discrete time intervals and thus in essence sequential data. How can we now convert sequential data into functional data? The answer is *smoothness*. Sequential data samples are considered as 'functional' if they reflect the smooth curves that we assume are at the origin [60]. Just like sequential data imply dependencies between adjacent samples, smooth basis functions also imply such dependencies. In fact, smooth functions can be interpreted as solutions of a differential equation.

Musical Relevance of a dynamical systems approach

The approach leads to novel research methodologies for movement analysis in relation to music. It allows the description of model parameters that capture expression in terms of basic movement patterns that are extracted from real movements, rather than in terms of single values that capture a particular feature of a particular movement segment (e.g. [57] [38] [65]). Such an approach has recently been explored by Leman and Navada [39], who use periodicity analysis to capture spatiotemporal representations of gestures, and by Camariaux et al. [66], who use gesture templates.
To our knowledge functional data analysis is not so often applied in musical research. Exceptions are for example work from Toiviainen [67] and Almansa [68].

1.4 Methods and Goals

The essence is that we live in a complex world. To deal with this complexity we can collect data, learn a model out of it and use this model as a representation for the complex reality. The model is then used as a tool for gaining insight and/or prediction.

This vision is perfectly in line with the definition of *Machine Learning* as given by Tom Mitchell [69] : “A computer program is said to learn from experience E with respect to some task T and performance measure P, if its performance at task T, as measured by P, improves with experience E.” We just have to replace a computer program by a model, task T by gaining insight and/or prediction and experience E by learning.

The above definition of Machine Learning dominates this section. In the reminder, we give a summary of Machine Learning methods, followed by a subsection that focuses on the task T and eventually we explain how to express the performance measure P.

1.4.1 Machine learning methods

Machine Learning tools can be organized into the following taxonomy:

Supervised Learning

For Supervised Learning the training data consists of a set of input data and the corresponding set of output data. The task of Supervised Learning is to construct an algorithm or model that links input data with output data. For continuous output data this is called *regression*, for discrete output data the task is called *classification.* The major focus is on generalizing from the training data to new, unseen data, in other words on predicting. For completeness we have to state that nowadays in addition to black box models (like Reservoir Computing, Support Vector Machines (SVM) or Neural Networks) other models (e.g. Decision Trees, Bayesian Models, . . .) are used to allow the integration of process knowledge. In these cases Supervised Learning combines prediction with gaining insight (insight in process knowledge).

Unsupervised Learning

For Unsupervised Learning the training data is just one entire set of data. We do not distinguish between output data and input data. The task of Unsupervised Learning is to discover patterns and structure in a data set. The most commonly used techniques are clustering (e.g. K-means clustering, Gaussian Mixture Models (GMM’s)) and dimension reduction techniques (e.g. Principal Component Analysis (PCA), Multi Dimensional Scaling (MDS)). The major focus is on gaining insight in process knowledge.

Reinforcement Learning

Reinforcement Learning is about taking actions (output) in an environment that maximizes some notion of future cumulative reward. Because of the presence of actions (output) Reinforcement Learning is much closer to Supervised than Unsupervised Learning. It differs from standard Supervised Learning in the sense that correct input/output pairs are never presented. Further, there is a focus on on-line performance, which implies a trade-off between discovery of new actions and execution of known actions.

1.4.2 Tasks - Goals

Although the subdivision of machine learning methods determines the basic tasks (prediction, gaining insight, maximizing future reward), it is worthwhile to consider an extended set of tasks as this will determine what the most appropriate model(s)/method(s) are.

Prediction

1. *Classification* is a technique to identify to which class (category) a new data sample (item) belongs.

   The set of classes can be anything that can be enumerated and that can range from an easy to understand class set (dog, cat, horse, cow, ...) to complex sets where each member is a model on itself. In the latter case classification is synonym for model selection.

2. *Regression* is a technique to estimate the relationship between a (continuous) dependent variable and one or more independent variables.

Representation

Representation (Modeling) is related to gaining insight by making a low dimensional (compact) representation of the reality. The low dimensionality helps understanding.

Verification

Verification checks how well a set of observations fits a trained model. This information can be used in a various number of ways. If a set of observations fits one model better compared to another model, we can use this information for model selection.
Another use of verification is for *data validation.* This will tell how well a model generalizes to unseen data. The base problem is that from a small training set only simple models can be learned. In case a model is too complex the model will also describe the random error or noise instead of the underlying truth alone. This phenomenon is known as overfitting. To avoid overfitting data validation offers a solution. Data validation sets a part of the data aside for testing. This is the test data set. The remainder is the learning data set. Both data sets should be large enough: the learning data set to allow complex models and the test data to check the validity of the model: The test data is used to indicate statistically significant model improvement for minor model changes. In practice data sets are often too small for having a separate test data set. For those cases an alternative solution is k-fold cross validation.
k-fold cross validation partitions a data set in k parts. $k-1$ parts are used for learning and 1 part is used for testing (validating). This process is repeated k times, with each of the k parts acting exactly once as a test data set.

Generation

Generation is the process of generating a set of new observations in line with a model or a training set. Generation usually involves some restrictions like for example starting from an initial condition and quite often a process of randomization is involved.

Rewarding

Optimizing future reward is a key element for Reinforcement Learning. Each time the learner performs an action, he receives feedback (reward) about the appropriateness of his response. For example think about a baby learning to walk by stumbling and falling. This makes Reinforcement Learning, although not used in this thesis, of interest for musical research and more specific for educational games [70]. For example the game-play of an educational musical game like the Music Paint Machine [71] could benefit from it.

1.4.3 Performance measures

- A simple performance measure for the classification-task is to calculate the number of misclassified samples divided by the total number of classified samples. This number can directly be used to compare classification algorithms but in some cases it is better to assign different costs to different types of misclassification. Take for example the binary classification case where the cost associated with a false positive or a false negative can be different.

- For regression the performance measure is usually calculated as the mean square of the residuals, with the residuals defined as the difference between the predicted and the observed value.

- For the representation task the same performance measures like for classification and regression can be used. Instead of predicted values one now uses the values as reconstructed from the low dimensional model.

- For the verification task the goodness of fit can be calculated by many criteria. These can also be the same criteria like for classification and regression. However, if the trained model is a probabilistic model, then in addition the goodness of fit can also be calculated by a measure called the production probability. The production probability allows to compare sets of observations in terms of their fit to a model and additionally it can be used as discriminator for model selection (as mentioned under classification).

- There exist no general performance measures for a generation task. It requires dedicated algorithms and usually also human intervention to judge its appropriateness.

1.5 Research methods for Musical Gestures

At the basis of this thesis are the analyses of four experiments, all set-up serving a specific goal. These analyses are discussed in detail from Chapter 2 to Chapter 5. Here, in this section we explain and motivate why we used some particular Machine Learning methods for the specific tasks (goals) we wanted to accomplish.

1.5.1 Towards E-Motion Based Music Retrieval

Here, the purpose of the research is to extract valence and arousal information from a musical gesture and to use this to retrieve music from a valence and arousal annotated library. The goal is to develop a model that can calculate from gestural data the values for valence and arousal (See Fig. 1.21).

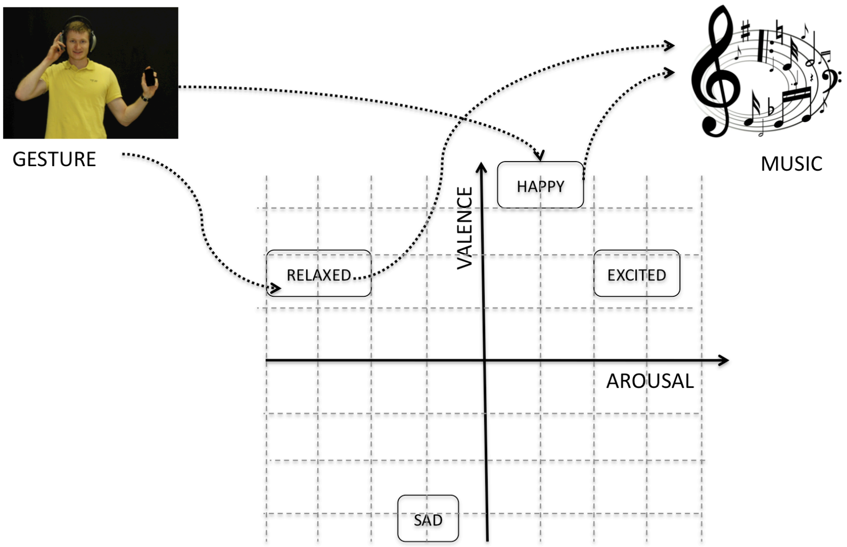


Figure 1.21: E-motion based music retrieval. The purpose of this research is to (1) extract arousal and valence information from a musical gesture and (2) to retrieve music from an arousal/valence annotated library using the extracted arousal and valence information.

Verification of the model was the next important task and this was accomplished by setting data aside for the sole purpose of data validation. The performance measure used for data validation was the mean square of the residuals.

The research was set-up as a proof of concept. The step towards a full working application would have required a more meticulous handling of variability and constraints and this would have lead us away from the fundamental research path. Handling variability means dealing with different brands and types of sensing devices, dealing with calibration issues and dealing with human variability. Constraints can come from lack of some sensing devices, like lack of gyroscopes making it impossible to measure orientation. The number of subjects (32) participating in this experiment was also rather low, making a high dimensional model with lots of parameters impossible.

The eventual choice was on a simple linear regression model for predicting valence and arousal figures using just a small number of explaining variables (dependent variables). The explaining variables needed to have a direct physical meaning (like e.g. speed) to facilitate the portability to another setting.

Major strength of our approach is that the valence/arousal plane acts as a Marshalling panel between gestures and the music library (See Fig. 1.21). This implementation allows to “solve” faulty maps not by making changes to the model but by overriding annotations in the music library. In a real-life application a music library will be annotated by an algorithm calculating a valence/arousal value for every entry. We refer to these values as being the default values. The default values will not have a 100% match for every individual user. Our set-up can use the valence/arousal values calculated from an individual’s gestures to overwrite these default values. This makes it straightforward to tune the application to the individual’s needs.

Disadvantage of our approach is that valence and arousal are usually measured on a continuous scale and in our case it is a 1-5 Likert scale. The validity of handling the Likert scale as a continuous scale is in this context debatable. For a discussion consult for example [72].

The chosen model is a simple linear regression model covering the whole valence/arousal domain. An improvement could come from dividing the valence/arousal plane in sub-areas with a dedicated model per sub-area.

1.5.2 Expressive Body Movement Responses to Music are Coherent, Consistent, and Low Dimensional

The main task for this research topic is to represent music evoked body movement data in a low dimensional datamodel (Fig. 1.22) and to evaluate if there was a group difference between musically trained and musically untrained subjects. We consider the collected data as functional data and the *variables* can therefore be represented as $f(1), f(2), ...f(t), ...f(N)$, with N the number of samples in the time domain. We have in other words as many variables as there are time stamps. For a 5 minute signal sampled at 100Hz this means 30.000 variables!

Working with all these variables means that we can use standard multivariate techniques like Principal Component Analysis (PCA), Analysis of Variance (ANOVA), ... That is a major asset as these methods are beneficial to our envisioned tasks. PCA handles for example dimension reduction with the average as a reference and this is in line with our definition of expressiveness. ANOVA allows to compare several groups of observations (in our case two groups, namely the musically trained and musically untrained group) with possibly a different mean for each group.

However there are some problems with this multivariate approach: (1) The large number of variables makes computation on computers infeasible. (2) It treats the function values $f(t)$ and $f(t+1)$ as independent variables and we know that this is not the case. The value of $f(t+1)$ does depend

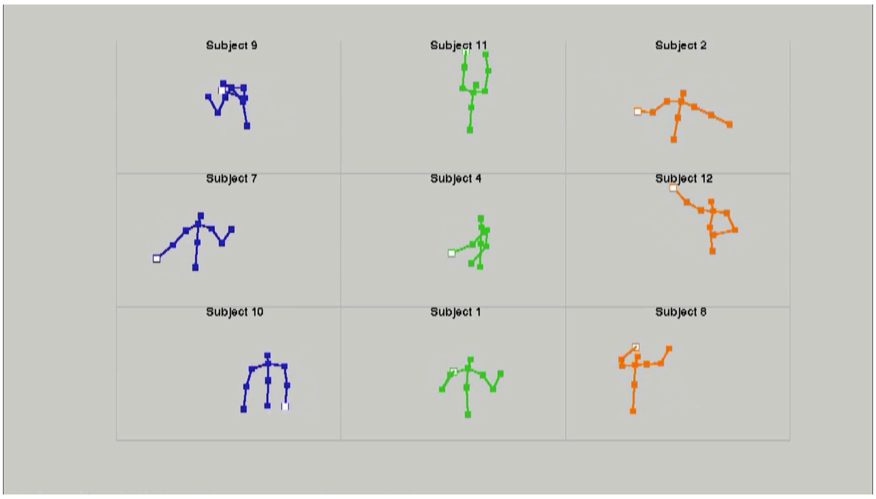


Figure 1.22: Low dimensional representation of expressiveness in music-driven spontaneous dance movements. The purpose of this research is to generate a low dimension model to describe a group of dancers. In the shown figure the model allows to cluster subjects in three groups (low dimensional) based upon their level of expressiveness: from low expressiveness(left) to high expressiveness(right).

on $f(t)$.

To solve these problems we use the technique of Functional Data Analysis (FDA) as worked out by Ramsay and Silverman [60]. This technique decomposes a signal into a number of basis functions. In our case we use B-splines as basis functions because our signal (a low pass filtered speed signal) was not periodic. Ramsay and Silverman adapted the standard multivariate algorithms for PCA and ANOVA so that these methods work for decomposed signals. This makes the whole implementation feasible on an average computer system. They labeled these new algorithms “functional PCA” (FPCA) respectively “functional ANOVA” (fANOVA). Because of the decomposition into continuous basis systems, the values f(t) and f(t+1) are now also dependent as the bases are continuous.

Ramsay and Silverman enhanced the decomposition even further by penalizing decomposition solutions that are further away from an assumed dynamical system. Our hypothesis is that a human subject, its environment, and their interaction, is best modeled as a dynamic system that is determined by a set of quantitative variables changing simultaneously and interdependently over time. This is mathematically translated into a differential equation. Common practice is to describe a human gesture as smooth by having its second derivative as small as possible. In this case the second derivative is used as penalty term for decomposition.

1.5.3 Beating-Time Gestures: Imitation Learning for Humanoid Robots

The work here builds further on findings from Maes et al. [73]. One of the problems they faced was extracting a reference beating-time gesture out of a series of performances. Underlying difficulty was the temporal variation in a beating-time gesture: Sometimes a gesture arrived too fast at the first beat but this got corrected by slowing down in the subsequent inter-beat interval.

Our research provides a solution that handles this type of temporal variation as well as positional variation. The procedure creates a so-called generalized gesture, which is “optimal” with respect to the temporal and spatial characteristics of a set of performances (Fig. 1.23). In this sense we can refer to this task as a generation task. Additionally, we want this generalized gesture to be suitable for use with humanoid robots. This means that the gesture should be smooth and easily adaptable to all kinds of music.

Our solution is inspired by the Programming by Demonstration (PbD)-solution from [74]. It handles spatial variation by cubic spline regression. This has as additional advantage that it deals well with periodic boundaries. A beating-time gesture is part of a continuously repeated sequence, and so we want the beginning and the end of the generalized gesture to coincide. Cubic spline regression is often done with a set of equidistant knots (uniform splines). Then, extrema in the trajectory can or can not coincide with the knots. If they do not coincide, the consequence is that the extrema of the trajectory are flattened out resulting in a compressed shape. Because beating-time gestures use the extrema to convey beat information, we do not go that path and we choose for non-uniform splines instead.

We handle the temporal variation by adding a dynamical time warping (DTW) step. This is achieved by warping all demonstrations non-linearly in the time dimension to a reference signal. Here, the challenge comes from the calculation of a reference signal.

We propose to handle the remaining issues by fitting an HMM. The average timestamps of where the HMM state transitions happen are then used (i) for setting the non-equidistant knots for cubic spline regression and (ii) for the creation of a reference signal for DTW.

As we prefer to keep the set of demonstrations low we need a simple model, in our case a HMM with few parameters. The number of HMM states and the initial values for Baum-Welch training of the HMM parameters follow from a Dirichlet Process Gaussian Mixture Model (DPGMM) that

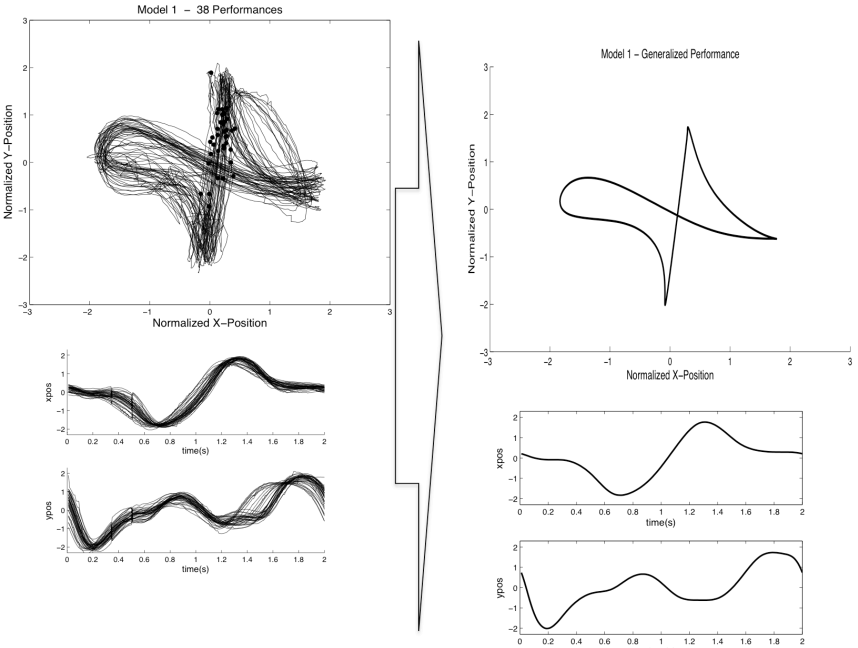


Figure 1.23: Beating-time Gestures. The task is to produce a single generalized trajectory using all information (temporal and spatial) from a set of performances. The spatial variation is shown at the left top. The temporal variation is shown at the bottom left. The generalized trajectory is shown in a spatial dimension (top right) and in a temporal dimension (bottom right).

we fit to the data. DPGMM is a Bayesian method using a Dirichlet process as prior. The prior acts as a regularizer preventing overfitting and resulting in models that usually generalize better. This is an asset, as in our case we have few data and model fitting with few data is prone to overfitting. For more information on DPGMM we refer to existing literature (e.g. Teh [75] and El-Arini [76]) or appendix A.

A critique on the above solution is that the gesture is not human anymore. This is correct, a human solution would revert to selecting one performance out of a set. This would make that the generalized trajectory is based upon the information of one single performance and it would not take into account the temporal and spatial information that exists in all the performances.

The calculated generalized trajectory lasts exactly one measure and the beat points (metronomic ticks) are known for this trajectory. The beat points can be used to adapt a generalized trajectory to any music provided that the beat timestamps of the musical piece are identified.

1.5.4 The Surprising Character of Music. A search for sparsity in music evoked body movements

The main task for this research topic is “representation”. We search for a low dimensional model that helps in finding sparsity in movement data (Fig.1.24). Sparsity is synonym for low density areas. A classical way to model a density is by applying a mixture model. This is a model which comprises a number of component functions (clusters). The component functions are then combined to model the density as a multimodal density. As component functions we use Gaussians for modeling spatial data and Multinomial distributions for modeling directional data. Spatial data is represented in a 3 dimensional space, hence the use of 3D Gaussians. Directional data is represented as a directional mix over a time interval, hence the use of Multinomial distributions.

For this research we use also the Dirichlet Process Mixture Models which were introduced in the previous section: For the positional data we use a Dirichlet Process Gaussian Mixture model (DPGMM) and for the directional data a Dirichlet Process Multinomial Mixture Model (DPMMM).

1.6 Results

So far we have provided the background, the concepts and the methods. In this section we link the results of our empirical research to the previously defined main concepts discussed in section 1.2. Additionally, we show that

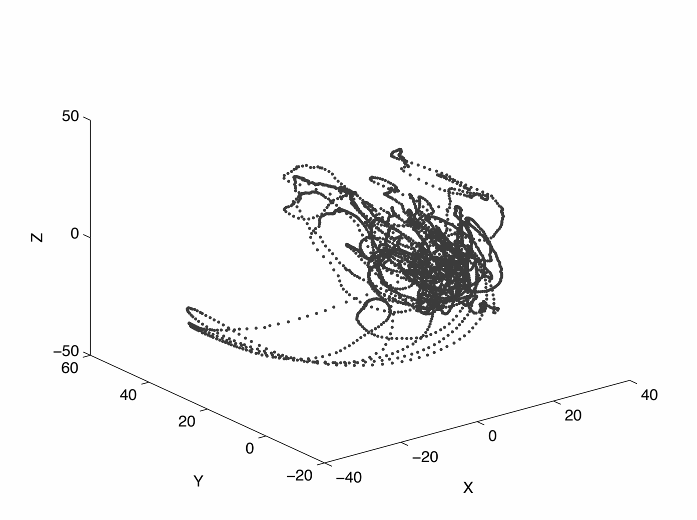


Figure 1.24: Sparsity in Spontaneous Dance Movement. The task is to model a density and to identify low density areas. Low density areas can be an indicator for "surprise". The density shown in the picture here represents the spatial density of the right hand movement. Visual inspection reveals an abnormal movement (in a low density area) presented at the bottom left of the figure.

our results are in support of the dynamic model for the action-perception coupling system (Figure 1.2) presented in section 1.1. For readers requiring more information, we refer to the subsequent Chapters for in-depth results and detailed information.

1.6.1 Gestural Descriptors

As specified in section 1.1, gestures play a core role in the encoding and decoding of musical expressiveness [6]. Quite often, movement in response to music is seen as a gestural expression of a particular emotion that is assumed to be imitated by the music [77] [78] [79] [30].

Concentration.

The way we defined concentration (as locations and/or directions that are more and longer frequented than others) makes that it correlates immediately with spatial and directional density. We investigated this by means of a spontaneous dance experiment on music of Johannes Brahms' First Piano Concerto Opus 15 in D minor (Chapter 5). The analysis of the experiment showed that these densities can best be modeled by a mixture model, resulting in multiple areas of high concentration (Fig. 1.25).

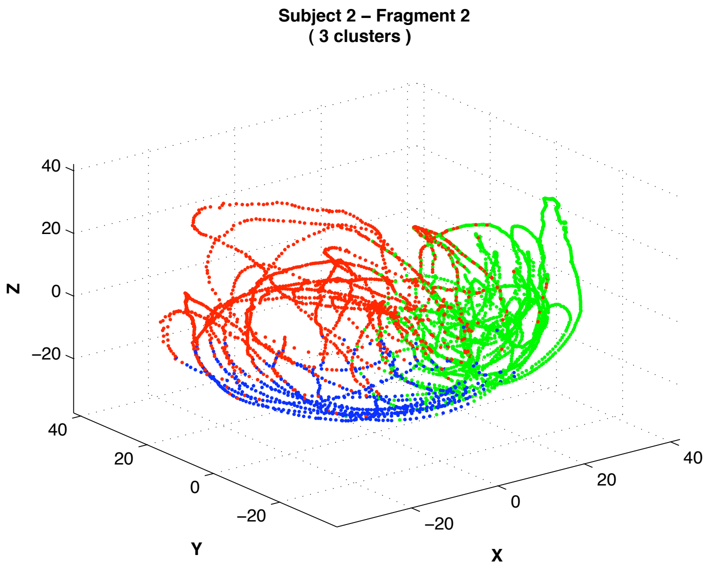


Figure 1.25: The spatial representation of spontaneous right hand movement on music of Johannes Brahms reveals three clusters of high density areas for the first lyric fragment. For more details, we refer to Chapter 5.

Volume.

The descriptor volume was studied during the same experiment. Volume was found to be a descriptor for the complexity of the dance movement. A small volume stood for mainly repetitive movement, large volumes stood for more variety in movement. This linked also to the musical style intervals present inside the fragment. Higher volumes matched the heroic style intervals and lower volumes matched the lyric style intervals. Additionally, there was a difference between the musically trained group (labeled here as the MTr-group) and the musically untrained group (MunTr-group). Within a style fragment the MTr-group reached higher volumes than the MunTr-group (Fig. 1.26). This correlated with more expressiveness of the dominant hand for the MTr-group.

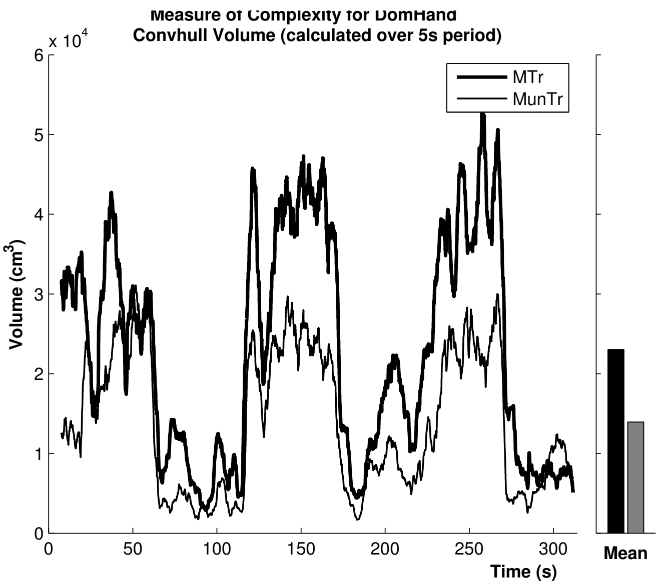


Figure 1.26: Volume of dominant hand movement averaged for two groups of dancers : in dark the musically trained group (MTr) and in light gray the musically untrained group (MunTr). The volume is expressed over time showing a correspondence between volume and expressive style intervals in the music. Here, we have three time intervals where the music is labeled as heroic and three time intervals labeled as lyric. The higher volumes match with the heroic style intervals, the lower volumes with the lyric style intervals.The group averages are also different, the musically trained group reaching higher volumes.

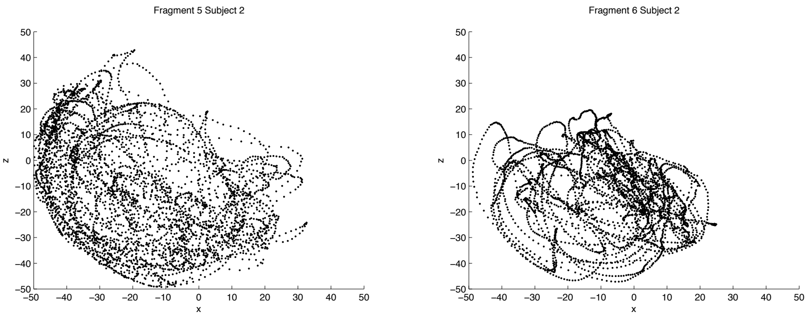


Figure 1.27: Elevation of right hand for the same dancer in two musical fragments, respectively a heroic (left) and a lyric style (right) fragment. The origin is placed at the shoulder. The elevation (z-axis) is found significantly higher in the heroic style fragment compared to the lyric style fragment. See Chapter 5 for more details.

Elevation.

Elevation was also found to be a differentiator between the heroic and lyric musical style intervals in that same experiment. The maximum in absolute terms or the average maximum of the high density areas (clusters) was found to be higher in heroic style intervals compared to lyric style intervals (Fig. 1.27).

Dimensionality.

Dimensionality was found to be a good predictor for the emotional content of a musical fragment. Our research on emotion based music retrieval (See Chapter 2 or [29]) revealed for example that speed (for arousal) and jerk (for valence) were important. These were however no novel descriptors. The properties of the descriptor dimensionality were less anticipated. Arousal and valence did both correlate with dimensionality. High dimensionality (more 3D movement) correlated with high arousal. Low dimensionality (movement along a line) correlated well with low valence.

We also gained new insight in dimensionality from our research on a spontaneous dance experiment on music of Johannes Brahms (Chapter 5). The positional analysis revealed that all movement happened on the surface of a three dimensional ellipsoid, centered at the body (Fig. 1.28). This is interesting because the surface of an ellipsoid is a manifold of dimension two. This low dimensionality partly finds its origin in physiological restrictions of the body, being a chain of pendulums but here it is also due to the lack of spoke-like (punching) movement. In modern dance styles as for example hip-hop, we do encounter these spoke-like movements. This makes

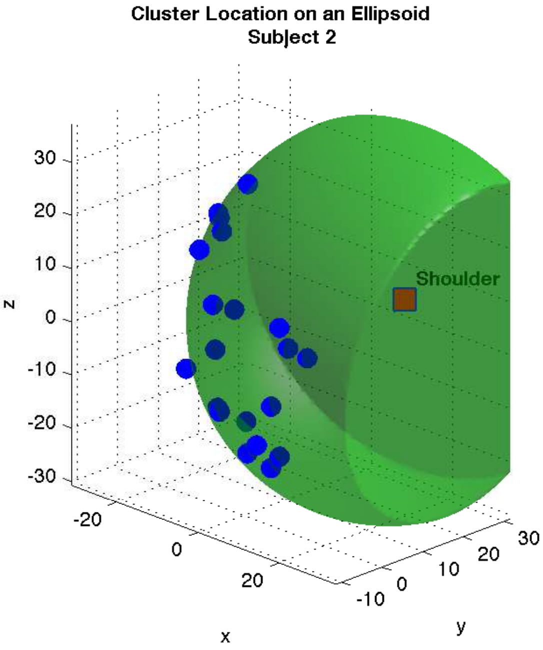


Figure 1.28: Example of dimensionality.

Visual inspection learns that the centers of high density areas of right hand dance movement fit on the surface of an ellipsoid. These centers are represented by blue dots. For more details on how these high density areas were found we refer to Chapter 5. The shoulder was added as reference point. The dimensionality of this movement is "two" as two coordinates suffice to identify a spot on the ellipsoid. Size and shape of the ellipsoid are also of help for the descriptor proximity.

dimensionality a key element in comparing the musical intentionality of different musical styles.

Proximity.

The descriptor proximity was related to the ellipsoid mentioned in the previous paragraph under "dimensionality". The size of the semi-axes give an indication of how close movement was to the body and changes in the size of these axes could link to concepts like 'opening' and 'closing'. As can be learned from Chapter 5 the major problem is that fitting an ellipsoid is not evident given the fact that all movement is concentrated on a limited area of the surface of the ellipsoid. Future research should investigate other approaches to deal with this concept.

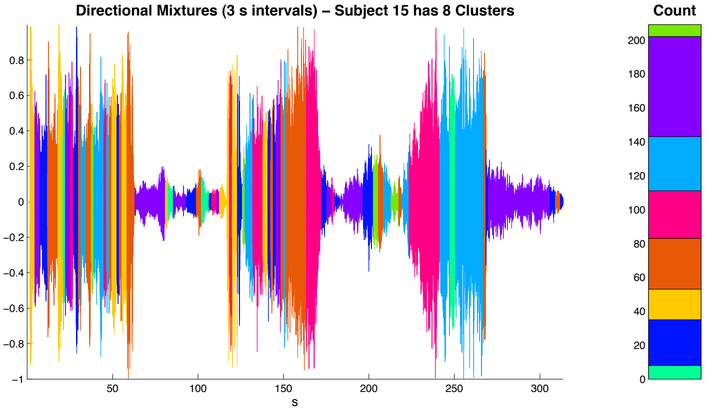


Figure 1.29: This figure visualizes the directional mixture of right hand dance movement by means of the musical wav-file. The mixture is calculated over a three second time interval. Clustering based upon these directional mixtures divided the dance movement for this subject in 8 clusters. The lyric intervals (recognizable by their small amplitude) are dominated by a single cluster, the heroic intervals (large amplitude) do not show this behavior. See Chapter 5.

Direction.

In our research on emotion based music retrieval (See Chapter 2 or [29]) we evaluated observers' findings describing the arm movement in their own wordings. A discriminating indicator that was shared amongst all of the observers was whether there was more horizontal movement compared to vertical movement. This gave us the idea to define a directional mix over a fixed time interval as gestural descriptor. We investigated this mix in a music evoked body movement experiment using music of Johannes Brahms. It was of particular interest to see how this mix evolved over the musical fragment (Fig. 1.29). The analysis revealed that for all lyrical style fragments the directional mixture was dominated by one and the same cluster. This was not the case for the heroic style fragments. The link with the surprising character of the music is there but needs more research. We refer to Chapter 5 for more detailed information.

1.6.2 Key Points and Goal Points

The concept of key points and goal points was at the basis of our research to produce a generalized conducting gesture from a set of demonstrations (Chapter 4). This research is of major interest because it reveals a problem immanent to all musical gesture research. The problem is that there exist two sources of variability, namely spatial and temporal variability. Spatial variability can not be handled or studied without removing temporal variability. The results of our research on this topic are summarized below. For more in depth information we refer to Chapter 4.

Key Points.

To remove temporal variability, we first reduced a performance of a conducting gesture to a set of key points: spatial landmarks that are common to all performances. This was done by fitting a continuous HMM. The key points were set at the internal state transitions of the fitted HMM. The HMM was fitted on an augmented feature space using beside the positional variables also the velocity variables (Fig. 1.30).

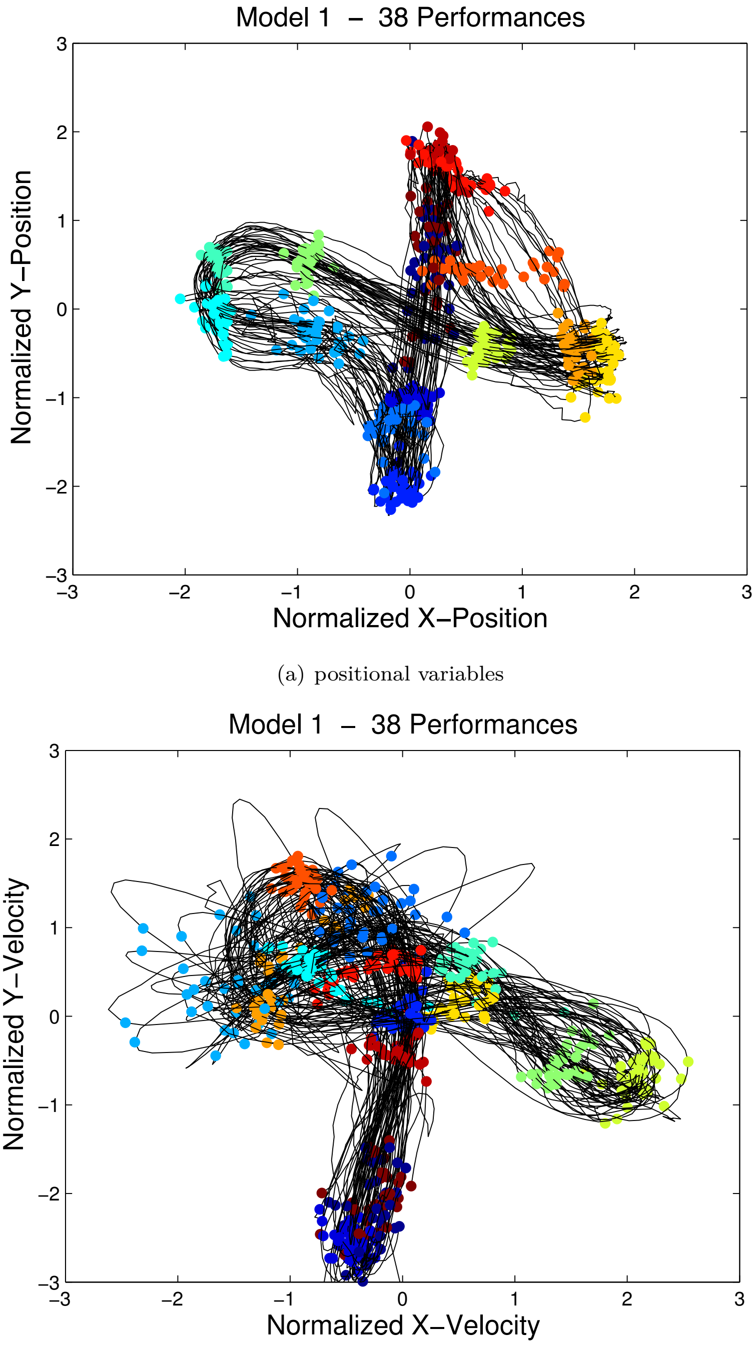
The concept of key points reduces a gesture to a small subset of important (key) samples. This facilitates for example the comparisons of gestures. Our application however used this concept to remove temporal variance. This was achieved by temporally aligning the key points of a gesture to a reference gesture using Dynamic Time Warping (DTW).

Goal Points.

The concept of goal points was of high importance for the synthesis part in our work on conducting gestures (Chapter 4). In the synthesis part we adapted a generalized trajectory to match a real musical fragment using goal points as anchor points (Fig. 1.31). The goal points were mapped to the beat time stamps. For the generalized trajectory these time stamps were at the metronome ticks. For a musical fragment these were the time stamps at which the beats occurred. An initial solution for the synthesis problem is to time stretch the generalized trajectory so that the goal points (originally at the metronome ticks) match now the beat time stamps of the musical fragment. More advanced solutions would use probabilistic models (e.g. semi-Markov models) to achieve this.

1.6.3 Descriptors for gestures in social context

In this section we present the analysis results from a spontaneous dance experiment on music of Johannes Brahms' First Piano Concerto Opus 15 in D minor. The data was analyzed by functional data analysis. The logarithm of a low pass-filtered speed signal was a sufficient and relevant marker for analysis and modeling. Considering this signal as a function was the only assumption. No a priori segmentation of movement signals was required. The analysis used decomposition with an empirical basis system instead of



]

(a) positional variables

(b) velocity variables

Figure 1.30: Trajectories of 38 performances of a conducting gesture reduced to 15 key points per trajectory. Key points coincide with the hidden state transitions of a fitted cHMM model. Figure uses the positional variables (a) and the velocity variables (b) to visualize the key points. (See also Chapter 4)

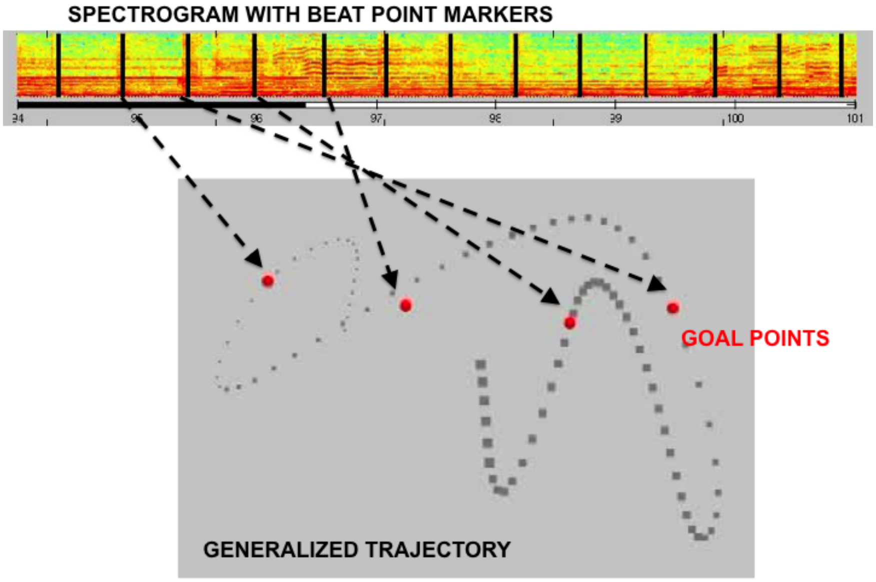


Figure 1.31: An idealized trajectory of a conducting gesture with its goal points (red dots). The goal points are mapped to the beat time stamps of a real musical fragment (spectrogram shown) in a 4/4 measure. An initial and easy solution towards solving the synthesis problem is to keep the shape (form) of the trajectory and to adjust the speed in function of the goal points. More advanced solutions make use of semi-Markov models. (See also Chapter 4).

segmentation. The results are here discussed with respect to the concepts of coherence/consistency and expressiveness. For additional detail we refer to Chapter 3.

Expressiveness.

We formerly defined the concept of expressiveness (Section 1.2.3.1) as "a variation around a reference". In our study we looked at *reference* from two different perspectives, namely (i) *reference* in relation to a number of musical characteristics such as pure periodic movement in the tempo of the music or as a musical amplitude, and (ii) *reference* as the average expressive movement of a group of performers. In other words, variation in body-movements is compared either with music, or either with a group average.

Functional Principal Component Analysis (FPCA) is a tool that is supportive to handling expressiveness. It allows to decompose a signal into a linear combination of a group average (the reference) and a number of eigenfunctions (the variation). In our case three eigenfunctions sufficed to cover 70% of the variance present in a group of subjects dancing spontaneously (not choreographed!) on music of Brahms. The decomposition for the right hand movement is presented in Fig. 1.32.

Coherence and Consistency.

The low dimensionality of the FPCA based model is explained by the presence of structure in the correlation matrix. This leads us to introducing the concepts of coherence and consistency. In section 1.2.3.2 we defined coherence as high correlation between subject performances at every two distinct time stamps in a continuous time interval. Consistency was standing for high correlations between remote coherent time intervals. Coherence and consistency are best visualized by means of a correlation diagram as in Fig. 1.33.

Using the concepts of coherence and consistency, we conclude that the musically untrained group focused on torso movement expressing the tempo of the music whereas the musically trained group focused on the dominant hand expressing additional structural elements as for example indicated by the musical amplitude. For more details we refer to Chapter 3.

1.6.4 Action-perception coupling system

Beside results for descriptors, we present here three results that are in support of our dynamic model for the action-perception coupling system (Fig. 1.2).

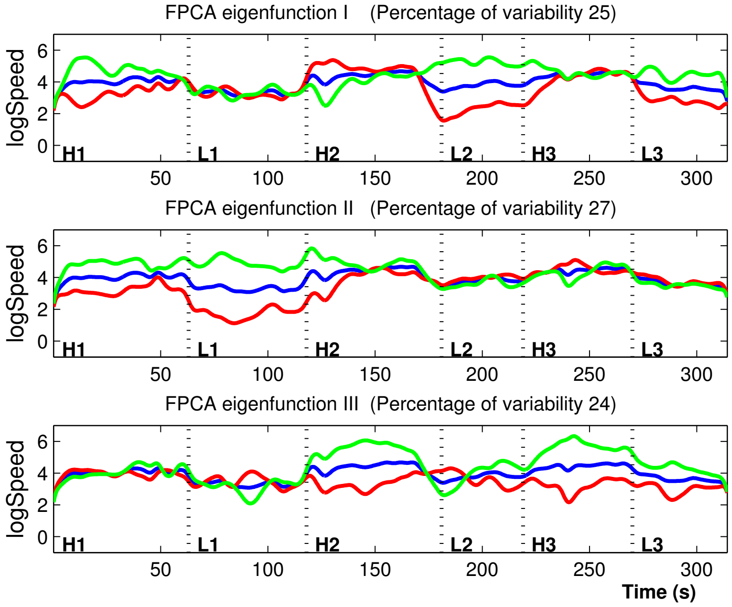


Figure 1.32: Dance movement (logspeed) of the dominant hand can be decomposed in an average (blue line on the three subplots) and a linear combination of eigenfunctions. The eigenfunctions are calculated from a functional principal component analysis and here three eigenfunctions are sufficient to cover over 70% of the variance. Every subplot displays one eigenfunctions twice, once with a positive offset (green) compared to the mean and once with a negative offset (red) to the mean. The drawn offset for an eigenfunction is proportional to the square root of its corresponding eigenvalue indicating its importance. The plots learn us for example that the first eigenfunction is important in interval H1 (the first heroic style interval) but not important in interval L1 (the first lyric style interval). Consequently, subjects with a high positive score for the first eigenfunction will be highly expressive in heroic style interval 1. For more details we refer to Chapter 3.

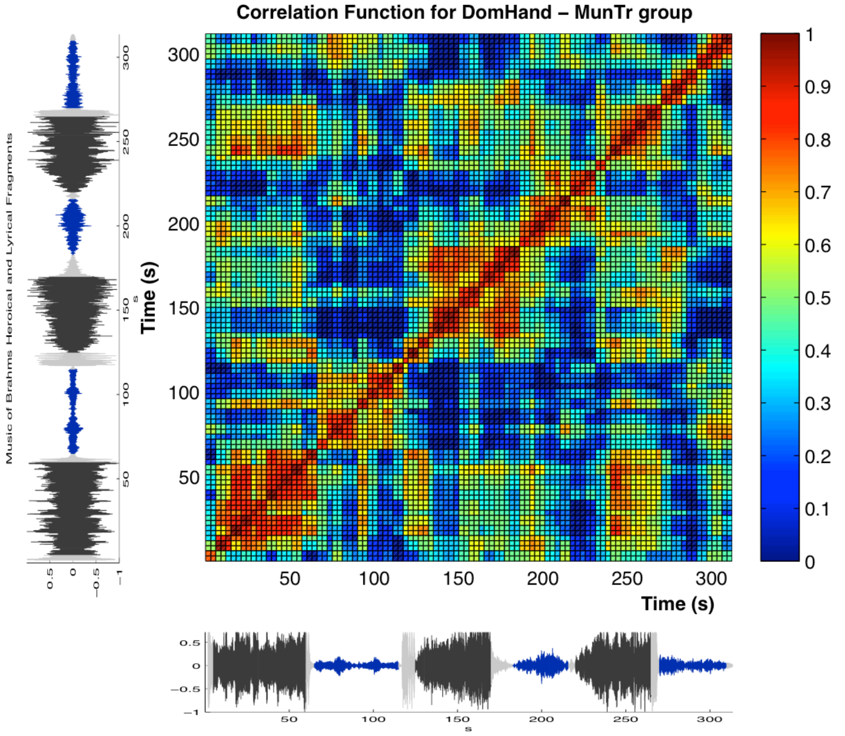


Figure 1.33: Coherence and consistency. The plot shown here is a correlation plot for the logspeed of the dominant hand and this for a group of musically untrained dancers. Warmer colors (red) indicate a high correlation between two timestamps (x-axis and y-axis). Cold colors (blue) indicate lack of correlation. We define then coherence as high correlation areas along the diagonal. These areas link back here to the different musical styles of the musical fragment. Musical style intervals alternate between 3 heroic and 3 lyric style intervals and that is visualized by means of the wav-file that is displayed below and to the right of the correlation diagram axes. Intervals with high amplitude stand here for heroic style and low amplitude intervals refer to lyric style intervals. Off-diagonal high correlation areas represent areas of consistency between remote time intervals. For example we identify high correlation (consistency) between the first and third heroic style interval. See Chapter 3.

The first result that confirms the existence of an action pattern repertoire comes from an experiment on music evoked bodily movement (see Chapter 5) where we found that the directional mix (amount of upward/downward/left/right) of movement was not arbitrary but could be clustered in a limited set (See Fig. 1.29). This endorses the existence of a distinct set of action patterns and one way to identify these action patterns is by looking at their directional mixture.

The second result comes from the same experiment where we noticed a link between the musical style interval and the directional mixture. The link is twofold: (i) Subjects change movement (in terms of directional mix) when musical style changes and (ii) some distant intervals having the same style (in our case lyric style intervals) reveal identical behavior. These phenomenons are best visualized by means of a *directogram*, being a density matrix (See Fig. 1.34) where the value of every cell (i, j) is calculated as follows: If for one subject the directional mixture for time i (row index) equals the directional mixture for time j (column index) the value of the cell is augmented by one. The end result is that every cell indicates how many subjects had “equal” behavior at the two time stamps given by the indices of the cell. The density matrix portrays areas of high values for the lyrical style intervals and low values elsewhere. This result back ups the existence of the outer feedback loop (the action-perception loop) in our model, where a change (comparison) in musical style (perception) causes a change in movement direction mixture (action).

A third result that is in line with our model comes from that same experiment where we identified common behavior within a group of musically untrained and within a group of musically trained dancers. The common behavior was explained by the terms coherence and consistency. The high degrees of coherence and consistency prove that a common underlying scheme must be at the basis. The difference between the two groups can be explained by the musically trained group having a larger and/or more complex repertoire that allows for a greater variety of action patterns.

1.7 Outline

This thesis is organized as follows. Chapters 2, 3, 4 and 5 are in essence extended versions of articles submitted to international peer reviewed journals:

- Chapter 2 Towards E-Motion Based Music Retrieval.
- Chapter 3 Expressive Body Movement Responses to Music are Coherent, Consistent, and Low Dimensional.

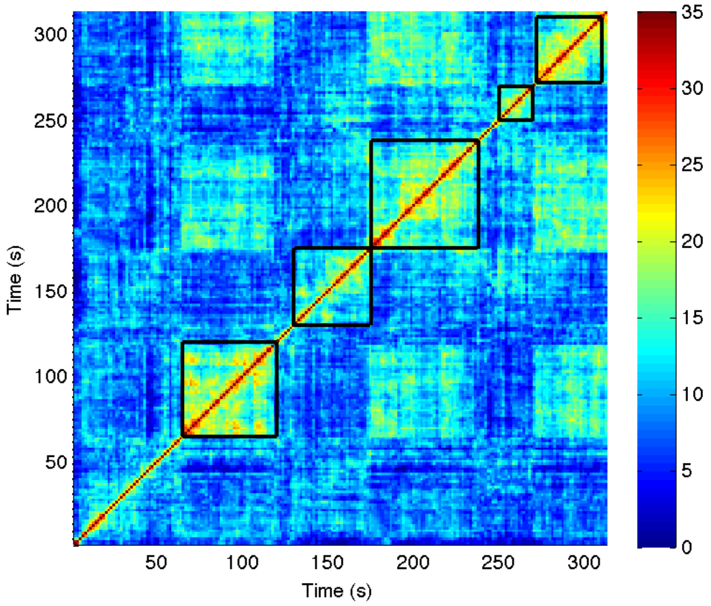


Figure 1.34: This directogram visualizes the directional mixtures (amount of upward/downward/left/right) of bodily movement for a group of subjects. The directogram is built by means of a density matrix where the value of every cell (i, j) is calculated as follows: If for one subject the directional mixture for time i (row index) equals the directional mixture for time j (column index) the value of the cell is augmented by one. The end result is that every cell indicates how many subjects had "equal" (equal in the sense of directional mix) behavior at the two time stamps given by the indices of the cell. The density matrix portrays now several areas of high density. High density areas correspond for example with the lyrical style intervals (squares with black edges) in the musical fragment. Also off-diagonal we notice high density areas revealing an identical directional mix in distant lyric style intervals. For more in depth information we refer to Chapter 5.

- Chapter 4 Beating-Time Gestures: Imitation Learning for Humanoid Robots.
- Chapter 5 The Surprising Character of Music. A search for sparsity in music evoked body movements.

Chapter 6 on the contrary, is a bit particular as we look there at recent developments in machine learning. The challenge of this Chapter was to go beyond a pure technical discussion of new techniques and to come with an overview that suits also less-technically skilled readers. The main underlaying idea was to provide the reader with an understanding of the new developments in Machine Learning and this in perspective to the embodied music cognition theory.

References

[1] F. Heylighen. *Cognitive systems - a cybernetic perspective on the new science of the mind.* ECCO: Evolution, Complexity and Cognition - Vrije Universiteit Brussel, 2009.

[2] T. Ziemke. *What's that thing called embodiment.* In Proceedings of the 25th Annual meeting of the Cognitive Science Society, pages 1305–1310. Mahwah, NJ: Lawrence Erlbaum, 2003.

[3] A. Clark. *An embodied cognitive science?* Trends in cognitive sciences, 3(9):345–351, 1999.

[4] F.J. Varela, E.T. Thompson, and E. Rosch. *The embodied mind: Cognitive science and human experience.* MIT press, 1992.

[5] JJ Gibson. *The concept of affordances.* Perceiving, acting, and knowing, pages 67–82, 1977.

[6] M. Leman. *Embodied music cognition and mediation technology.* The MIT Press, 2008.

[7] A. Gabrielsson and P.N. Juslin. *Emotional expression in music performance: between the performer's intention and the listener's experience.* Psychology of music, 24(1):68–91, 1996.

[8] N.L. Wallin and B. Merker. *The origins of music.* The MIT Press, 2001.

[9] A.P. Merriam. *The anthropology of music*, volume 11. Northwestern Univ Pr, 1964.

[10] R.I. Godøy and Marc Leman. *Musical gestures: Sound, movement, and meaning.* Routledge, 2009.

[11] M. Argyle. *Bodily communication (2nd ed.).* Madison, WI: International Universities Press, 1988.

[12] E. Hanslick. *Vom musikalisch-schönen: Ein beitrag zur revision der ästhetik der tonkunst.* Johann Ambrosius Barth, 1896.

[13] R.I. Godøy. *Motor-mimetic music cognition.* Leonardo, 36(4):317–319, 2003.

[14] A. Cox. *Embodying music: Principles of the mimetic hypothesis.* Music Theory Online, 17(2):1–24, 2011.

[15] F. Heylighen. *Brain in a vat cannot break out.* Journal of Consciousness Studies, 19(1-2):1–2, 2012.

[16] L.W. Barsalou. *Grounded cognition.* Annu. Rev. Psychol., 59:617–645, 2008.

[17] M. Leman. *Fundamentals of Embodied Music Cognition: a Basis for Studying the Power of Music.* In The Power of Music, Researching Musical Experiences:a Viewpoint from IPEM, pages 17–34. ACCO Leuven Den Haag, 2013.

[18] R.I. Godøy. *Gestural Affordances of Musical Sound.* In Musical gestures: Sound, movement, and meaning, chapter 5, pages 103–104. Routledge, 2009.

[19] M. Leman. *Music, Gesture, and the Formation of Embodied Meaning.* In Musical gestures: Sound, movement, and meaning, chapter 6, pages 139–142. Routledge, 2009.

[20] M. Csikszentmihalyi. *Flow: The psychology of optimal performance*, 1990.

[21] M. Lombard and T. Ditton. *At the heart of it all: The concept of presence.* Journal of Computer-Mediated Communication, 3(2):0–0, 1997.

[22] M. Leman and R.I. Godøy. *Why Study Musical Gestures?* In Musical gestures: Sound, movement, and meaning, chapter 1, page 5. Routledge, 2009.

[23] R.I. Godøy and Marc Leman. *Gestural affordances of musical sound.* In Musical gestures: Sound, movement, and meaning, pages 108–110. Routledge, 2009.

[24] R. Laban. *Modern Educational Dance, 2d ed., rev.* Lisa Ullman (London: Macdonald and Evans, 1963), pages 29–49, 1963.

[25] R.I. Godøy, A.R. Jensenius, and K. Nymoen. *Production and perception of goal-points and coarticulations in music.* Journal of the Acoustical Society of America, 123(5):3657, 2008.

[26] P. Hodgins. *Relationships between score and choreography in twentieth-century dance: Music, movement, and metaphor.* E. Mellen Press, 1992.

[27] A. Camurri, S. Hashimoto, K. Suzuki, and R. Trocca. *KANSEI analysis of dance performance.* In Systems, Man, and Cybernetics, 1999. IEEE SMC'99 Conference Proceedings. 1999 IEEE International Conference on, volume 4, pages 327–332. IEEE, 1999.

[28] E. Ruud. *Music and identity.* Nordic Journal of Music Therapy, 6(1):3–13, 1997.

[29] D. Amelynck, M. Grachten, L. Van Noorden, and M. Leman. *Toward E-Motion-Based Music Retrieval a Study of Affective Gesture Recognition.* Affective Computing, IEEE Transactions on, 3(2):250–259, 2012.

[30] A. Camurri, I. Lagerlöf, and G. Volpe. *Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques.* International journal of human-computer studies, 59(1):213–225, 2003.

[31] E. Rusconi, B. Kwan, B.L. Giordano, C. Umilta, and B. Butterworth. *Spatial representation of pitch height: the SMARC effect.* Cognition, 99(2):113–129, 2006.

[32] J.C. Sprott. *Chaos and time-series analysis*, volume 69. Oxford University Press Oxford, UK:, 2003.

[33] B. Burger, M. R Thompson, G. Luck, S. Saarikallio, and P. Toiviainen. *Music moves us: Beat-related musical features influence regularity of music-induced movement.* In 12th International Conference on Music Perception and Cognition, Thessaloniki, Greece, 2012.

[34] S. Dahl and A. Friberg. *Visual perception of expressiveness in musicians' body movements.* Music Perception, 24(5):433–454, 2007.

[35] H.G. Wallbott. *Bodily expression of emotion.* European journal of social psychology, 28(6):879–896, 1998.

[36] J. Bicknell. *Explaining strong emotional responses to music: Sociality and intimacy.* Journal of Consciousness Studies, 2007.

[37] T.W. Schubert. *Your highness: vertical positions as perceptual symbols of power.* Journal of personality and social psychology, 89(1):1, 2005.

[38] P.-J. Maes, E. Van Dyck, M. Lesaffre, P.M.-J. Kroonenberg, and M. Leman. *The coupling of action and perception in musical meaning formation.* Music Perception, in press. submitted.

[39] M. Leman and L. Naveda. *Basic gestures as spatiotemporal reference frames for repetitive dance/music patterns in Samba and Charleston.* Music Perception, 28(1):71–91, 2010.

[40] D.A. Rosenbaum, R.G. Cohen, S.A. Jax, D.J. Weiss, and R. van der Wel. *The problem of serial order in behavior: Lashley's legacy.* Human Movement Science, 26(4):525–554, 2007.

[41] E. Carl. *Seashore, The Psychology of Music*, 1938.

[42] G. De Poli. *Methodologies for expressiveness modelling of and for music performance.* Journal of New Music Research, 33(3):189–202, 2004.

[43] J.W. Davidson. *Qualitative insights into the use of expressive body movement in solo piano performance: a case study approach.* Psychology of Music, 35(3):381–401, 2007.

[44] A. Camurri, B. Mazzarino, M. Ricchetti, R. Timmers, and G. Volpe. *Multimodal analysis of expressive gesture in music and dance performances.* Gesture-based communication in human-computer interaction, pages 357–358, 2004.

[45] J.L. Broeckx. *Muziek, ratio en affect: Over de wisselwerking van rationeel denken en affectief beleven bij voortbrengst en ontvangst van muziek.* Metropolis, 1981.

[46] I. Molnar-Szakacs and K. Overy. *Music and mirror neurons: from motion to'e'motion.* Social Cognitive and Affective Neuroscience, 1(3):235–241, 2006.

[47] V. Sevdalis and P.E. Keller. *Captured by motion: dance, action understanding, and social cognition.* Brain and cognition, 2011.

[48] E. Delaherche, M. Chetouani, A. Mahdhaoui, C. Saint-Georges, S. Viaux, and D. Cohen. *Interpersonal Synchrony: A Survey Of Evaluation Methods Across Disciplines.* IEEE Transactions on Affective Computing, 2012.

[49] K. Overy. *Making music in a group: synchronization and shared experience.* Annals of the New York Academy of Sciences, 1252(1):65–68, 2012.

[50] M.R.L. Clayton. *Observing entrainment in music performance: Video-based observational analysis of Indian musicians' tanpura playing and beat marking.* Musicae Scientiae, 11(1):27–59, 2007.

[51] G. Van Belle. *Statistical rules of thumb*, volume 699. John Wiley & Sons, 2011.

[52] F. Desmet. *A statistical Framework For Embodied Music Cognition.* PhD thesis, University of Ghent, Department of Arts, 2011.

[53] M. Leman, M. Demey, M. Lesaffre, L. Van Noorden, and D. Moelants. *Concepts, Technology, and Assessment of the Social Music Game.* In Computational Science and Engineering, 2009. CSE'09. International Conference on, volume 4, pages 837–842. IEEE, 2009.

[54] SB Kotsiantis, D Kanellopoulos, and PE Pintelas. *Data preprocessing for supervised leaning.* International Journal of Computer Science, 1(2):111–117, 2006.

[55] L. van Noorden. *The Functional Role and Bio-kinetics of Basic and Expressive Gestures in Activation and Sonification.* In Musical gestures: Sound, movement, and meaning, chapter 7, page 162. Routledge, 2009.

[56] L. Henbing and M. Leman. *A gesture-based typology of sliding-tones in guqin music.* Journal of New Music Research, 36(2):61–82, 2007.

[57] F. Desmet, L. Nijs, M. Demey, M. Lesaffre, J.P. Martens, and M. Leman. *Assessing a Clarinet Player's Performer Gestures in Relation to Locally Intended Musical Targets.* Journal of New Music Research, 41(1):31–48, 2012.

[58] D. Glowinski, A. Camurri, G. Volpe, N. Dael, and K. Scherer. *Technique for automatic emotion recognition by body gesture analysis.* In Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on, pages 1–6. IEEE, 2008.

[59] C. De Boor. *A practical guide to splines.* Applied Mathematical Sciences, 27, 2001.

[60] J.O. Ramsay. *Functional data analysis.* Wiley Online Library, 111 River Street Hoboken NJ 07030-5774 USA, 2006.

[61] W.A. Sethares and T.W. Staley. *Periodicity transforms.* Signal Processing, IEEE Transactions on, 47(11):2953–2964, 1999.

[62] B.W.W. Vines, R.L. L Nuzzo, and D.J. Levitin. *Analyzing temporal dynamics in music: Differential calculus, physics, and functional data analysis techniques.* Music Perception: An Interdisciplinary Journal, 23(2):137–152, 2005.

[63] J. Ramsay. *Functional data analysis.* Wiley Online Library, 111 River Street Hoboken NJ 07030-5774 USA, 2005.

[64] T.S. Wotton. *A dictionary of foreign musical terms and handbook of orchestral instruments.* Breitkopf & Härtel, 1907.

[65] B. Burger, M. R. Thompson, G. Luck, S. Saarikallio, and P. Toiviainen. *Music Moves Us: Beat-Related Musical Features Influence Regularity of Music-Induced Movement.* In Proceedings of the 12th International Conference on Music Perception and Cognition (ICMPC) - 8th Triennial Conference of the European Society for the Cognitive Sciences of Music (ESCOM)., 2012.

[66] B. Caramiaux. *Studies on The Gesture-Sound Relationship in Musical Performance).* PhD thesis, France, Institut de recherche et coordination acoustique et musique, 2011.

[67] P. Toiviainen, G. Luck, and M.R. Thompson. *Embodied meter: Hierarchical eigenmodes in music-induced movement.* Music Perception, 28(1):59–70, 2010.

[68] J. Almansa and P. Delicado. *Analysing musical performance through functional data analysis: rhythmic structure in Schumann's Träumerei.* Connection Science, 21(2-3):207–225, 2009.

[69] T.M. Mitchell. *Machine learning. WCB*, 1997.

[70] Y. Björnsson, V. Hafsteinsson, A Jóhannsson, and E. Jónsson. *Efficient use of reinforcement learning in a computer game.* Proceedings of International Journal of Intelligent Games & Simulation, 2008.

[71] L. Nijs, P. Coussement, C. Müller, M. Lesaffre, and M. Leman. *The Music Paint Machine: A multimodal interactive platform to stimulate musical creativity in instrumental practice.* In 2nd International conference on Computer Supported Education (CSEDU 2010). IEEE, 2010.

[72] E. Schubert. *Measuring emotion continuously: Validity and reliability of the two-dimensional emotion-space.* Australian Journal of Psychology, 51(3):154–165, 1999.

[73] P.-J. Maes, D. Amelynck, M. Lesaffre, M. Leman, and DK Arvind. *The "Conducting Master": An interactive, real-time gesture monitoring system based on spatiotemporal motion templates.* International Journal of Human-Computer Interaction, 2012.

[74] A. Vakanski, I. Mantegh, A. Irish, and F. Janabi-Sharifi. *Trajectory Learning for Robot Programming by Demonstration Using Hidden Markov Model and Dynamic Time Warping.* Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 42(4):1039–1052, 2012.

[75] Y.W. Teh. *Dirichlet Process.* Submitted to Encyclopedia of Machine Learning, 2007.

[76] K. El-Arini. *Dirichlet Process : A gentle tutorial.* Select Lab Meeting, 10 2008.

[77] M. Clynes and Y. Menuhin. *Sentics: The touch of emotions.* Anchor Press Garden City, NY, 1977.

[78] A. Friberg and J. Sundberg. *Does music performance allude to locomotion? A model of final ritardandi derived from measurements of stopping runners.* The Journal of the Acoustical Society of America, 105:1469, 1999.

[79] A. Friberg, J. Sundberg, and L. Frydén. *Music from motion: Sound level envelopes of tones expressing human locomotion.* Journal of New Music Research, 29(3):199–210, 2000.

Essentially, all models are wrong,
but some are useful.
George E. P. Box , 1987

2
Towards E-Motion Based Music Retrieval

A study of Affective Gesture Recognition

Abstract

The widespread availability of digitized music collections and mobile music players have enabled us to enjoy music during many of our daily activities, such as physical exercise, commuting, relaxation. A practical problem that comes along with the wish to listen to music is that of music retrieval, the selection of desired music from a music collection. In this paper we propose a new approach to facilitate music retrieval. Modern smart phones are commonly used as music players, and are already equipped with inertial sensors that are suitable for obtaining motion information. In the proposed approach, emotion is derived automatically from arm gestures, and is used to query a music collection. We derive predictive models for valence and arousal from empirical data, gathered in an experimental setup where inertial data recorded from arm movements is coupled to musical emotion. Part of the experiment is a preliminary study confirming that human subjects are generally capable of recognizing affect from arm gestures. Model validation in the main study confirmed the predictive capabilities of the models.

2.1 Introduction

The widespread availability of digitized music collections and mobile music players has enabled us to listen to music during many of our daily activities, such as physical exercise, commuting, relaxation, and many people enjoy this. A practical problem people face when they want to listen to music is the selection of desired music from a music collection. The bibliographic, text-based interface to music-collections that is prevalent in mobile music players today, is not an optimal solution to this problem for two major reasons. Firstly, bibliographic indices to music, such as artist and album names, are only useful when the user is familiar with the music he/she is looking for. Secondly, a text-based visual interface on small screens of mobile devices is often impractical to use. It requires a lot of attention and a fine motor control of the user, which can be cumbersome and even dangerous in everyday life situations.

The basic tenet of affective computing, as stated by Calvo and D'Mello [1], is that automatically recognizing and responding to a user's affective states during interactions with a computer can enhance the quality of the interaction, thereby making a computer interface more usable, enjoyable, and effective. This may be particularly true in the context of interfaces for music players, since music and affect are strongly related. Not only is it natural for people to describe music in affective terms; studies have also suggested that the most common purpose of musical experiences and in particular of music listening is to influence emotions: People use music to influence their emotions, to enjoy or comfort themselves, and to relieve stress [2].

Of the various forms an affection-based interface to music players might take, motion-driven approaches seem especially promising. One reason for this is that there is ample evidence that corporal gestures are a very effective way of communicating affect among people (see section 2.2). Another, more pragmatic reason is that many smart phones that people use as music players nowadays, are equipped with inertial sensors that make it possible to capture movements of the user. This opened a whole new world of applications and we refer to Synch'n'Move [3] as being just one example.

In our envisioned interface, users can search through music collections based on the affective character of the music, where the character of the desired music is expressed through corporal gestures. In this way we aim to implement the conceptual framework of *embodied music cognition* and *mediation technology* [4], and reduce the gap between the fundamentally corporeal aspects of music and the disembodied, bibliographical way of interacting with music collections that is common practice today.

The work presented in this paper is intended as the foundation for such a motion based affective user interface for music retrieval. We present a linear regression model that predicts the affective character of music, based on the arm movements of people expressing that character. The model is derived from empirical data that is gathered from an experiment, as described in section 2.3. A motion based interface can employ this model to interpret arm movements of the user in terms of affective character, so that the movement can be matched to the affective character of music. For this, it is also necessary to have a music collection that is annotated in affective terms. Automated affective description of music is beyond the scope of this paper, but this is an active field of research in its own right (see e.g. [5–7]).

It is commonly acknowledged that the notion of *affect*, and subsumed notions such as *emotion*, and *mood*, are notoriously intricate. The study of affect in combination with music is by no means less intricate and controversial. First of all, some studies question whether the emotions music evokes should be considered as basic emotions [8], whereas others consider this view mislead [2]. There is also some disagreement about the question whether music is more adequately described as inducing *mood* (a relatively vague and long-lasting form of affect), or *emotion* (more instantaneous and focused forms of affect) [2]. Furthermore, perception of music has been shown to influence neuroaffective processes [9]. Other studies show, however, that the strength of emotions induced by music is relatively low compared to emotions induced by personal memories [10].

In light of these controversies, it is useful to clarify our use of notions of affect in this paper, and the corresponding assumptions we make. To begin with, we focus on the affective character of the music, as expressed by the listener. We will also refer to this as the *emotional* character of the music, because of the instantaneous and concrete nature of the music. More specifically, we adhere to the *valence-arousal* model [11] to represent the emotional character of music, in line with other studies about emotion in music [5]. Although extended versions of this model, including a third dimension representing *tension*, have been proposed there is evidence that the three-dimensional model does not account better for music related emotions [12]. Our focus is on *expressed emotions*, expressed emotions being the embodiment of emotions through movement. Expressed emotions can find their origin in perceived or induced emotions but are by nature distinct as the link between perceived and expressed emotions or between induced and expressed emotions is not always transitive [13].

The next section contains a brief review of related work on both human and automated affect detection from human movement. In section 2.3, we describe the experiments carried out to gather the movement data for affect recognition. In section 2.4, we present the data-modeling process, and the results of the model validation. A discussion of the results is presented in section 2.5, followed by conclusions and directions for future work, in section 2.6.

2.2 Related Work

A natural way for humans to express affect is by corporal gestures [14, 15]. Communication of affect through gestures (both static and dynamic) is arguably an intrinsic part of social behavior. This is reflected in numerous studies showing the capability of humans to recognize affect from the corporeal behavior of others. To a large extent, movement seems to convey affective information. For example, to recognize emotion from gait, a small number of features describing joint angles and spatial trajectories is sufficient for humans to recognize emotions in animated avatars [16]. Furthermore, Atkinson et. al [17] show that even with very reduced visual representations of the body, such as point-light displays, recognition of emotion by human subjects is still possible (though to different degrees for different emotions). Point-light displays of arm movements of actors expressing affect in everyday movements, like drinking and knocking, also enable observers to recognize the expressed affects [18]. Pollick et al. also found that movement features such as average velocity, peak velocity, acceleration, and jerk were all correlated with the level of activation.

Music-related body movement, such as that of dancers, and the performing musicians, also conveys affect. Brownlow and Dixon [19] state that observers easily can judge happy dances as happier and stronger than sad dances. Again, the observers in their experiment based their judgment solely upon point-light displays of dance, thus excluding recognition of affect by facial expression or other cues. Successful *automatic* recognition of emotions of dance movements has been reported [20]. Vines and Wanderley [21] analyzed gestures from professional clarinet performers. They confirm that the visual component (body movement) of the performance carries much of the same structural information as the audio. In some conditions, removing the visual component decreases the judgment of tension (emotion).

These studies strengthen the view that affect can be effectively communicated through human body movement, and therefore, that automatic affect recognition from human motion, even if it is a challenging problem (see [1] for a survey of current research), is feasible.

2.3 Experimental Set-Up

An experiment was carried out with the goal of building a data set of arm movements expressing the affective character of different pieces of music. The design of the experiment is oriented towards the use case of gesture based music retrieval in mobile devices, in the sense that arm motion is captured using a wireless handheld device equipped with 3D inertial sensors, comparable to the motion-capture technology available in smart phones.

The following setup was designed to link movements to affective descriptions of music: Participants were asked to listen to a musical fragment. Then, they were asked to listen to the music again, and simultaneously express the emotional character of the music as clearly as possible through the movement of their arm (either the left or right arm, depending on preference). Eventually they were asked to describe the emotional character of the music, in terms of valence and arousal. The movement of the arm was observed by three other participants, who had to guess the emotional character of the music being expressed, judging only on the arm movement. One reason for including observing participants in the setup is to encourage the observed participants (called *performers* henceforth) to *communicate* the intended emotion through the movement, rather than making just any movements associated with the music. A second reason is that the degree of agreement between the intended emotion and the emotion recognized by observers serves as an indicator of how clearly the intended emotion is expressed by the movement.

The rest of this section describes the experimental setup in more detail.

Participants

In total 32 persons participated in the experiment. Among these participants, five groups of four persons were made who participated in the main part of the experiment. The remaining 12 persons participated individually. Their responses were used to validate the model derived from the data obtained in the main experiment, as described below.

Stimuli

The musical material was selected from a pre-existing library of 30 second musical excerpts [22]. In total 24 musical fragments (table 2.1 and 2.2) were selected divided over four similar sets of six fragments. The sets are separated by a double line in both tables. Similarity of the sets was controlled after the experiment. From the results shown in table 2.3 it can be verified that the sets were indeed homogeneous and that they spanned the whole

Performer - Title	**Arousal**	**Valence**
New Zealand Symphony Orchestra - Many Meetings	1.4±0.5	3.6±0.5
Midori/Berliner Philharmoniker / Claudio Abbado - Canzonetta. Andante (Concerto for Violin and Orchestra in D major op. 35)	2.6±0.9	2.0±0.7
Tam Echo Tam - One Step	4.2±0.4	4.4±0.5
L' Arpeggiata / Christina Pluhar - Ah, vita bella	1.6±0.9	1.8±0.8
Blur - Song 2	4.8±0.4	4.4±0.9
DJ Tiësto - Traffic	4.6±0.5	3.4±1.1
Metallica - St. Anger	4.0±0.7	2.8±0.8
De Nieuwe Snaar - Achterbank	3.8±0.8	5.0±0.0
Enya - Orinoco Flow	2.2±0.8	3.4±0.5
The Cleveland Orchestra/Pierre Boulez - Le Sacre du Printemps	5.0±0.0	2.4±1.1
Alberto Gilberto - The girl from Ipanema	1.6±0.9	3.8±1.1
New Philharmonia Orchestra/Sir John Barbirolli - Adagietto, Sehr langsam Symphony No. 5 in C sharp minor	1.4±0.5	1.8±0.8

Table 2.1: Musical Fragments (part 1) and their average Arousal/Valence appraisal scores as given by the Performers. (Scores ± SD are on a 1 to 5 scale)

valence/arousal range. The arousal and valence scores mentioned in both tables are average appraisal scores collected from the performers.

Each of the 24 musical fragments was rated once in each of the five participant group, resulting in 120 ratings in total, and five ratings per fragment.

Material

For capturing arm movement, a Wii Remote was used. This is a wireless, handheld device commercially available as a gaming interface from Nin-

Performer - Title	**Arousal**	**Valence**
Esa-Pekka Salonen / Philharmonia Orchestra - Car Horn Prelude (Le Grand Macabre)	4.0±1.0	1.4±0.5
Bob Marley - Corner stone	3.0±0.7	5.0±0.0
Beyoncé - Naughty Girl	3.8±0.4	3.8±0.4
Astor Piazzolla - Oblivion	1.2±0.4	1.8±0.8
Metallica - My World	4.8±0.4	1.6±0.9
Manu Chao - Mr. Bobby	2.4±1.1	4.6±0.5
Novastar - Never back down	2.6±1.1	3.4±1.5
David Hill / Westminster Cathedral Choir - Motectum (Requiem, Officium defunctorum)	1.2±0.4	1.2±0.4
Usher - Usher	4.6±0.5	4.2±0.8
Vladimir Ashkenazy - Nocturne in F major op.15 No.1	1.4±0.5	2.8±1.3
St. Germain - Land of ...	3.8±0.4	4.4±0.5
Collegium Vocale & La Chapelle Royale/Orchestre des Champs Elysées/Philippe Herreweghe - Dies Irae (Requiem KV 626)	4.4±0.5	1.6±0.5

Table 2.2: Musical Fragments (part 2) and their average Arousal/Valence appraisal scores as given by the Performers. (Scores ± SD are on a 1 to 5 scale)

tendo. It transmits 3D inertial sensor data in realtime via Bluetooth at a sample rate of 100Hz. Musical material was played to the participants from a computer, using wired headphones. Visual recordings of arm movements were transmitted in real-time, using a digital video camera. Judgments of emotional character were obtained from participants through printed questionnaires.

Procedure

Within a group of four participants, one set of six musical fragments was assigned to each participant, such that each fragment was uniquely assigned

Set	**Arousal : mean ± stdev**	**Valence : mean ± stdev**
1	3.2 ± 1.5	3.3 ± 1.1
2	3.0 ±1.5	3.2 ± 1.1
3	3.2 ± 1.3	3.0 ± 1.6
4	3.0 ± 1.5	2.9 ± 1.3

Table 2.3: Sets of fragments and their statistical Arousal/Valence dispersion. (Scores on a 1 to 5 scale)

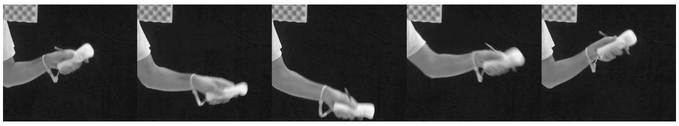


Figure 2.1: Video capture of performance as monitored by the observers

to a participant within the group. Every participant was asked in turn to listen to each of the fragments assigned to him/her and to express this character by arm movements while listening to the fragment again, and judge its emotional character (dealing with one fragment at a time). The arm movements were made while holding the Wii Remote, in front of a camera that was positioned in such a way that only the arm was monitored, as illustrated in figure 2.1. A small shield was used to prevent the performers' faces from occasionally appearing on the screen.

The instructions for the performer were as follows:

1. Listen to a short musical fragment

2. While listening a second time, express the emotional character of the music as accurately as possible through the movements of your arm

3. Rate the emotional character of the music on the provided form

We focus on *expressed* emotions (being the embodiment of emotions through movement). By instructing the performers to rate the emotional character of the music immediately following the performance, we anticipated that the performers would rate expressed emotions. In our experiment, this was confirmed by having observers validate performers' rating scores. It is worthwhile mentioning that for experiments on induced or perceived emotions where verification by an external source is unachievable a more thorough instruction set is required : we would like to refer, for example, to the instruction set used by GEMS [23].

The emotional character of musical fragments was rated in terms of valence and arousal on a 1 to 5 scale. Rather than using the terms valence and arousal directly, the semantics of the two scales was indicated by labeling the extremes of the scales with corresponding adjectives. The adjectives were given in Dutch: *kalm*, *vermoeid* (calm, tired) versus *energetisch*, *gespannen* (energetic,tense) to label the low and high extremes of arousal respectively, and *droevig*, *kwaad* (sad, angry) versus *blij*, *tevreden* (happy, pleased) for valence. It was explained to the participants that a single matching adjective was sufficient to rate a musical fragment correspondingly. For example, it is sufficient for either *kalm* or *vermoeid* to apply, in order to choose that rating.

The three other participants, referred to as *observers*, watched the arm movements of the performer via a monitor in a separate space (figure 2.1). They did not hear the music fragments the performer heard.

The observers were instructed as follows:

1. Monitor the (arm movement of the) performer.

2. Rate the emotions expressed by the arm movement.

3. Describe any cues in the motion that helped you to make your rating (free text)

The remaining 12 subjects participated in the role of performer, as described above. Each subject was assigned again a group of six fragments. This time, no observers were present. The arm movements and the subject's rating of the emotional character of the fragments was recorded as before. The data obtained in this way is used for validation, as described in section 2.4.

2.4 Results

The data obtained in the experiment was used to create a regression model for predicting the expressed arousal and valence from arm movements as captured by the 3D inertial sensors of the Wii remote. We aim at a general data model that can easily be ported to other devices, possibly using other sensing technologies. Therefore only predictor variables with a relatively straightforward relationship to movement were considered. This preference of general validity over a fine-tuned model leads us to consider only models with at most five predictor variables.

The Wii Remote measures the acceleration of the device in the direction of three perpendicular axes, relative to the device. Since the way of holding the device was not constrained, similar arm movements may lead to different data, as an effect of the Wii Remote being held in different ways. To compensate for this, the acceleration data for each fragment was projected onto its three principal components by performing a principal component analysis (PCA). In figure 2.2 we show acceleration data collected from two different subjects performing on the same musical fragment. In this figure an acceleration value of 25 corresponds with 1G (gravity). From this figure it is very difficult to see similarities between the two performances. When the data is translated and rotated to the PCA-axes, the similarity between these two performances becomes more apparent (Figure 2.3). Apart from making data from different subjects easier to compare, an advantage of the PCA transform is that it reveals the intrinsic dimensionality of the movement.

To determine a set of candidate features to compute from the accelerometer data, we made an inventory of the free text responses in which subjects reported useful cues for judging the emotional character of the movements. The cues can be roughly grouped into five complementary aspects of the movement:

1. Roughness: gracefulness, multiple short moves
2. Rhythm: tempo of the music
3. Speed: high speed, low speed, and acceleration
4. Size: large versus small movements
5. Location of the arm: high = happy

Ideally, each of the cue categories should be represented by at least one predictor variable.

We extract various features that describe various properties of the distribution of acceleration data, in terms of geometry and density (Figure 2.3). The same was done for jerk (derivative of acceleration) and speed (integral of acceleration). Beside these spatial features we calculated also a number of time related features such as peak-rates, zero crossings and randomness (runs test). Summarized the following features were extracted :

- Distribution properties for acceleration, jerk and speed along the 3 PCA axes : mean, range, standard deviation, kurtosis, skewness.
- Distribution properties for direction: circular standard deviation, concentration parameter kappa (Von Mises distribution: $\boldsymbol{\kappa}$).

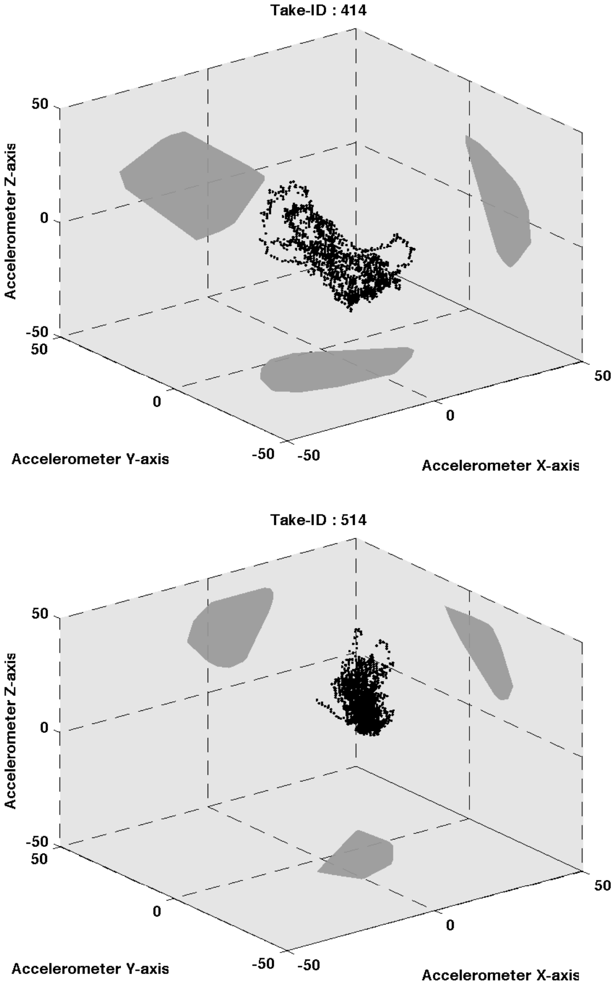


Figure 2.2: Raw acceleration data from accelerometer : Performances of two subjects on the same musical fragment (set 1, fragment 4)

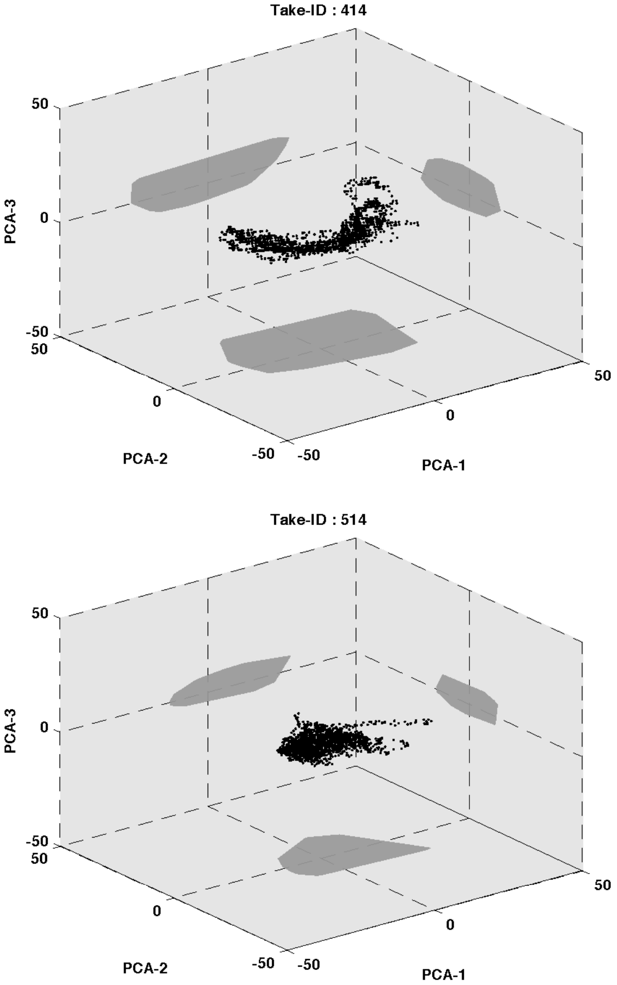


Figure 2.3: Acceleration data translated and rotated to PCA-axes : For same performances as shown in Figure 2.2.

Directional data is analyzed by means of unit length vectors. All angular representations of the acceleration vector are converted to vectors on the unit sphere setting their radius to one. From these representations the mean resultant length ($\bar{r}$) is calculated. The sample circular standard deviation $\boldsymbol{s}$ is then calculated as $\boldsymbol{s = \sqrt{-2 * ln(\bar{r})}}$. The concentration parameter $\boldsymbol{\kappa}$ is calculated by solving the equation $\boldsymbol{\rho = coth(\kappa) - 1/\kappa}$ substituting $\bar{r}$ for $\boldsymbol{\rho}$ [24].

- Volume (convhull) of acceleration point cloud.

  We define volume as the volume inside the convex hull or convex envelope of the acceleration trajectories deployed in space.

- Time related variables: speed peak rate, zero crossing rate, randomness (runs test).

  Starting from a peak detection algorithm, the local minima and maxima in a signal are detected. The number of peaks over a time interval is then the peak rate. Zero crossing checks the number of alternations between a positive and a negative signal value. The number of crossings divided by the time interval results in the zero crossing rate. The runs test is a test based on the number of runs of consecutive values above or below the mean of x. Too few runs indicate a tendency for high and low values to cluster. Too many runs indicate a tendency for high and low values to alternate. The number of runs is taken here as a feature variable.

These features (149 in total) can be linked to the cue categories identified before, with the exception of location cues[1].

- Roughness (total: 18): all jerk related features
- Rhythm (total: 16): time related variables
- Speed (total: 75): Speed and acceleration features
- Size (total: 40): Direction parameters, volume of acceleration point cloud.

To remove any transient effects due to subjects starting or stopping to move, the features are extracted after removing the first and last 5 seconds of each data stream. The remaining stream spanned 20 seconds (corresponding to 2000 samples).

[1] Although the position of the Wii Remote can in principle be estimated by assuming an initial position and tracking acceleration over time, this estimation is unusable in practice, due to cumulative estimation errors.

Because we are faced with a large number of features, the next step comprised feature selection or Feature Subset Selection (FSS) and aimed at selecting a subset of relevant features for use in model construction. FSS methods belong to two categories: filters and wrappers. In the filter approach, features are selected based upon data properties independent of the learning [2] algorithms. In the wrapper approach, feature selection does use the learning algorithms. Our approach uses both methods: first a filter algorithm, followed by a wrapper algorithm.

For the filter algorithm, we correlated the extracted features with valence and arousal. There were strong correlations between some features and arousal ($|\boldsymbol{r}| > \mathbf{0.6}$) but in general weaker correlations with valence (all $|\boldsymbol{r}| < \mathbf{0.4}$). Using the nomenclature from Bell [25] filtering should look for what he calls relevant features. These are features that have an influence on the output and their role can not be assumed by the rest. The criterion we used for relevance, was to discard all uncorrelated features ($|\boldsymbol{r}| < \mathbf{0.2}$) from the analysis. Technically spoken, this is not 100% correct as also a linear combination of two uncorrelated features still might correlate but it was sufficient for our goal: a good model that validates.

Dropping these irrelevant features, we are still left with a number of features that are highly inter-correlated ($|\boldsymbol{r}| > \mathbf{0.9}$). Therefore it was also necessary to take precautions against multicollinearity.

Using all (the remaining) features in a least squares estimate as a regression model would present us with two problems [26]. The first problem is prediction accuracy: the least squares estimates often have low bias but large variance. Prediction accuracy can however sometimes be improved by shrinking or setting some coefficients to zero. By doing so we sacrifice a little bit of bias to reduce the variance of the predicted values, and hence may improve the overall prediction accuracy. The second problem is interpretation. With a large number of predictors, we often would like to determine a smaller subset that exhibits the strongest effects.

Because of our restriction to models of maximally five predictor variables, the method of 'best subset selection' was used. By retaining a subset of the predictors and discarding the rest, subset selection produces a model that is interpretable but has a possibly higher prediction error than the full model. Best subset selection also gives a hard threshold on how many parameters to keep. Other shrinkage methods (like lasso or ridge regression) may give less variability but the number of parameters is soft-thresholded. The choice for least squares estimates and for best subset selection implied the need of a data validation step to check the generalization capabilities of the calculated data model.

[2] In our case: a regression algorithm

The filtering algorithm reduced the set of variables to a smaller subset. It is this smaller subset that was presented to SPSS for calculating a regression model using the stepwise method. This method is in essence a wrapper method: Each variable is entered in sequence and its value assessed. If adding the variable contributes to the model then it is retained, but all other variables in the model are then re-tested to see if they are still contributing to the success of the model. If they no longer contribute significantly they are removed. In this way this method ensures that the smallest possible set of predictor variables is included in the model [27].

From now onwards we will make a distinction between data modeling for arousal and data modeling for valence.

2.4.1 The Regression Model for Arousal

Set-Up of the Arousal Model

The regression model for Arousal was derived from 97 performances out of a total of 120 performances. 23 cases where performer and observers did not agree (Difference > 1) were discarded . In other words, in over 80% of the cases there was an agreement between performer and observer. We use SPSS and the stepwise method to enter the predictor variables, resulting in a model with three predictor variables. The variables (listed in their order of contribution importance) are the following (beta-values in table 2.4):

SpeedPeakrate: The number of local maxima and local minima for speed (integral of acceleration) divided by the time interval. Speed is calculated as an Euclidean norm.

KurtPCA1Speed: Kurtosis of the distribution of speed along the first (main) principal axis. This variable is negatively correlated with arousal. A high value means that intermediate values have become less likely and the central (higher peak) and extreme values (fat tails) have become more likely. In other words low arousal corresponds with long periods of low speed (central values) and other periods of high speed (extreme values). High arousal corresponds with periods of nearly constant speed or where the variation in speed is not huge (intermediate values).

PCA3Std: Standard deviation of the distribution of acceleration along the third principal axis. A small value indicates that the acceleration mainly happens in a plane formed by the two main principal axes.

The regression analysis did not reveal any outliers. (Criterion used: more than three standard deviations difference). There was however one influential case (group 5 subject 3 fragment 5) that ended up with a high

Model	**B**	**SE B**	**Beta**	**Sig.**
(Constant)	**2.096**	**0.410**		**0.000**
Speedpeakrate	**0.560**	**0.056**	**0.680**	**0.000**
PCA3Std	**0.043**	**0.014**	**0.213**	**0.002**
kurtPCA1Speed	-0.585	**0.149**	-0.212	**0.000**

R = 0.854 (R^2 = 0.730 adjusted R^2 = 0.721)

Table 2.4: Regression analysis for Arousal (Training Data).
- B = Parameter values of the regression model
- SE B = Standard Error for the parameters
- Beta = Standardized version of the B-values. Tells us the number of standard deviations that the outcome will change as a result of one standard deviation change in the predictor.
- Sig. = if $<$ **0.05** *the B-value is significant different from zero and the parameter contributes to the model.*

value for DFFit (Difference in Fit). In order to preserve the general character of our model we removed this case and recalculated our regression model. An overview of the recalculated model can be found in table 2.4.

Because of the correlations between features, the following assumptions were checked:

1. Multicollinearity: VIF (variance inflation factor) average was close to 1 (1.4) and indicated absence of multicollinearity between the 3 predictor variables.

2. Normality for distributed errors: Probability plot for the residuals confirmed normality.

Validation of the Arousal model:

Model validation was done using the data gathered from the 12 individual subjects, who did not participate during the main part of the experiment.

Explanatory capabilities of the model: The variance of the validation data explained by the model : R = 0.754 (R^2 = 0.568). Compared to R = 0.854 (R^2 = 0.730) for the original data, this means a shrinkage with 16%.

The predictive capabilities are presented in Fig. 2.4. We see that the average prediction from the model deviates most for low arousal values. For other arousal values, the prediction is in line with the target value, although

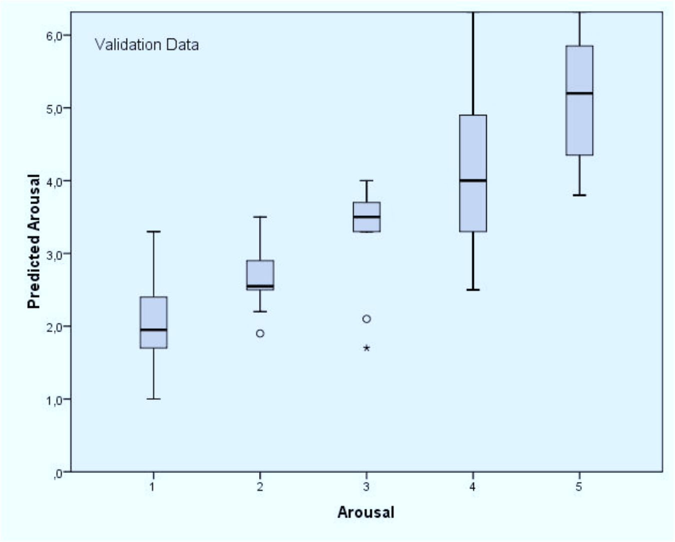


Figure 2.4: Validation data.
X-axis: Arousal values as judged by the performer (Likert scale 1 to 5).
Y-axis: Predicted arousal values using the model from table 2.4

Residuals	Minimum	Maximum	Mean	StdDev
Model	-1.553	**1.880**	**0**	**0.740**
Validation	-4.888	**1.506**	-0.494	**1.083**

Table 2.5: Residual statistics.

the variation is higher for high arousal values. We checked for sparsity in the training database as a possible explanation but this was not the case. A part of the effect is explained by the censoring mechanism due to cutoff limits at arousal value 1 (minimum) and at arousal value 5 (maximum). It was also in these areas that we spotted most of the disagreement between performers and observers.

Further investigation was done by having a closer look at the residuals. The results of this analysis are mentioned in table 2.5. The large residual value of -4.888 is due to an out of scale prediction of 9.888 for an arousal value of 5. Allowing non-linearity by replacing values outside the boundaries of 1 and 5 with their respective boundary values, reduces the error, and the explanatory value of the model is increased ($R^2 = 0.685$). This leads to a reduction of only 4.5 % compared to the original data model.

2.4.2 The Regression Model for Valence

The regression model for valence was derived from 88 out of a total of 120 performances. 32 cases where performer and observers did not agree (Difference > 1) were discarded. We discarded considerably more samples (27 %) than for arousal (19 %).

As was the case with the arousal model, we started with the stepwise method to add variables to the model. In a first step we obtained a model with five predictor variables. There were no outliers but there was one influential case. Group 2 Subject 3 Fragment 5 ended up with a high value for DFFit. In order to preserve the general character of our model we removed that case and recalculated our model. Recalculation led to the removal of two more predictor variables that had no significant contribution. The final result was a model with again three predictor variables. The variables are hereafter listed according to the importance of their contribution(beta-value), see table 2.6:

stdPCA1Jerk: Standard deviation of the derivative of the acceleration (jerk) of the first PCA component. This variable correlated negatively with valence. If acceleration changes nearly have a random pattern (high standard deviation), this will result into a lower valence.

SpeedPeakrate: See subsection 2.4.1.

PCA2std: Standard deviation of the second principal component. A small value means that acceleration/movement happens mainly along the axis of the first principal component rather than in a plane. This variable correlated positively with valence. In other words : for low valence, the movement is rather one dimensional (1D).

A complete overview of the model can be found in table 2.6.

Assumptions checked:

1. Multicollinearity: (VIF variance inflation factor). The VIF never exceeded 10, but the average over all variables is situated well above 1 (4.4). So there might be some moderate bias in the model.

2. Normality for distributed errors: Probability plot indicates that the distribution is slightly skewed left.

The R^2 value of 0.367 for valence is relatively low compared to a value of 0.730 for arousal. A possible reason for this is that the model contains no predictor variable representing location, although observers reported this cue as indicative for valence. Even if location cannot be estimated directly from the accelerometer data, an estimate of position can be made indirectly: Because of the fact that the Wii Remote device is ergonomically designed for one particular way of grasping, in practice subjects held the Wii Remote all

Model	B	SE B	Beta	Sig.
(Constant)	**1.083**	**0.354**		**0.003**
Speedpeakrate	**0.751**	**0.116**	**1.018**	**0.000**
StdPCA1Jerk	-0.238	**0.046**	-1.148	**0.000**
PCA2Std	**0.058**	**0.021**	**0.461**	**0.008**

R = 0.606 (R^2 = 0.367 adjusted R^2 = 0.344)

Table 2.6: Regression analysis for Valence (Training Data).
- B = Parameter values of the regression model
- SE B = Standard Error for the parameters
- Beta = Standardized version of the B-values. Tells us the number of standard deviations that the outcome will change as a result of one standard deviation change in the predictor.
- Sig. = if < **0.05** *the B-value is significant different from zero and the parameter contributes to the model.*

in the same position. Additionally, it is reasonable to assume that raising the arm leads to a different angle of the hand than lowering it, due to physiological constraints. By making these extra assumptions, location can be estimated as the rotation of the device along its *pitch* axis, comparable with nose up (pitch>0) or nose down (pitch<0) for a plane.

The contribution of the pitch variable to the regression model was slightly below the contribution of the strongest variables. Because the pitch variable did not explain more or additional variance, we did not include it in the model here.

Validation of the Valence model:

Model validation was done again on the validation set.

Explanatory capabilities of the model: The variance of the validation data explained by the model : R = 0.532 (R^2 = 0.284). Compared with R = 0.606 (R^2 = 0.367) for the original data, this means a shrinkage with 8.3 %.

The predictive capabilities of the model are presented in Fig. 2.5. As expected with the lower R^2 values, the predicted (main) values for valence are closer to the mean. Most variation in prediction is found for low valence values.

Further investigation was done by having a closer look at the residuals. The results are in table 2.7.

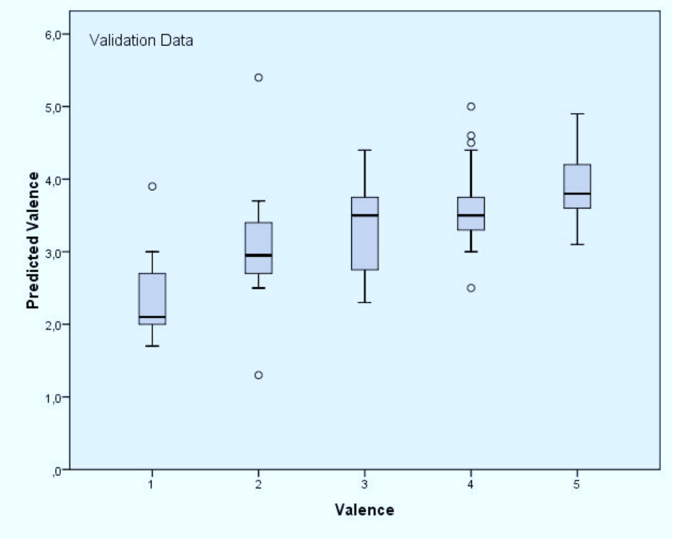


Figure 2.5: Validation data.
X-axis: Valence as judged by the performer (Likert scale 1 to 5).
Y-axis: Predicted valence values using the model from table 2.6.

Residuals	**Minimum**	**Maximum**	**Mean**	**StdDev**
Model	-2.031	**2.203**	**0**	**0.970**
Validation	-3.394	**1.901**	-0.255	**1.073**

Table 2.7: Residual statistics.

Standard Deviation for residuals is 1.073 and that is close to the standard deviation of the model. There is one residual with an excessive value of -3.394. This case is associated with a low valence value (value = 2).

2.5 Discussion

The data model for arousal explains 73 % ($\mathbf{R^2}$-statistic) of the variance for the original sample and 68.5 % ($\mathbf{R^2}$-statistic) of the variance for the validation samples. These are high values and endorse the good predicting capabilities of the model. The small shrinkage (4.5 %) from the original sample to the validation data confirms the generalization capability of this model. The data model for valence resulted in a value for the $\mathbf{R^2}$ statistic of 36.7 % for the original sample and of 28.4 % for the validated data. This is a shrinkage with 8.3 %. The generalization of the valence model is clearly less than for arousal and its predicting capabilities are also clearly less.

We have tried to remedy a possible cause for this, namely that the accelerometer data do not allow for a good estimate of location to be made. However, an post hoc heuristic to estimate location indirectly did not improve results.

Another explanation for the lower prediction results of valence is that valence related aspects of movement are ambiguous, in the sense that human observers are also less successful in recognizing valence accurately. This is reflected in the fact that for valence, a larger proportion of the experimental data was discarded due to lack of agreement between intended and observed valence. The higher ambiguity of valence compared to arousal is also on a par with the findings of Pollick et al. who stated that the second dimension of affect, pleasantness, was less correlated with any of the considered movement features [18]. A possible explanation is that sad music is not systematically associated with negative valence [28] [29]. Although sadness is generally considered to be an unpleasant emotion, the classification is not straightforwardly applied to music. Sad music is often considered beautiful, and therefore it may be difficult to perceive sadness in music as unpleasant [12].

The models presented here are based upon motion data from arm gestures as input. To our knowledge, experiments attempting to detect musical affect from movement using inertial sensors are as of yet very scarce. What has been done before is affect detection from music audio signals. Lie Lu [30] obtained classification results from the four quadrants of the arousal-valence space with a resulting accuracy of 76-94 %. The results of a classification study (containing 4 classes) can not be compared directly with a regression study but they do confirm the feasibility of an affect based music retrieval system. Another reference is for example made to the study of Yi-Hsuan Yang [31]. In his research a support vector regression model was used based on timbral texture features (spectral centroid, spectral rolloff, spectral flux and MFCC) and MPEG-7 features.

They obtained an $R^{\mathbf{2}}$-statistic of 79.3 % for arousal, and 33.4 % for valence, which corresponds well to the results presented here (68.5 % for arousal and 28.4 % for valence).

Apart from the accuracies obtained for predicting arousal and valence, the cue categories identified in observers' responses are likewise similar to those reported in other studies, such as a study on dance movements, where full-body movement was judged [20]. Similar movement cues (irregularity, fluency, speed, amount) were also identified in a study on the visual perception of expressiveness in musicians' body movements [32].

The data regression models for predicting arousal and valence are the key building blocks to form the envisioned application of an affect based music retrieval system. The data models project arm movement data into a point onto the valence/arousal plane used to describe emotion. What is missing for a complete affect based music retrieval system is the annotation of a music library and the construction of playlists. For automatic mood-annotation, there are several applications possible such as the model of Yi-Hsuan Yang [31] mentioned before. However, all these applications can not capture the possibly idiosyncratic relation between the expressed emotions and the music. The models we developed can overcome this issue because they allow annotation of music collections by movement, rather than by textual annotation. As this is potentially a tedious task, a realistic implementation will start with a default library that is automatically annotated and will personalize the annotations by movement as per requirement. A straight-forward method for constructing a playlist is to select songs that come close to the projected point in the valence/arousal plane. Such a method could possibly reflect more differences in valence than in arousal, to compensate for the lesser quality of the data model for valence. This, however, is an issue of playlist-creation, and what constitutes a ”good” playlist also depends on the expectations of the end-user. An end-user study would be required to gain more insight in this domain.

The regression models were derived from movement data limited to data streams of 20 seconds. For a music retrieval system, requiring 20 seconds of movement to retrieve music is probably unacceptably long. Therefore it is worthwhile to investigate the impact of reducing the query duration. We simulated a shorter query time by reducing the analysis interval from 2000 samples (20 seconds) to 500 samples (5 seconds) and investigating the prediction errors at every step. In figure 2.6, the impact of using shorter retrieval intervals is visualized. The data set used for this investigation is the validation data set. The impact on the prediction errors was measured by the normalized root mean square error (NRMSE):

$$\boldsymbol{NRMSE = \frac{\sqrt{\frac{\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2}{n}}}{(Ymax - Ymin)}} \tag{2.1}$$

With $\boldsymbol{\hat{Y}_i}$ being the value for valence/arousal calculated from the models and $\boldsymbol{Y_i}$ being the real valence/arousal value (appraisal by the performer). Reducing the retrieval time clearly results in less accurate predictions. The loss of precision is however rather small. A usability study should determine the right ratio between retrieval time duration and precision. These findings are valid for the arousal model as well as for the valence model.

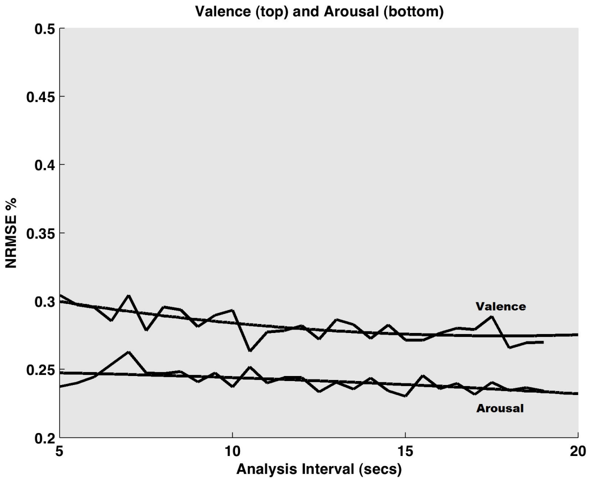


Figure 2.6: NRMSE for varying retrieval/analysis intervals. (For convenience of the reader polynomial regression lines were added.)

The data models for valence and arousal were derived from an experiment where subjects could hear the music. This differs from the typical situation of musical playlist creation, where the user does not hear music while making an arm gesture. Instead, he or she must “think” music. Most people intuitively understand what it means to hear a tune in your head. This can be considered as a form of musical (auditory) imagery. Converging evidence indicates that auditory cortical areas can be activated even in the absence of sound and that this corresponds to phenomenological experience of imagining music [33]. Auditory imagery preserves many structural and temporal properties of auditory stimuli, and generation of auditory imagery appears to involve activation of many brain areas involved in perception of auditory stimuli [34]. We hypothesize that gestures made by subjects to emotionally express the music they hear is triggered by activating these brain areas. As a consequence, we expect arm movements made in absence of music but triggered by musical imagery, will be essentially similar to movements that would have been made when the imagined music would have been physically audible. In particular, we assume here that musical emotion can be transmitted through movement independent of the actual presence of the music.

Additional research is needed to gain insight into the role of musical imagery for our application. One important research question is: What is the impact of arousal and valence on musical imagery ? In a study with words, emotional words were consistently better recalled than the neutral words [35]. Does this also apply to music imagery? Can we more easily imagine music that triggers extreme values for valence and arousal?

2.6 Conclusions and Future Work

The work presented in this paper is intended as a foundation for a motion based affective user interface for music retrieval. We have derived predictive models for valence and arousal from empirical data, gathered in an experimental setup where inertial data recorded from arm movements is coupled to emotion ratings. This experiment firstly extends previous findings that state that human subjects are generally capable of recognizing affect by means of arm gestures to the capability of recognizing affect by means of gestures originated by the mood of a musical fragment. Secondly, model validation in the main study confirmed the predictive capabilities of the model, regressing musical emotion ratings to arm movement. In line with previous studies, we find that arousal is more directly related to arm movement than is the case for valence.

To our knowledge, attempts to detect affect from movement using inertial sensors are as of yet very scarce[3]. The use of inertial sensors for affect recognition has the crucial advantage that such sensors are readily available in mobile devices nowadays, which makes the use of the developed method in commercial applications a viable option.

Several improvements to the models can be made. A first improvement would be an individual calibration of the model. Movement on music is an individual expression. Although our general model works, fine-tuning to individual traits of users may increase its accuracy. Studies revealed indeed that for example gender, age, musical expertise, active musicianship, broadness of taste and familiarity with music have an influence on the semantic description of music [37].

A second improvement would be the use of other more sophisticated statistical models. In this study, we used linear regression models, but more complex models like support vector regression [38] or reservoir computing [39] may achieve higher prediction accuracies.

A last improvement can result from other and/or more sensing devices. The observers in the experiment indicated that the physical location where the arm movement takes place plays an important role for the determination of the valence. Arm movement performed at higher locations are indicators of joy and consequently of high valence values. Since accelerometer data alone are not sufficient to accurately estimate position, additional sensing techniques (e.g. gyroscopic sensing) will be required.

[3] The exception that proves the rule is [36].

References

[1] R.A. Calvo and S. D'Mello. *Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications.* Affective Computing, IEEE Transactions on, 1(1):18–37, jan. 2010 2010.

[2] P.N. Juslin and D. Västfjäll. *Emotional responses to music: The need to consider underlying mechanisms.* Behavioral and Brain Sciences, 31(05):559–575, 2008.

[3] G. Varni, G. Volpe, and A. Camurri. *A system for real-time multimodal analysis of nonverbal affective social interaction in user-centric media.* Multimedia, IEEE Transactions on, 12(6):576–590, 2010.

[4] M. Leman. *Embodied music cognition and mediation technology.* The MIT Press, 2008.

[5] T. Eerola, O. Lartillot, and P. Toiviainen. *Prediction of Multidimensional Emotional Ratings in Music from Audio Using Multivariate Regression Models.* In 10th International Society for Music Information Retrieval Conference, pages 621–626. ISMIR, 2009.

[6] E. Schmidt and Y. Kim. *Projection of Acoustic Features to Continuous Valence-Arousal Mood Labels via Regression.* In 10th International Society for Music Information Retrieval Conference. ISMIR, 2009.

[7] E. Schubert. *Modeling Perceived Emotion With Continuous Musical Features.* Music Perception, 21(4):561–585, 2004.

[8] K. R. Scherer. *Why music does not produce basic emotions: A plea for a new approach to measuring emotional effects of music.* In Proceedings of the Stockholm Music Acoustics Conference 2003, pages 25–28, 2003.

[9] J. Panksepp and G. Bernatzky. *Emotional sounds and the brain: the neuro-affective foundations of musical appreciation.* Behavioural Processes, 60(2):133–155, 2002.

[10] V. J. Konečni, A. Brown, and R. A. Wanic. *Comparative Effects of Musicand Recalled Life-events on Emotional State.* Psychology of Music, 36(3):289–308, 2007.

[11] J.A. Russell. *A circumplex model of affect.* Journal of Personality and Social Psychology, 39(6):1161–1178, 1980.

[12] T. Eerola and J.K. Vuoskoski. *A comparison of the discrete and dimensional models of emotion in music.* Psychology of Music, 39(1):18, 2011.

[13] S.R. Livingstone, R. M
”uhlberger, A.R. Brown, and A. Loch. *Controlling musical emotionality: An affective computational architecture for influencing musical emotions.* Digital Creativity, 18(1):43–53, 2007.

[14] P. Ekman and H. Oster. *Facial expressions of emotion.* Annual review of psychology, 30(1):527–554, 1979.

[15] H.G. Wallbott. *Bodily expression of emotion.* European journal of social psychology, 28(6):879–896, 1998.

[16] C.L. Roether, L. Omlor, A. Christensen, and M.A. Giese. *Critical features for the perception of emotion from gait.* Journal of Vision, 9(6), 2009.

[17] A.P. Atkinson, W.H. Dittrich, A.J. Gemmell, and A.W. Young. *Emotion perception from dynamic and static body expressions in point-light and full-light displays.* Perception, 33:717–746, 2004.

[18] F.E. Pollick, H.M. Paterson, A. Bruderlin, and A.J. Sanford. *Perceiving affect from arm movement.* Cognition, 82(2):B51–B61, 2001.

[19] S. Brownlow, A.R. Dixon, C.A. Egbert, and R.D. Radcliffe. *Perception of movement and dancer characteristics from point-light displays of dance.* The Psychological Record, 47(3), 1997.

[20] A. Camurri, I. Lagerlöf, and G. Volpe. *Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques.* International journal of human-computer studies, 59(1):213–225, 2003.

[21] B.W. Vines, M.M. Wanderley, C.L. Krumhansl, R.L. Nuzzo, and D.J. Levitin. *Performance gestures of musicians: What structural and emotional information do they convey?* Gesture-based communication in human-computer interaction, pages 3887–3887, 2004.

[22] M. Lesaffre. *Music Information Retrieval. Conceptual framework, Annotation and User Behaviour.* Unpublished PhD, 2005.

[23] M. Zentner, D. Grandjean, and K. R. Scherer. *Emotions evoked by the sound of music: characterization, classification, and measurement.* Emotion, 8(4):494, 2008.

[24] JP Marques de Sá. *Applied Statistics: Using SPSS, Statistica, MATLAB, and R.* Springer, 2007.

[25] D. A. Bell and H. Wang. *A formalism for relevance and its application in feature subset selection.* Machine learning, 41(2):175–195, 2000.

[26] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin. *The elements of statistical learning: data mining, inference and prediction.* The Mathematical Intelligencer, 27(2):43–94, 2005.

[27] N. Brace, R. Kemp, and R. Snelgar. *SPSS for Psychologists.* Palgrave, 2003.

[28] E. Bigand, S. Vieillard, F. Madurell, J. Marozeau, and A. Dacquet. *Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts.* Cognition and emotion, 19(8):1113–1139, 2005.

[29] G. Kreutz, U. Ott, D. Teichmann, P. Osawa, and D. Vaitl. *Using music to induce emotions: Influences of musical preference and absorption.* Psychology of music, 36(1):101, 2008.

[30] L. Lu, D. Liu, and H.J. Zhang. *Automatic mood detection and tracking of music audio signals.* Audio, Speech, and Language Processing, IEEE Transactions on, 14(1):5–18, 2005.

[31] Y.H. Yang, Y.C. Lin, H.T. Cheng, and H.H. Chen. *Mr. emo: Music retrieval in the emotion plane.* In Proceeding of the 16th ACM international conference on Multimedia, pages 1003–1004. ACM, 2008.

[32] S. Dahl and A. Friberg. *Visual perception of expressiveness in musicians' body movements.* Music Perception, 24(5):433–454, 2007.

[33] R. J. Zatorre and A.R. Halpern. *Mental Concerts: Musical Imagery and Auditory Cortex.* Neuron, 47(1):9 – 12, 2005.

[34] T.L. Hubbard. *Auditory imagery: empirical findings.* Psychological bulletin, 136(2):302–329, 2010.

[35] R.J. Maddock, A.S. Garrett, and M.H. Buonocore. *Posterior cingulate cortex activation by emotional words: fMRI evidence from a valence decision task.* Human Brain Mapping, 18(1):30–41, 2003.

[36] J. Wagner, E. André, and F. Jung. *Smart sensor integration: A framework for multimodal emotion recognition in real-time.* In Affective Computing and Intelligent Interaction, 2009.

[37] M. Lesaffre, L. De Voogdt, M. Leman, B. De Baets, H. De Meyer, and J.P. Martens. *How potential users of music search and retrieval systems*

describe the semantic quality of music. Journal of the American Society for Information Science and Technology, 59(5):695–707, 2008.

[38] S. Golowich V. Vapnik and A. Smola. *Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing.* In M. Mozer, M. Jordan, and T. Petsche, editors, Neural Information Processing Systems, volume 9. MIT Press, Cambridge, MA., 1997.

[39] D. Verstraeten, B. Schrauwen, M. D'Haene, and D. Stroobandt. *An experimental unification of reservoir computing methods.* Neural Networks, 20(3):391–403, 4 2007.

The world is its own best model.
Rodney A. Brooks, 1990

3 Expressive Body Movement Responses to Music are Coherent, Consistent, and Low Dimensional

Abstract

Embodied music cognition stresses the role of the human body as mediator for the encoding and decoding of musical expression. In the present study we set up a low dimensional functional model that accounts for 70% of the variability in the expressive body movement responses to music. With Functional Principal Component Analysis (FPCA) we modeled individual body movements as a linear combination of a group average and a number of eigenfunctions. The group average and the eigenfunctions are common to all subjects and make up the commonalities. An individual performance is then characterized by a set of weights (the individualities), one per eigenfunction. The model is based on experimental data which finds high levels of coherence/consistency between participants when grouped according to musical education. This shows an ontogenetic effect. Participants without formal musical education focus on the torso for the expression of basic musical structure (tempo). Musically trained participants decode additional structural elements in the music and focus on body parts having more degrees of freedom (such as the hands). Our results confirm earlier studies that different body parts move differently along with the music.

3.1 Introduction

THE power of music as a non-verbal expressive communication system is widely recognized [1–4]. Yet, the mechanisms that support the encoding and decoding of musical expression are still poorly understood, especially in social contexts (e.g., pop concerts, joint action, etc.). Embodied approaches to music have defined the human body and body movements as core aspects of these encoding-decoding mechanisms [1, 5]. In general, body movements are considered to facilitate the non-verbal expression and communication of emotions, feelings, ideas and intentions [6].

In the context of music production, expressive movements can be encoded into sound (e.g. [5, 7–15]), typically through the use of a musical instrument. Accordingly, the structural features inherent to a musical composition (e.g., melodic lines, rhythm, etc.) combined with the expressive performance of a musician (e.g., timing, dynamics, etc.) create, what has been called, "moving sonic forms" (cf. [1]). When listening to music, people can mirror the expressive aspects of moving sonic forms back into actual movement patterns. Synchronization of movement to the musical beat is known to be based on brain regions that associate sounds with motor activity [16]. However, people are also capable of generating smoother body movement patterns that go along with the musical expression [15, 17–19]. These movement patterns can be further connected to other modes with which actions are typically associated, like emotions, situations, and images. By mirroring sound to movement, music can be experienced and understood as intentional, expressive, and semantically meaningful [1].

A successful and effective communication of musical expression requires that human expressive movement responses to music are at least partly coherent and consistent. Therefore, the study of patterns of coherence and consistency in music-evoked body movements is important in order to provide deeper insights into musical signification processes in a social context. In addition, we want to be able to define what is common in the expressive response of a population (commonality), as well as what is different in the expressive responses of the individuals of this population (individuality).

So far, only a small number of studies have addressed coherence/consistency in expressive movement responses to music. Leman et al. [18] used a regression model to study the coherence of listeners' movements in response to Guqin music. It was shown that there was a trend depending on learning. Desmet et al. [20] applied dynamic time warping (DTW) and cross correlation to estimate group coherence of spontaneous movements to music. Based on recurrence analysis, Varni et al. [21] calculated the level of synchronization established among the affective behaviors of each single subject in the group on the basis of a generalized autocorrelation function. Several other studies have tried to capture the coherence/consistency of group movements [22]. However, in several of the recent studies that consider music-related experiments in relation to expressiveness [9, 23], measurements of movement deployment are often reduced to single-value measures, and less attention is devoted to the particular dynamic features of the expressive movement. As a result, the commonality and individuality of the expressive responses of a population have remained hard to define.

3.2 Aim of study

The aim of this study is to develop a dynamic approach to analyze coherence, consistency, commonality and individuality in how people mirror musical expression in their free and spontaneous body movement in response to music.

Because we anticipate that people will mimic the musical expression in their movement, we study movement also in terms of expressiveness. Although we do not define this concept until section 3.4, we assume that the reader's intuitive knowledge about this concept will be sufficient here.

Coherence is a group effect and describes how well individual subjects correlate in terms of expressiveness at two distinct timestamps. Coherent time intervals are then continuous time intervals where we see high coherence for every pair of timestamps. Consistency means high correlations between distant coherent intervals. Commonality stands for what is common in the expressive response of a population and individuality stands for what is different in the expressive responses of the individuals of this population. Analyzing the music in terms of these concepts poses some challenges:

- Firstly, since body movement synchronization to music is a dynamic phenomenon, the method should allow for describing model parameters in terms of basic movement patterns extracted from real movements, rather than in terms of single values that capture a particular feature of a particular movement segment. Such an approach has recently been explored by Leman and Naveda [24] and Fan et al. [25], who use periodicity analysis to capture spatiotemporal representations of gestures. Their approach however, still requires a segmentation step using periodicity (beats) to determine segment boundaries. Our newly proposed method avoids the cumbersome segmentation process and focuses directly on the process dynamics.

- Secondly, it has been shown that people do not synchronize all their body parts to the music at all frequency levels at any given time [26, 27]. We may assume that parts of the human body (such as the hands) are better suited for capturing expressiveness than, say, a leg [28–31]: (i) There is the known privileged role of hand gestures in reasoning and conveying emotion and expressivity [19, 32–34]. (ii) The human body can be modeled as a chain of rigid bodies connected at joints that provide a number of degrees of freedom (DOFs). Chest has few DOFs and hands have many DOFs. (iii) Hands make up the personal space, i.e. the space within reach of the performer's body [35] whereas the chest belongs to the intimate space, i.e. the space occupied by the performer's body.
  All this makes it challenging to find out the role of individual body parts in movement to music. We provide results for all individual body parts, but the focus will be on hands and chest as they contrast for the explanations just mentioned.

- Thirdly, there may be a considerable degree of inter-individual and intra-individual variability among peoples' movements. We believe that it is possible to capture this aspect by extracting a common model from the population. Starting from this model we can then rebuild the individual responses with some additional parameters that characterize each individual. The number of parameters needed for rebuilding the individual expressive responses to music defines the dimensionality of the expression space for the population. We assume that the expression space is low-dimensional because otherwise, populations would have more difficulties in mirroring the musical expression.

To cope with the aforementioned challenges, we developed a statistical method that captures the essential functional features of human expressive movements as a dynamic model. We assumed that coherence and consistency in these movements subsume a low-dimensional model of expressiveness. Such a model would support the theory that expressive embodiment of music has a firm social foundation. The method followed is based on functional data analysis, in particular correlation analysis and functional principal component analysis. It enables us to capture free body movement responses to music as a dynamic phenomenon with a focus on basic overall model parameters. Thereby, we hypothesized that movement *speed* is a sufficient and relevant marker for measuring the coherence/consistency of free human movement responses to music. It makes abstraction of the position and focuses on the rate of change of the movement position. No further assumptions are made about segmentation of gestures. Instead, we believe that the analysis should come up with the segmentations from the bottom-up [36].

The presented method will be validated on a data set originating from an experiment assessing people's free body movement responses to music (i.e., *The First Piano Concerto* of Brahms). The participants of the experiment consisted of a group of musically trained people and a group of musically untrained people. Applying our statistical method, we investigate patterns of coherence/consistency within the groups of participants and investigate the role of individual body parts in how participants synchronize their movements to the music. Accordingly, based on this case study, we demonstrate how our statistical method can contribute to advances in knowledge on social aspects of embodied music cognition. We believe that such an approach may lead to a novel research methodology for movement analysis in relation to music. Moreover, we believe that the method may be useful for other domains (e.g. dance) where coherence/consistency of non-verbal communication and human movement is studied.

3.3 Experimental set-up

3.3.1 Participants

Distinct groups of participants were formed on the basis of their musical background. A first group (i.e., musically trained MTr group) was composed of 18 participants (10 male, 8 female) with a mean age of 23.83 years (SD=3.71). A second group (i.e., musically untrained MunTr group) was composed of 18 participants (10 male, 8 female) with a mean age of 24.60 (SD=4.81). Subjects from the MTr group declared to have had music education with a mean number of years of 9.7 (SD=5.4) and declared to play a musical instrument. Musical education refers here to additional lessons compared to the obligatory courses in primary and secondary school that are taught from the age of 6 till 14 (1 hour/week). The additional lessons comprise solfège, instrument learning, ensemble playing and take up more than 3 hours a week. The MunTr group declared to have received only the obligatory courses.

A Big Five Inventory test [37] revealed no significant difference between the rate of extraversion/introversion between the MTr group (M=3.49, SD=0.62) and the MunTr group (M=3.51, SD=0.67), t(34)$= -0.099$, p$>$.05 (p=.92).

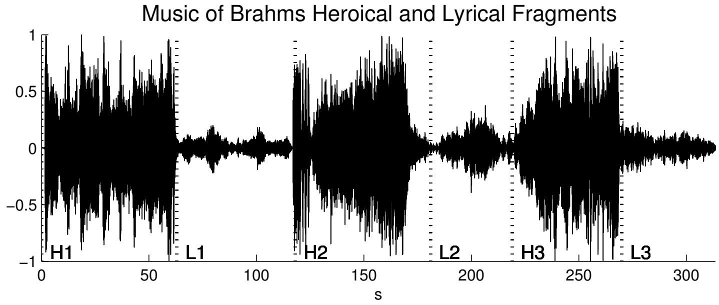


Figure 3.1: Musical excerpt: Johannes Brahms' First Piano Concerto, Opus 15 in D minor. Dashed lines separate Heroic (Hx) and Lyric style (Lx) intervals.

Moreover, we asked how familiar participants were with the type of music used in the experiment. This question was rated on a five-point Likert scale, with 1 as not at all familiar, 3 as somewhat familiar, and 5 as extremely familiar. Familiarity for the MTr group (Mdn=4) was significantly higher than for the MunTr group (Mdn=2), U=0, z=−5.45, p<.001, r=−.91.

Participants also filled in a questionnaire based on the semantic differential method [38] to test the emotional experience of musical fragments (e.g., [39–44]). No differentiation in responses was found related to the musical background of the participants. This finding is in line with studies of Bigand et al. [45, 46] indicating that emotional responses to music are stable and only weakly influenced by musical expertise.

3.3.2 Musical stimulus

The musical stimulus is based on the first 6 minutes and 10 seconds of the Maestoso movement of Johannes Brahms' First Piano Concerto, Opus 15 in D minor from 1858 (in a recording by Krystian Zimmerman and the Berlin Philharmonic Orchestra, conducted by Simon Rattle). The musical piece is characterized by passages that articulate extreme contrasts in physical acoustic energy, reflecting two contrasting expressions, namely a Heroic and Lyric expression. In the stimulus three Heroic passages were alternately presented with three Lyric passages. Because the first Lyric passage is relatively long in comparison with the other we deleted some portion of that passage (1 min 56 s - 2 min 46 s of the recording) in a way that was not audible for people that do not know the musical piece well. The remaining musical stimulus had a duration of approximately 5 min (See Fig. 3.1).

The contrasts between the Heroic and Lyric passages were checked on the basis of a (psycho-)acoustical analysis. The properties that were extracted from the audio signal encompassed an energy property (i.e., amplitude), a

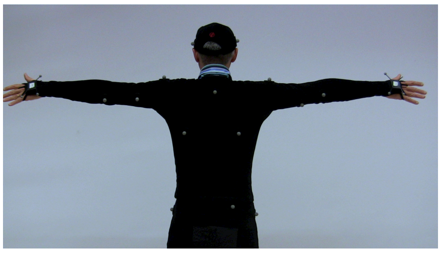


Figure 3.2: Marker settings for Motion Capture System.

rhythm property (i.e., onset likelihood), and spectrum properties (i.e., irregularity, spectral flatness, spectral kurtosis, spectral sharpness, spectral variance) (See [47] for further details). To statistically test the differences between the heroic and lyric fragments on the various acoustic properties, we applied nonparametric Mann-Whitney U tests (normality was violated). In summary, the results show that the levels of amplitude, onset likelihood, irregularity, spectral sharpness, and spectral variance were significantly higher ($p < .001$) in the heroic fragments. In contrast, the levels of spectral flatness and spectral kurtosis were significantly higher ($p < .001$) in the lyric fragments.

3.3.3 Procedure

Participants were invited *individually* to take part in an experiment where their movements were recorded. The participants received the task of moving spontaneously to the music. This was formulated as: "Translate your experience of the music into free full-body movement. Try to become absorbed by the music that is presented and express your feelings into body movement. There is no good or wrong way of doing it. Just perform what comes up in you." They could thereby use the space indicated by a round carpet with a diameter of 4 meters. Furthermore, we made the room completely dark, as the pilot study had indicated that this made the participants more comfortable and less constrained to execute their task.

3.3.4 Measurement

The data for this experiment was collected by a Motion Capture System (Optitrack infrared optical system with Arena motion capture software) at a sampling rate of 100 Hz. Markers were attached to the upper body (hips-chest-neck-head-collarbones-shoulders-elbows-wrists) (See Fig. 3.2).

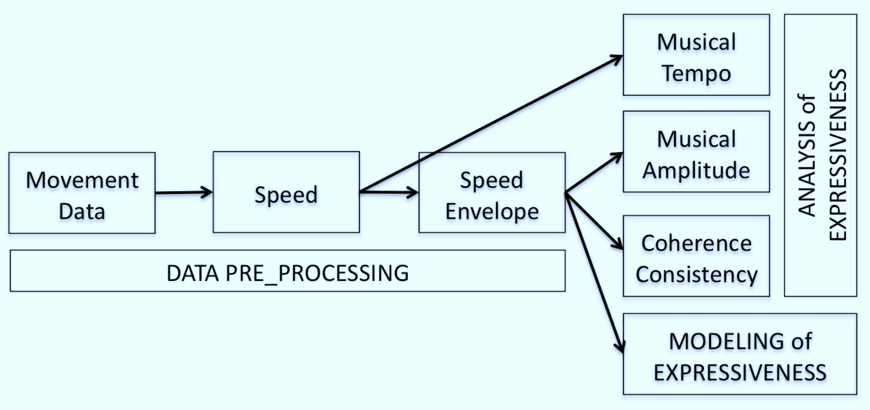


Figure 3.3: From movement data we calculate a speed signal and run this through a low pass filter to obtain the speed envelope. Analyses are done with reference to musical tempo and musical amplitude. A third analysis correlates speed at different timestamps which leads us to the concepts of coherence and consistency. The speed envelope is also used as input for the modeling process.

3.3.5 Block-diagram

Fig. 3.3 shows the structure of this chapter using data processing as the leading thread. For our data pre-processing steps we used the movement data to calculate a speed signal and ran this through a low pass filter to obtain the speed envelope. Subsequent analyses use these signals as input and have musical tempo and musical amplitude as points of reference. A third analysis correlates speed at different time stamps which leads us to the concepts of coherence and consistency. Eventually the speed envelope is used as input for the modeling process.

3.4 Analysis of Expressiveness

Music psychology pioneer Carl Seashore [48] introduced the idea that expressive performance consists of deviation from the regular in sound properties such as loudness, tempo (rubato), articulation and intonation. Davidson [8] adds the idea of embodiment to it saying that each movement type (for instance, the wiggle) can be executed in various ways giving the potential for a range of expressivity levels to be elicited. Their findings concern music performances but similar results are found in work on dance performances. For example, Camurri et al. [49] studied expressive gestures as gestures superimposing expressive content (deviation) to normal gestures (regular).

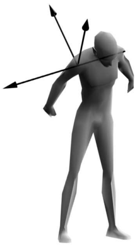


Figure 3.4: New axis system used for the analysis of elbows and hands(wrists) movement.

However, the terms "deviation" and "regular" may have negative connotations. Regular stands for normal behavior and deviation for abnormal behavior. This is the connotation we want to avoid. We do not want an association between music-evoked body movement and abnormal behavior. Therefore, we will use the terms variation and reference instead. As points of reference we consider various social norms. In Section 3.4.2 the reference is defined as the tempo of the music while in Section 3.4.3 we use the musical amplitude as reference. Tempo and amplitude are examples of conventional social norms describing normal, anticipated behavior. In section 3.4.4, we define the reference as being the group average of the log of the speed of the movement.

Let us start, however, by explaining how to pre-process the data.

3.4.1 Pre-processing of the data

Data collected from the markers of the Motion Capture System are converted to Cartesian coordinates referencing a fixed axis system with the origin located on a fixed spot on the floor. In such a system movement of the torso (translations and rotations) has influence on the speed of connected limbs like arms. To eliminate this influence a new axis system was defined for the analysis of the elbow and wrist markers. The origin of this new axis system was placed on the shoulder as shown on Fig. 3.4. Axis 1 was defined as the line going through the clavicle (shoulder-neck). Axis 2 was defined by the projection of the up-position (chest-neck) onto a plane perpendicular to axis 1 and eventually axis 3 was determined as the cross product of axis 1 and axis 2. All calculations and findings in this paper use this coordinate system for hands and elbows unless otherwise stated.

Samples were collected for 315 seconds corresponding to the duration of the musical fragment. The data from one subject in the MTr group were discarded due to technical problems during the recording. Bodily movements were free (no imposed choreography) and resulted in individual movements that were not directly comparable from subject to subject. Therefore, we decided not to compare the positional coordinates. Instead, we compared the speed of the bodily movements with as underlying motivation that speed is closely related to kinetic energy ($\boldsymbol{E_k = \frac{1}{2}mv^2}$ for every body part following Dempster's human body model [50]). But even then timing differences among subjects hindered the analysis. We solved this by (low-pass) filtering the speed signal with filter parameters set in line with musical characteristics (see below). The selected music of Brahms has measures varying from 1.6 to 1.7 s and the time interval for filtering was set to 5 s approximating three measures. The 5 s window was also checked with respect to Pöppel's theory of the 3 s window of temporal integration [51]. Using a 5 s window made sure that we captured this phenomenon.

Ultimately, the calculation of the speed signal was handled in two steps. Firstly, the speed signal was calculated from the Motion Capture positional coordinates and secondly it was low-pass filtered.

The first step, calculation of the speed signal, used a linear regression based *derivation* filter with a regression window of 0.175 s. The value of 0.175 s corresponds with a linear response of the derivation filter in the useful frequency band of 0-4 Hz. The useful frequency band was derived from spectrograms (an example for subject 1 is shown in Fig. 3.5). This derivation filter was applied to all coordinates of all subjects. Eventually the speed signal was set equal to the L2-norm of the derivatives. The regression filters used are identical to first order Savitzky-Golay smoothing filters [52]. Our motivation for using regression filters is that features of the temporal speed pattern, such as minima, maxima and slopes, are better preserved by filters having a clear temporal interpretation than by filters having a clear spectral interpretation.

In the second step we calculated from this speed signal a slowly varying speed signal (the speed envelope) using a moving average filter with a window of 5 s as explained above. The resulting speed envelope is presented in Fig. 3.6.

3.4.2 Motor-mimesis of musical tempo

Here, we investigate in what ways the periodicity of expressive movement responses to music equals or differs from the tempo of the music. The tempo of the music is used as reference here. It is calculated from the manually annotated inter-beat time intervals for the different musical style fragments. This is discussed in Subsection 3.4.2.1. To analyze periodicity in the bodily movement, the dominant frequencies are extracted by means of Fourier analysis (Subsection 3.4.2.2). In Subsection 3.4.2.3 we compare

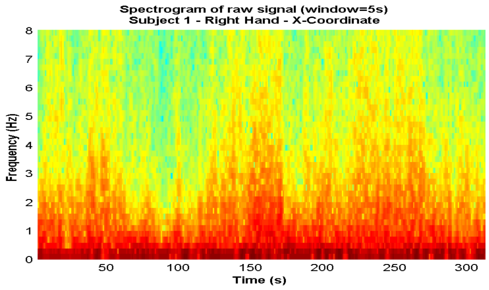


Figure 3.5: Spectrogram for the x-coordinate of the right hand (subject 1)

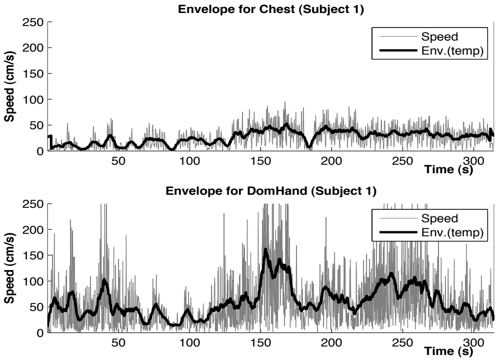


Figure 3.6: Speed and Speed Envelope for subject 1. Speed envelope is the result of running the speed signal through a moving average filter with a window of 5 s.

musical tempo with periodicity in movement.

3.4.2.1 Manual analysis of the musical signal

The time signature of the music is 6/4 with an emphasis on the first and fourth beat. The beat times of the musical piece were manually annotated using Audacity [53]. From the annotation we learn that the inter-beat time intervals slightly vary during the musical piece. In order to better estimate the varying tempo, we split the entire musical excerpt into six style fragments labeled respectively as Heroic 1-2-3 and Lyric 1-2-3. The inter-beat time intervals are considerably longer in the lyric fragments compared to the heroic fragments, pointing to a lower tempo. In addition, owing to the musical articulation, the inter-beat time intervals are not equally spaced resulting in a different time gap from beat 1 to beat 4 in the same measure compared to the time gap from beat 4 to beat 1 of the next measure (See Table 3.1). When the beat 1-4 time interval differs significantly from the 4-1 interval, one can expect a frequency peak at the "measure"-frequency compared to a peak at the double "measure"-frequency otherwise.

3.4.2.2 Fourier Analysis of the Speed Signal

Because the tempi differ from style fragment to style fragment, we do not make a Fourier analysis on the entire time interval of the speed movement signal but on the different style fragments. Given our interest in differences between the two groups, we apply a Fourier analysis twice, once for the MTr group and once for the MunTr group.

The Fourier analysis is done in two stages. Firstly, we analyze every individual subject (and every marker) and secondly we consolidate these results to end with one spectrum per group. The consolidation process consists of a normalization step followed by an averaging step. The normalization step is required to compensate for subjects performing at different speed levels.

For all groups and most markers we observe dominant frequency peaks in the range from 0.68 to 0.87 Hz (±0.02 Hz) matching a period from 1.47 to 1.15 s. These periods indicate a half measure and correspond to tempi between 122 and 157 BPM. To illustrate we show the spectra for chest and dominant hand in the second heroic style fragment (H2) for both groups (Fig 3.7 and Fig 3.8).

The coordinates of the dominant hand are relative to the shoulder and are as such not influenced by movement of the torso, as previously explained (Section 3.4.1). Still, a peak frequency corresponding to the tempo is clearly visible. (The frequency resolution depends on the number of samples in a fragment and is approximately equal to 0.02 Hz).

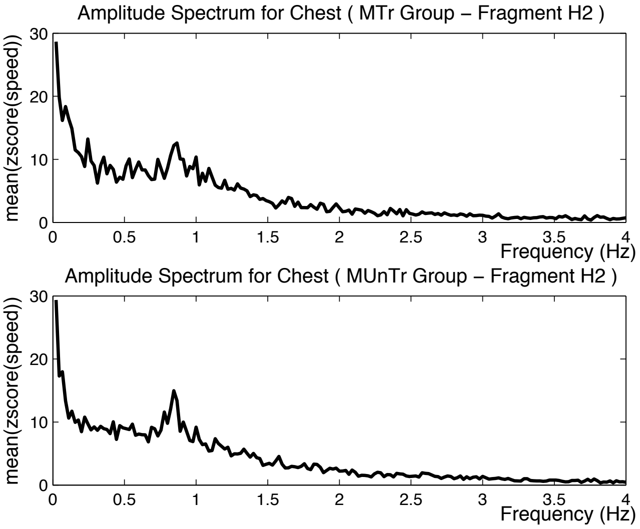


Figure 3.7: Amplitude Spectrum of Chest Speed for Heroic Style Fragment 2 (H2)

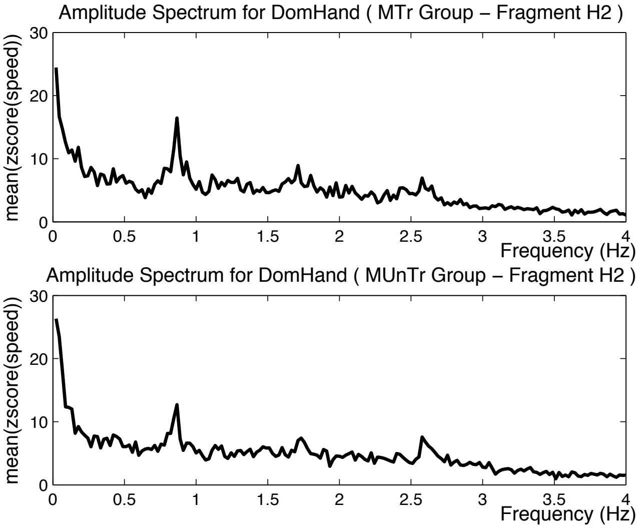


Figure 3.8: Amplitude Spectrum of Dominant Hand for Heroic Style Fragment 2 (H2)

Fragm.	**Annotated Beat 1-4 Interval**	**Annotated Beat 4-1 Interval**	**Fourrier Chest PeakPeriod**	**Fourrier Dom Hand PeakPeriod**
		Musically Trained Group		
$\boldsymbol{H1}$	$\mathbf{1.19 \pm 0.06}$	$\mathbf{1.23 \pm 0.07}$	**1.23**	**1.22**
$\boldsymbol{L1}$	$\mathbf{1.41 \pm 0.12}$	$\mathbf{1.52 \pm 0.07}$	**1.47(1.39)**	−−
$\boldsymbol{H2}$	$\mathbf{1.16 \pm 0.04}$	$\mathbf{1.18 \pm 0.05}$	**1.15**	**1.15**
$\boldsymbol{L2}$	$\mathbf{1.32 \pm 0.06}$	$\mathbf{1.37 \pm 0.15}$	**1.37**	**1.37**
$\boldsymbol{H3}$	$\mathbf{1.20 \pm 0.05}$	$\mathbf{1.22 \pm 0.07}$	**1.22**	**1.15(1.22)**
$\boldsymbol{L3}$	$\mathbf{1.25 \pm 0.05}$	$\mathbf{1.28 \pm 0.06}$	**1.29**	**1.29**
		Musically Untrained Group		
$\boldsymbol{H1}$	$\mathbf{1.19 \pm 0.06}$	$\mathbf{1.23 \pm 0.07}$	**1.22**	**1.22**
$\boldsymbol{L1}$	$\mathbf{1.41 \pm 0.12}$	$\mathbf{1.52 \pm 0.07}$	**1.35(2.00)**	
$\boldsymbol{H2}$	$\mathbf{1.16 \pm 0.04}$	$\mathbf{1.18 \pm 0.05}$	**1.18**	**1.15**
$\boldsymbol{L2}$	$\mathbf{1.32 \pm 0.06}$	$\mathbf{1.37 \pm 0.15}$	**1.37**	−−
$\boldsymbol{H3}$	$\mathbf{1.20 \pm 0.05}$	$\mathbf{1.22 \pm 0.07}$	**1.41(1.15)**	**1.18**
$\boldsymbol{L3}$	$\mathbf{1.25 \pm 0.05}$	$\mathbf{1.28 \pm 0.06}$	**1.29**	**1.29**

Table 3.1: Annotated Beat Interval versus Measured (Fourier) Peak Period expressed in seconds. The values inside brackets () represent a second, slightly smaller peak.

3.4.2.3 Musical Tempo versus Speed Spectrum

In Table 3.1 we compare for both groups the annotated beat-intervals with the dominant peak frequency found in the spectra from chest and dominant hand. We see that in general the measured (Fourier) peak frequency (expressed as a period) corresponds with the annotated beat intervals of the music. Note that the annotated beat 1-4 interval and the annotated beat 4-1 interval sum up into one measure. If the two intervals are similar in duration, we discover a clear peak at a half measure. For intervals having an unequal duration (for example the first lyric style fragment L1) the situation is less clear and the link with the musical tempo is blurred.

Overall, the above shows that the movements mimic the tempo of the musical signal. The tempo followed is the half measure tempo and not the full measure tempo which can be linked to the fact that the half measure tempo is closer to the natural resonance frequency (2 Hz or 120 BPM) of the human body [54]. Further, as a qualitative finding we notice clear peaks for the chest of the MunTr group in all fragments (except for the first lyric style fragment L1). For the MTr group, on the contrary, the peaks are less distinguishable as exemplified in Fig 3.7. The dominant hand shows a different picture. Here we discover clear peaks for both groups in all fragments (except now for the first and second lyric style fragment). This is illustrated in Fig 3.8 for heroic fragment 2. This finding hints at differences in expressiveness between chest and dominant hand and between the two groups.

3.4.3 Motor mimesis of musical amplitude

In this section we use another acoustical feature, namely the musical amplitude (an energy property) as a reference. Initially, we inspect the mean and standard deviation functions of movement (Section 3.4.3.1), followed by the coefficient of variation (Section 3.4.3.2). Further on we correlate movement with our reference, the musical amplitude (Section 3.4.3.3).

3.4.3.1 Mean and Standard deviation of speed amplitude

Here we look at the variation between the speed signals of the motor-mimetic responses and investigate if these reflect structural properties of the music. The speed signals used in this and subsequent sections are the envelope signals as calculated in Section 3.4.1. The mean function is calculated not as an average over time but as an average over subjects. Otherwise stated: we calculate a mean value over subjects at every distinct timestamp. The same applies for the standard deviation function. Both functions are displayed in Fig. 3.9 (MTr group) and Fig. 3.10 (MunTr group).

We see that for the dominant hand the ratio between speed in the heroic parts and speed in the lyric parts largely exceeds the ratio between their tempi and this is confirmed by Fig. 3.11. For the chest there is no noticeable difference. The ratio between the tempi is calculated from Table 3.1. This indicates that the dominant hands show more expressiveness in the heroic parts. So, apart from differences between subjects we find that expressiveness is also linked to body parts and musical style. Additionally, note also the variation in the speed ratio: for the Musically trained group (Mtr) the largest variation is in the hands, for the Musically untrained group (MunTr), it is in the chest. This explains where both groups put their focus on for expressiveness.

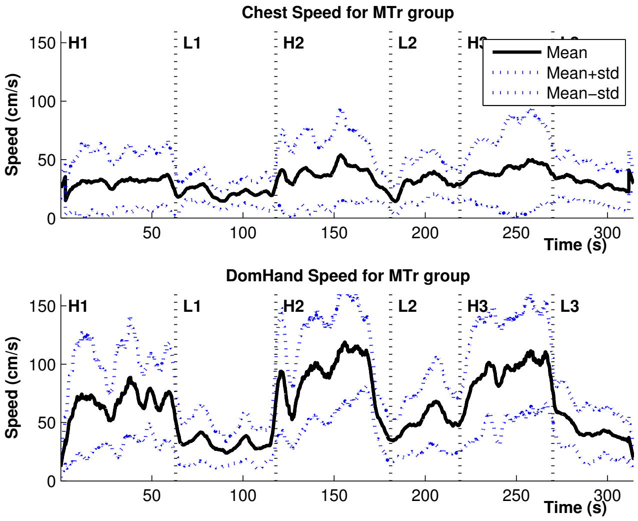


Figure 3.9: Mean and Standard Deviation of Speed for the musically trained group (MTr group)

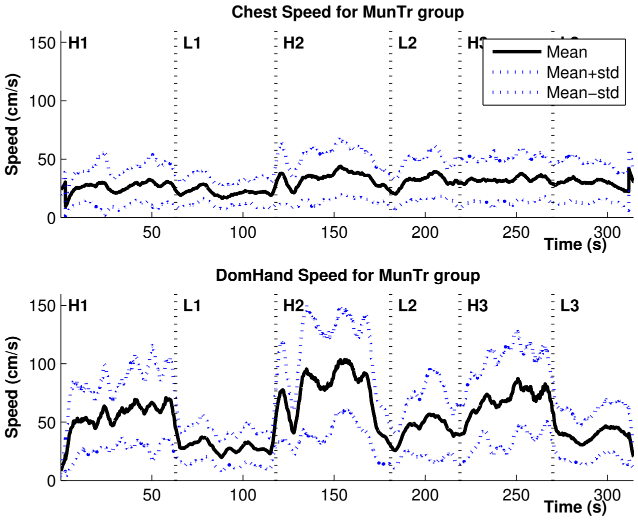
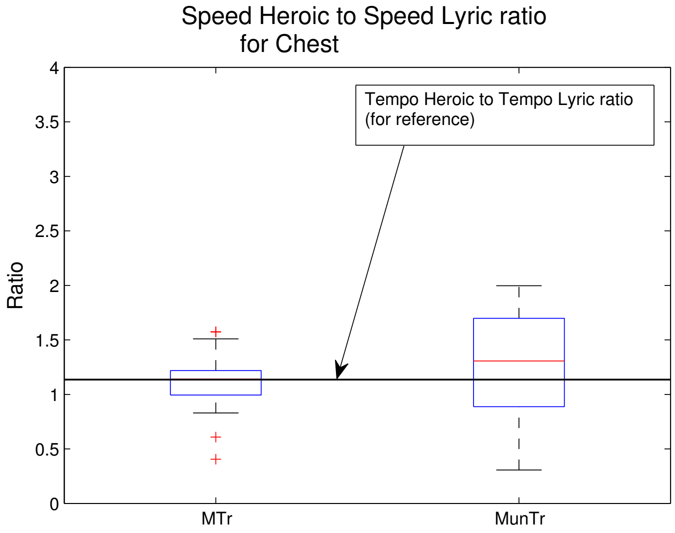


Figure 3.10: Mean and Standard Deviation of Speed for the musically untrained group (MunTr group)

3.4.3.2 Coefficient of Variation

At first glance, the ratio between standard deviation and mean appears to be constant. This ratio is defined as the coefficient of variation (CV). To esti-



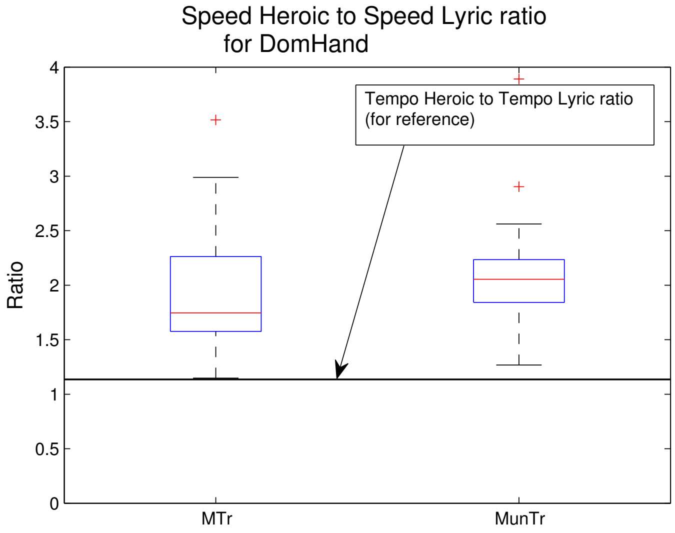


Figure 3.11: Ratio between the speed in heroic style parts compared to the speed in lyric style parts for the 35 subjects splitted up per group (Musically Trained (MTr) and Musically Untrained (MunTr)). The ratio between the tempi of heroic and lyric parts is added as reference. For the dominant hand the heroic style is more expressive as the speed ratio largely exceeds the tempi ratio. For the chest there is no noticeable difference between the two ratios.

mate the CV one could use the ratio of the sample standard deviation to the sample mean. However, this is not good practice since estimation depends highly on the distribution for small to moderately sized samples [55–58]. Therefore we investigate the distribution of the speed values across subjects at all moments in time. We do this by fitting a Weibull distribution. The shape parameter of the fitted Weibull distribution parameter discloses the underlying distribution. Values in the range of [1.25-2.75] hint at a lognormal distribution whereas values between [3-4] suggest a normal distribution [59, 60]. In our case, averaged over all timestamps the Weibull shape parameters for Chest and for Dominant Hand are **1.73 ± 0.38** and **2.08 ± 0.27** respectively. This points to a lognormal distribution as the best fitting distribution for the chest and dominant hand movements. For a lognormal distribution the estimation of the CV is given by equation (3.1) with $\boldsymbol{s_{log}^2}$ the sample standard deviation after a natural log transformation [55–58].

$$\boldsymbol{\hat{CV} = \sqrt{e^{s_{log}^2} - 1}} \tag{3.1}$$

The estimated CV is now presented per group and per marker in Fig. 3.12. The CV for the dominant hand is nearly constant or tends to just slightly decrease over the whole timeframe. The CV for Chest does not show this constant behavior. We see sudden CV increases at the start of the heroic style intervals. The CV increase is due to a proportionally higher increase in the standard deviation and points to a sudden variation in expressivity amongst the subjects. After that the CV slopes down to end up quite constant during the lyric style intervals. The negative slope within a style interval could be explained by a learning effect. Or, otherwise stated, the variation in expressiveness fades away together with the surprising character of the music.

3.4.3.3 Correlation between movement and musical amplitude

The speed envelope signals for the dominant hand and to a lesser extent for the chest show intervals of higher amplitude alternated with intervals of lower amplitude (Fig. 3.9 and Fig. 3.10). Based on a rough visual inspection, these intervals seem to correspond with the heroic and lyric fragments, respectively. As the heroic intervals coincide with high amplitude in the music, it is worthwhile investigating the correlation between the speed envelope signal, as an indicator of expressiveness, and the amplitude of the music. As we learned in section 3.4.3.2, the speed envelope signal is not normally but lognormally distributed. Therefore we correlate the logSpeed with the logAmplitude of the music.

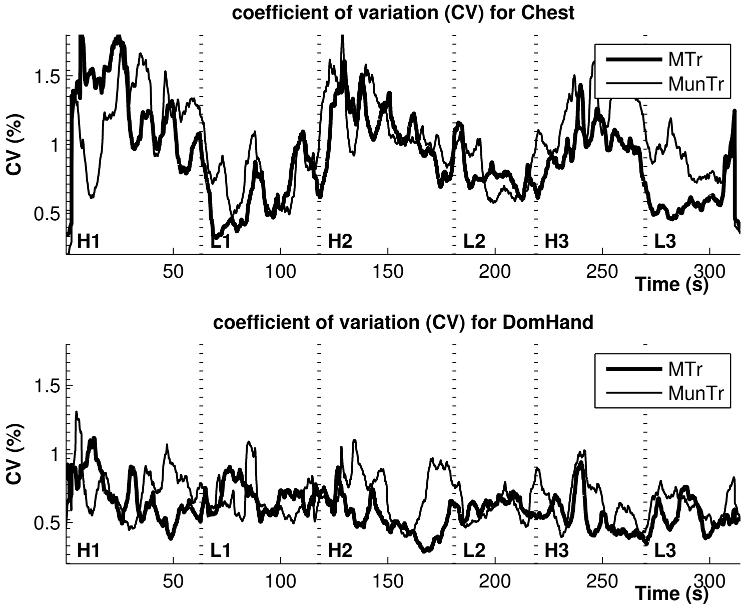


Figure 3.12: Coefficient of Variation for Chest (top) and Dominant Hand (bottom)

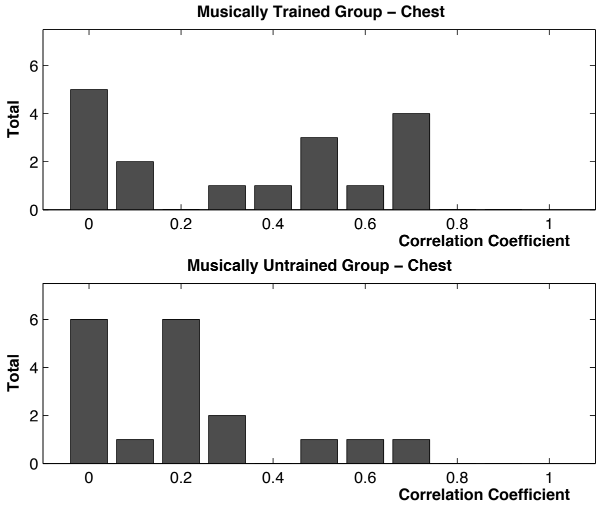


Figure 3.13: Histogram of correlation coefficients. The correlation coefficients are calculated per subject between the log(Music Amplitude) and the log(Chest Speed)

To calculate the logAmplitude of the music signal we use the same method as for the logSpeed. The absolute value of the music amplitude is run through a moving average filter with the same window size (5 s) as used for the calculation of the speed signal.

The logAmplitude is then correlated with the logSpeed signals from all subjects. A Wilcoxon rank sum test ($\boldsymbol{p < .05}$) confirms that the correlation coefficients for the chest (Fig. 3.13) are lower than those for the dominant hand (Fig. 3.14). Further on, the distribution for chest seems to point to an individual difference in chest use with some subjects correlating and some that choose not to do so. This is different from the dominant hand where all seem to correlate to some degree. Here, the distributions are concentrated around high values (0.6-0.7). Additionally, the correlation coefficient values are higher for the MTr group than for the MunTr group. A Wilcoxon rank sum test indicates a marginally significant difference between the two groups (Mdn=0.702, Mdn=0.578, p=.08). So, the dominant hand correlates to the musical amplitude (our reference in this section) and this phenomenon is more apparent with subjects that are musically trained.

3.4.4 Motor-mimesis of syntax and semantics in music

The Pearson correlations between logSpeed over participants at times $\boldsymbol{t_1}$ and $\boldsymbol{t_2}$ yield the entries of a correlation matrix PC($\boldsymbol{t_1}$,$\boldsymbol{t_2}$). We use logSpeed

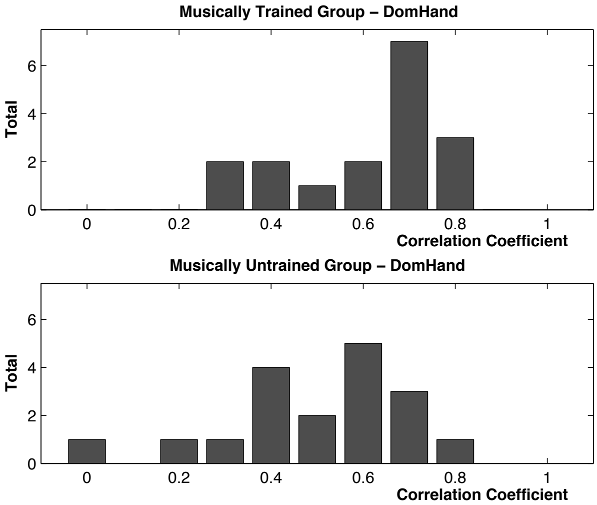


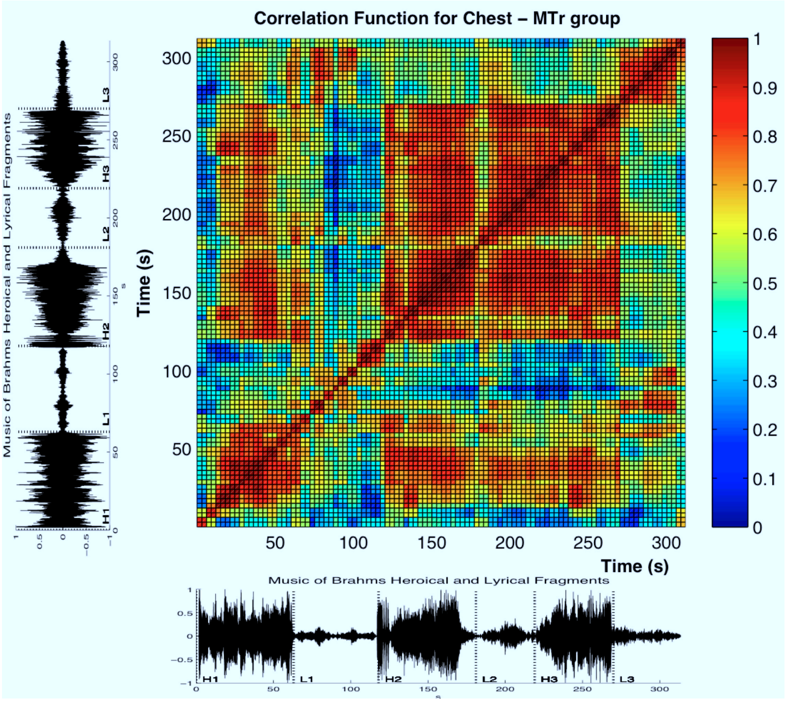
Figure 3.14: Histogram of correlation coefficients. The correlation coefficients are calculated per subject between the log(Music Amplitude) and the log(Dominant Hand Speed)

(log of the speed envelope signal) due to reasons of normality as discussed in Section 3.4.3. Furthermore, because of hypothesis testing, we run the analysis twice, once for the MTr group and once for the MunTr group. Using a correlation matrix implies that the group's mean movement is considered as the reference.

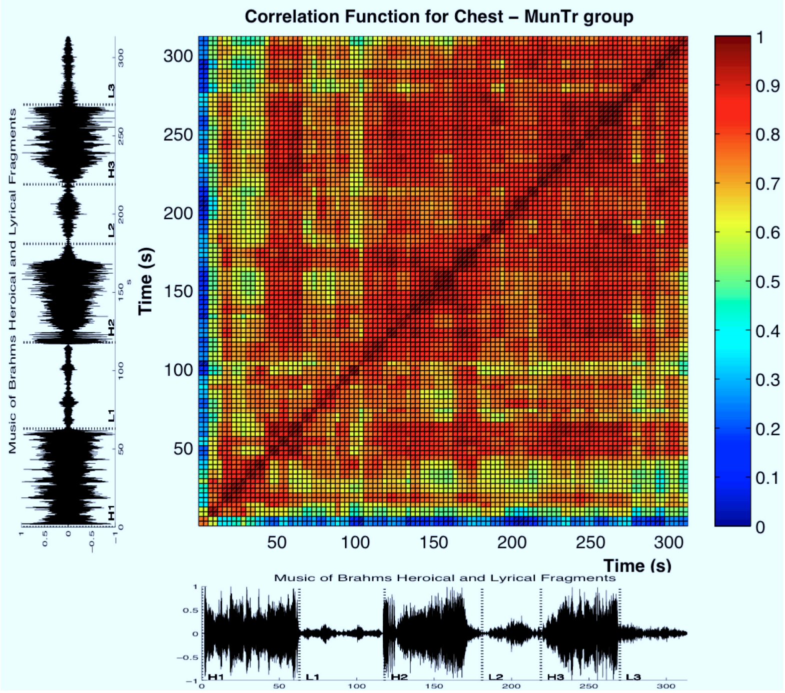
3.4.4.1 Analysis of Chest movement

The correlation matrix PC for chest is displayed as a density plot, once for the MTr group (Fig. 3.15(a)) and once for the MunTr group (Fig. 3.15(b)). The shown density plot is not the original full resolution plot but a low resolution version (this because of displaying and printing reasons). The resolution lowering was achieved by local averaging (over a 5 s surrounding square) and down-sampling to end with a 100x100 matrix. The same procedure is applied to all correlation density diagrams in this article. Results and discussions are however based on the original high resolution plots.

In a correlation density diagram, the diagonal running from lower left to upper right contains the unit values that are the correlation between identical or very close time values [61]. Directions perpendicular to this ridge of unit correlation indicate how rapidly the correlation falls off as the two timestamps separate. This finding is at the source of our definition for



(a) Musically Trained (MTr) Group.



(b) Musically Untrained (MunTr) Group.

Figure 3.15: Correlation Density Plots for Chest.
(Next to the time line is the musical signal shown as in Fig. 3.1)

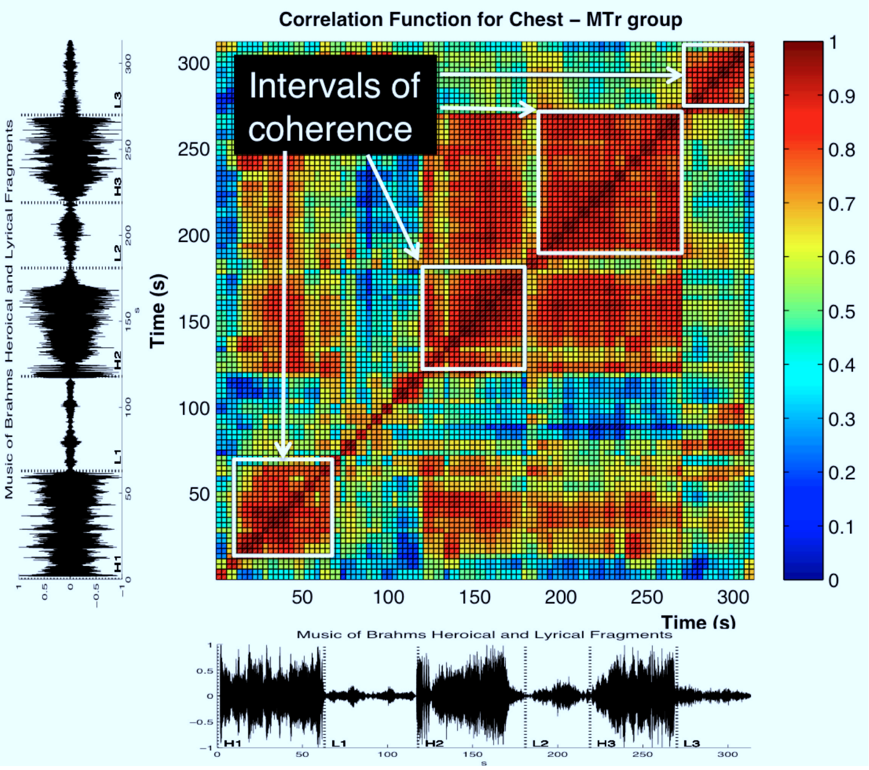


Figure 3.16: Correlation Density Plot for Chest with Intervals of Coherence for the Musically Trained (MTr) Group. The intervals of coherence stand for high correlations between levels of expressiveness in a group at every two distinct timestamps in a continuous time interval. The borders of these intervals can be determined by a user customized heuristic.

coherence. Coherence is identified by “squares of high correlation” along the diagonal (Fig. 3.16). Coherence stands for high correlations between levels of expressiveness (here, logSpeed is used) in a group at every two distinct timestamps belonging to a continuous time interval. It implies (i) synchronicity between subjects within this continuous time interval and (ii) preserving the ordering of subjects in levels of expressiveness (users with high expressiveness keep high expressiveness over the whole interval).

If required, the borders of these coherent intervals (the fall-off points) can be determined by a user customized heuristic: To determine these fall-off points we propose the following heuristic: for every timestamp we calculate the largest square centered on the diagonal where 90% of the correlation coefficients are above a value of 0.60. To locate the fall-off points we plot

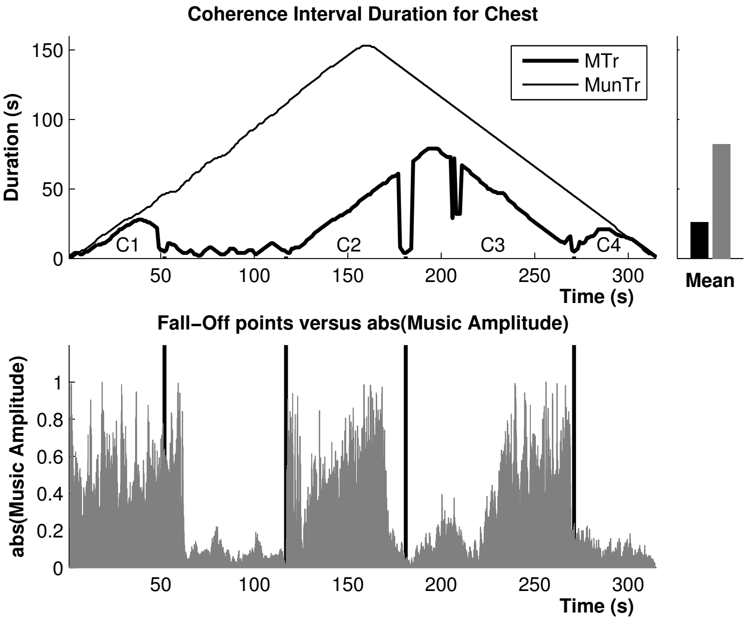


Figure 3.17: Chest Coherence : Correlation FallOff

then the size of these largest squares versus time (actually we plot the half of the length of a side). High peaks point to the middle of a period of high coherence and the height of the peak tells exactly how many seconds this period extends to the left and to the right. Low values correspond with drops in coherence and help in determining the fall-off points. For the chest we see in Fig. 3.17 that the MunTr group shows coherent behavior over the whole duration of the musical fragment. For the MTr group however, we distinguish four intervals of coherence (labeled C1,C2,C3 and C4). The most prominent fall-off points are displayed versus the music amplitude at the bottom of Fig. 3.17. It is clear that there is a certain match with the structural elements in the music like transitions of style (heroic-lyric).

Beside coherence we notice “squares of high correlation” in off-diagonal areas for the MTr group. This leads us to define consistency. Consistency is the phenomenon of having high correlation between remote coherent time intervals (Fig. 3.18).

Using the concepts of coherence and consistency, we notice an apparent difference between the MTr group and the MunTr group (Fig. 3.15). Coherence (of chest) for the MTr group is in line with the musical style (heroic/lyric) elements and this is less the case for the MunTr group. The

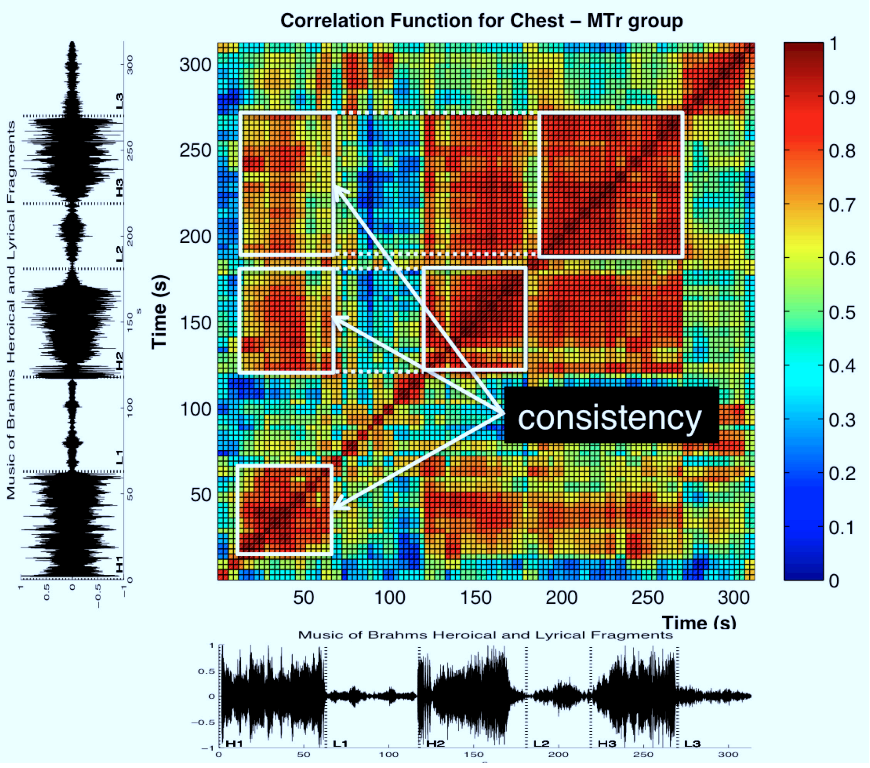


Figure 3.18: Correlation Density Plot for Chest illustrating some consistent intervals for the Musically Trained (MTr) Group. Consistency is the phenomenon of having high correlation between remote coherent time intervals.

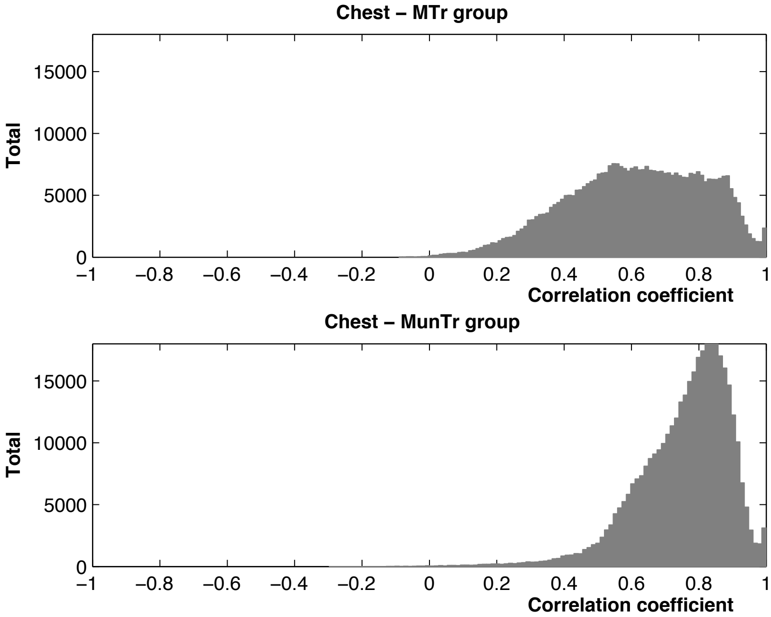


Figure 3.19: Histograms of correlation coefficients (Chest)

MunTr group keeps coherence over almost the whole musical fragment.

Consistency refers to coherent intervals in its definition and this dependency between consistency and coherence makes it difficult to compare performances from different groups. Therefore we introduce a new measure that combines coherence and consistency in one single value. We lose, in this case, the view on the individual components (coherence and consistency) but the advantage is that the new measure allows for comparison of group performances. The combined measure originates from an alternative view on the correlation density diagram: we look at it as being a histogram of correlation coefficients even if the link with time is lost in this representation. We display these results in Fig. 3.19 for the two conditions: the results for the MTr group at the top and the results for the MunTr group at the bottom. Here we see that the correlation coefficients for the MTr group (top) are indeed lower on average. A Wilcoxon rank sum test confirms that the correlation coefficients for the MTr group (Mdn=0.622) are significantly ($p<.001$) different from the MunTr group (Mdn=0.782). This convinces us to use the median of these distributions as a combined measure to compare performances from different groups.

At this stage we have only two groups, hence we can not make further traditional statistical inferences. However permutation tests for comparing two populations [62] offer an alternative. In our case we have 17 subjects in the MTr group and 18 subjects in the MunTr group. We can now make permutations of subjects between the two groups and assume that there is no group effect if none of these permutations yield an effect that is much larger than the one found for the initial group division. In our setting we calculate the effect between the two groups as the difference between the two medians (of the histograms of correlation coefficients). We permute then the 35 (17+18) subjects between the two groups (MTr and MunTr) so that there are always 17 subjects in the MTr group and 18 in the MunTr group. The number of permutations is given by formula (3.2). With a number that exceeds $\mathbf{10^9}$, handling all permutations is not feasible and therefore we limit ourselves to 2000 random permutations. For the permutation test we compute the significance (p-value) as the proportion of differences that are greater than our initial observation. The calculated p-value is .14, not low enough to talk about a significant effect ($> \mathbf{5}\%$) and the null hypothesis would not be rejected by a frequentist approach. However, what this value says is that if there was no difference between the two groups (null hypothesis) there is 14% chance that based on our test we would wrongly reject the null hypothesis (=making a type I error) and this value although not significant is worthwhile mentioning [63]. The 14% value is graphically depicted in Fig. 3.20 where the black bars indicate the values we would wrongly reject by making the null hypothesis. The red line indicates the decision border and is set to the difference between the two groups in their original composition.

$$\begin{pmatrix} \mathbf{35} \\ \mathbf{17} \end{pmatrix} = \frac{\mathbf{(35)!}}{\mathbf{18!\ 17!}} \tag{3.2}$$

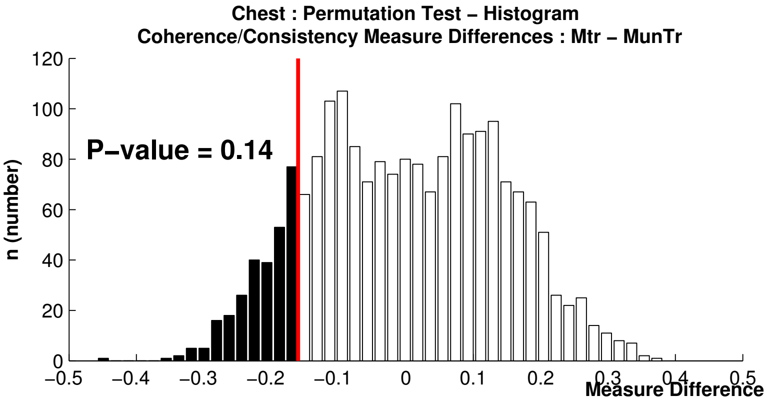


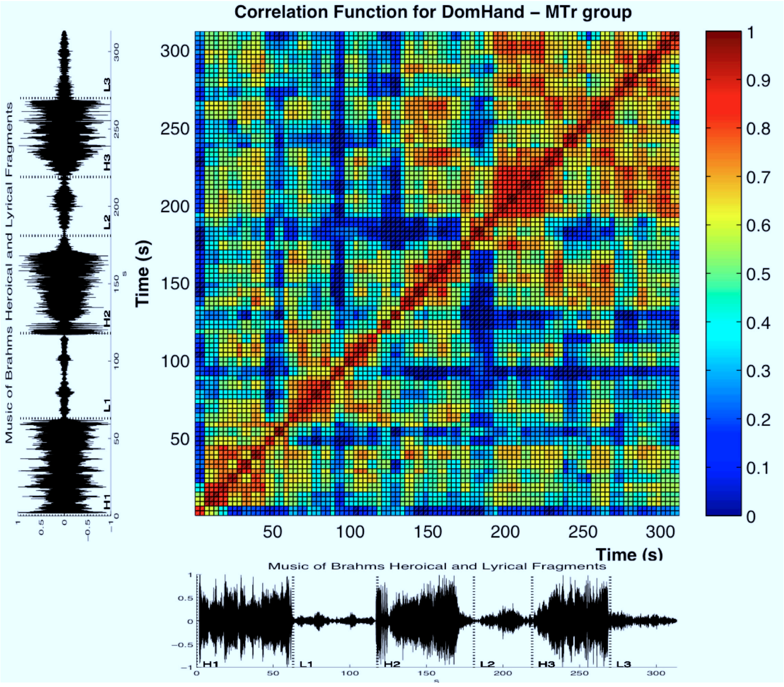
Figure 3.20: Chest : Permutation Tests to reveal MTr or MunTr Group Effect. The figure shows the histogram of the differences in coherence/consistency between the two groups. The subjects are randomly permuted between the two groups.

3.4.4.2 Analysis of Dominant Hand movement

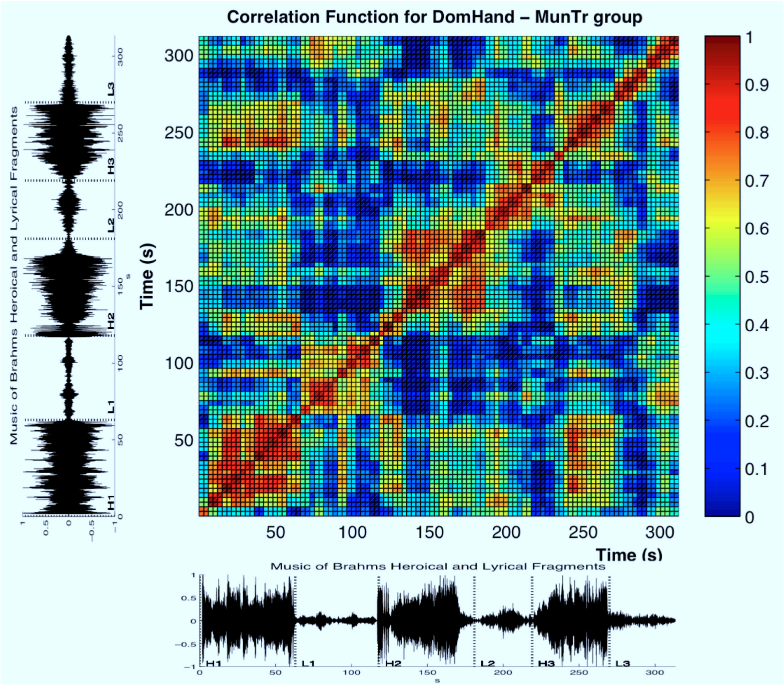
As mentioned before, hand movement is analyzed against a new axis system, eliminating all influences of the torso (Section 3.4.1). Additionally, literature (e.g. [64, 65]) makes a distinction between dominant and non-dominant hand and that discrimination is included in our analysis as well.

Correlation diagrams for the dominant hand (Fig. 3.21) reveal different patterns compared to chest (Fig. 3.15) and consequently suggest motor-imitation of other structural elements in the music (e.g. changes in musical style lyric/heroic).

Let us first discuss coherence (squares of high correlation along the diagonal). Visual inspection demonstrates that the total length of the coherence intervals does not differ much for both groups. An exact value can be calculated by some customized heuristic and the results using the same heuristic as for chest are displayed in Fig. 3.22.



(a) Musically Trained (MTr) Group.



(b) Musically Untrained (MunTr) Group.

Figure 3.21: Correlation Density Plots for Dominant Hand. (Next to the time line is the musical signal shown as in Fig. 3.1)

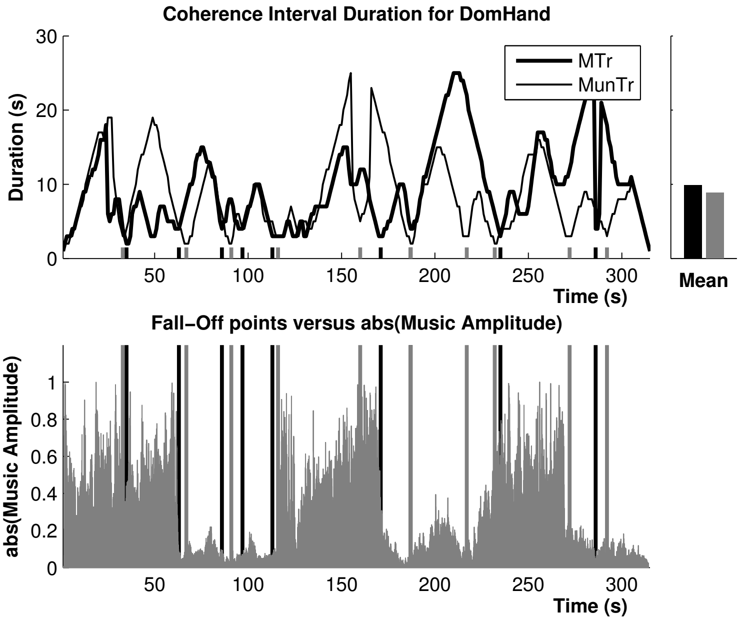


Figure 3.22: Dominant hand Coherence : Correlation Fall-Off

The main difference between both groups however, is not explained by coherence but by consistency. Consistency is characterized by high off-diagonal correlation areas. The MTr group has clearly more areas of consistency: Fig. 3.21 shows off-diagonal warmer colors for the MTr group than for the MunTr group. As the values for coherence are similar for both groups this difference in consistency should also pop-up in the previously defined combined coherence/consistency measure. This measure is the median of the distribution of the correlation coefficients shown in Fig. 3.23. A Wilcoxon rank sum test confirms that this measure is significantly (p<.001) different for the MTr Group (Mdn=0.491) than for the MunTr group (Mdn=0.395).

To check if there is a group effect, we repeat the permutation tests previously done for the Chest. The calculated p-value is now .27, not low enough to talk about a significant effect and the null hypothesis would not be rejected by a frequentist approach. However, this value says that if there was no difference between the two groups (null hypothesis) there is 27% chance that based on our test we would wrongly reject the null hypothesis (=making a type I error) and this value although not significant is worthwhile mentioning [63]. The 27% value is graphically depicted in

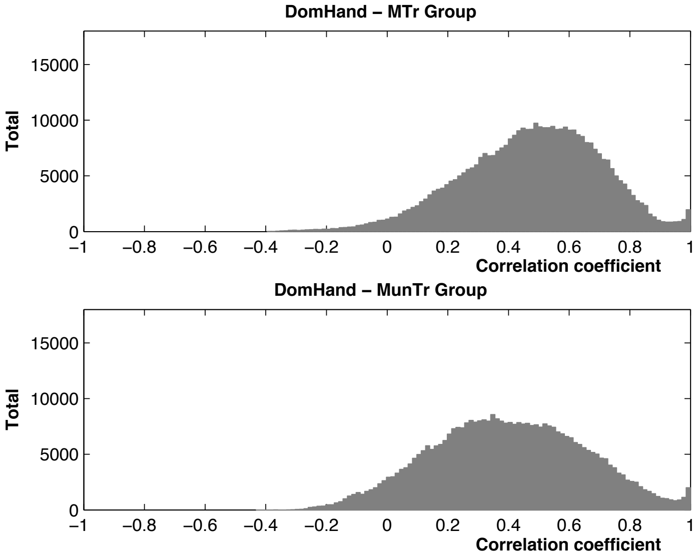


Figure 3.23: Histograms of correlation coefficients (dominant hand)

Fig. 3.24 where the black bars indicate at what values we would wrongly reject the null hypothesis. The red line indicates the decision border and is set to the difference between the two groups in their original composition.

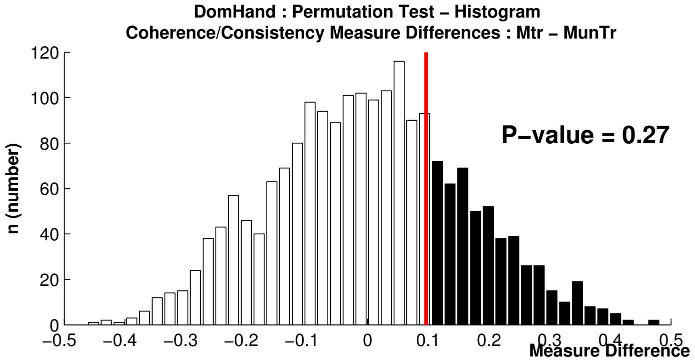


Figure 3.24: Dominant Hand : Permutation Tests to reveal MTr or MunTr Group Effect. The figure shows the histogram of the differences in coherence/consistency between the two groups. The subjects are randomly permuted between the two groups.

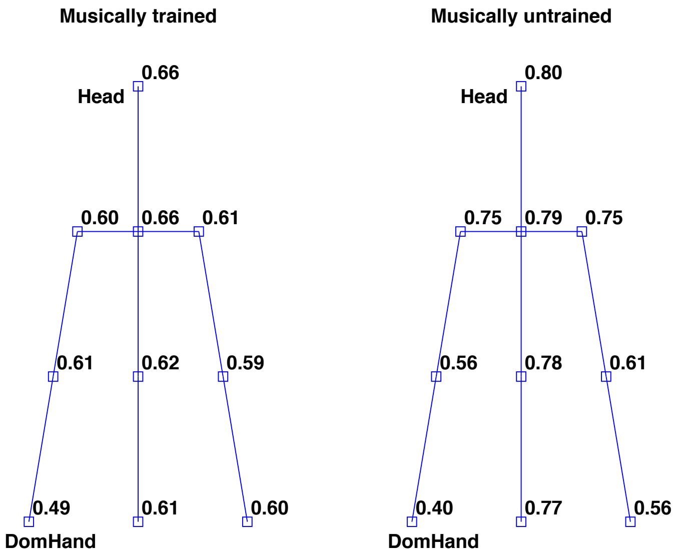


Figure 3.25: Here we show per marker the combined measure for Coherence/Consistency. (higher is more - dominant hand is shown left)

3.4.4.3 Summary

The results of all markers are summarized in an overview figure (Fig. 3.25). For every marker we calculated the single measure combining coherence and consistency. The results show different behavior for the torso (including chest) and for the arms (especially for the dominant hand).

The combined measure for chest reaches the highest value (0.78) for the MunTr group and the main contribution comes from a higher level of coherence. This can be explained by the simplicity of their movement. From Section 3.4.2.2 we know that chest movement tracks the tempo of the music and that this is even more apparent for the MunTr group. This phenomenon can be described as 'metronomic movement', periodic movement with the main focus on the tempo of the music. Movement for the MTr group is more complex, hence lower values for the combined measure (0.62). The presence of aperiodic movement is one cause for the higher complexity of movement.

To show the impact of aperiodic movement we ran all positional signals through a high pass filter (actually we removed the low frequency content with a low pass moving average filter having a 20 s window) (See Fig. 3.26).

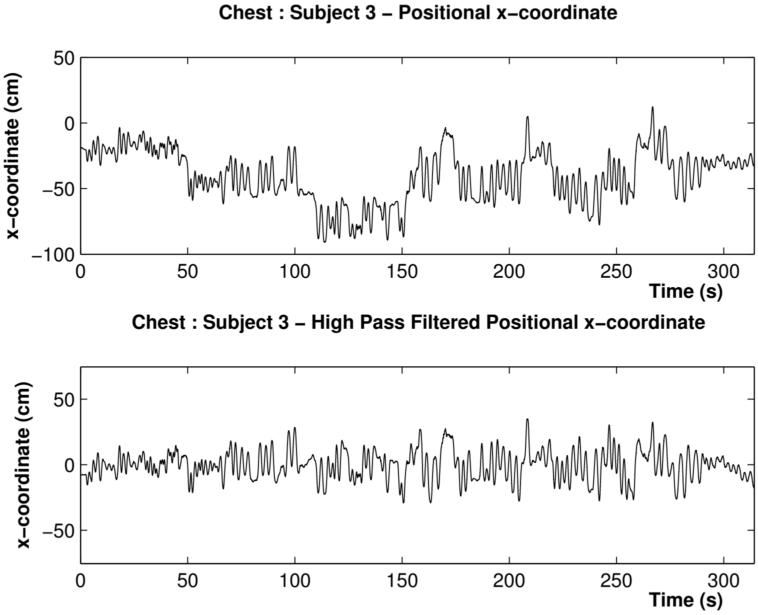


Figure 3.26: Chest : positional x-coordinate Original Signal (top) and High Pass Filtered Signal (bottom) where aperiodic displacements are removed.

We then used the filtered signals from all subjects as input to our analysis and we found indeed a significant increase for the MTr group in the combined measure (Wilcoxon rank sum test $\boldsymbol{p < .05}$) where the value raised to 0.64 from 0.62. For the MunTr group there was no significant difference.

The combined measure for the dominant hand shows a different picture. The high value for this measure is mainly due to a higher level of consistency for the MTr group and this despite the use of more complex movement. The fact that the movement is more complex can be understood from a study of the movement volume, defined as the volume within a convex hull made-up by positional coordinates over a fixed time period. A constant volume suggests a basic rhythmic movement. A higher volume stands for more expressiveness. The results are presented in Fig. 3.27.

How does it come that despite its utilization of more complex movements the MTr group reaches higher levels of consistency? We believe that this is due to the existence of some kind of absolute reference system. From section 3.4.3 we know that the MTr group correlates better with the musical amplitude. So using musical amplitude as a reference system could be an explanation for the high levels of consistency in the MTr group. However,

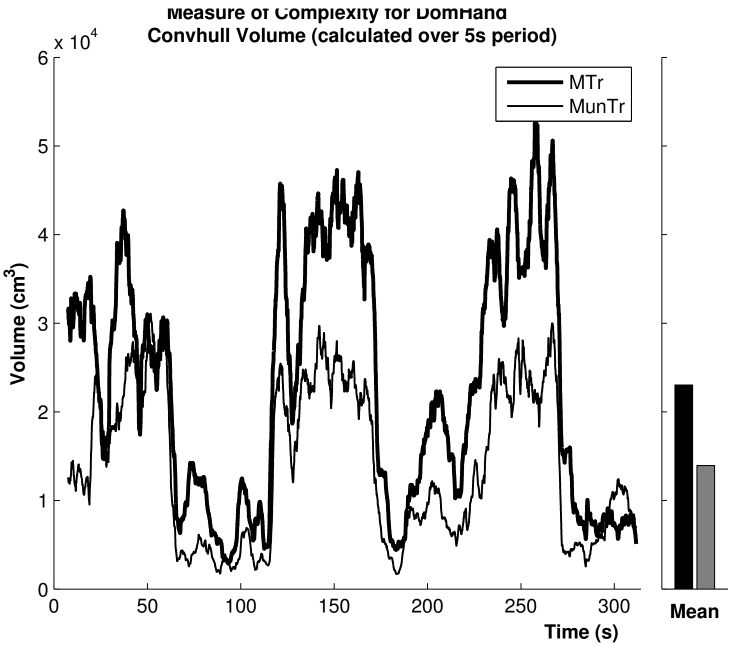


Figure 3.27: Dominant Hand : Difference in Complexity between MTr and MunTr group. The used complexity measure is the volume of the convex hull of the positional coordinates averaged over all subjects in the group. A constant volume suggests a rhythmic movement. A higher volumes means more expressivity.

more in depth research is required to confirm this.

3.5 Modeling expressiveness

In this section, we describe the process of modeling movement of a group of subjects. We discuss the models for the MTr group for both chest and dominant hand. We hereby use the log of the speed envelope signal as the input signal. The modeling process is based upon Functional Principal Component Analysis (FPCA). FPCA uses the correlation matrix as input and consequently the group's mean movement as the reference.

In our experiment we have 315 seconds of bodily movement sampled at 100 Hz which results in an equivalent multivariate data set of 31,500 variables. The advantage of a multivariate approach is that algorithms like PCA can be applied directly but this is computationally intricate. This means for example that a PCA analysis has to find the eigenvalues (called eigenfunctions) of a 31500x31500 (nxn) covariance matrix. The computational

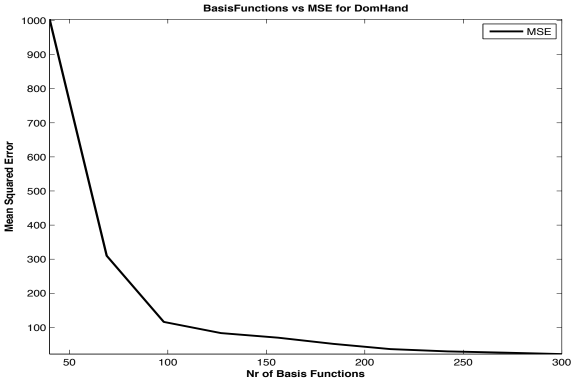


Figure 3.28: (dominant hand) Model error (MSE) versus number of basis functions. MSE is calculated as the sum of squared errors of a realization (signal) averaged over the subjects.

cost involved for most algorithms is O($\boldsymbol{n^3}$). This is a high cost.

Therefore we adopt a functional data analysis (FDA) approach that is based on one additional assumption and that is that adjacent samples are correlated. We handle this requirement by decomposing our data set in a set of basis functions as a mean of reducing the dimensionality of the problem. Uniform Cubic B-spline basis functions were chosen as the logSpeed signals lack periodicity. The number of basis functions is determined by means of a cross-validation exercise which calculates the error (MSE = Mean Squared Error) between model and signal for a different number of basis functions. The number of basis functions is set to 100 as can be understood from Fig. 3.28. This setting is valid for chest and dominant hand. Working with K basis functions reduces the computational cost for calculating eigenfunctions to O($\boldsymbol{K^3}$) [61], a considerable improvement.

FPCA calculates a set of eigenfunctions using a least square algorithm. In addition we place a penalty on the second order derivative of the eigenfunctions to favor smooth functions as suggested by Ramsay [61]. Smoothness is considered as a characteristic of human movement. All calculations were performed with the help of his matlab toolbox 'fda'. FPCA allows us to write a signal as the sum of an average signal $\boldsymbol{f\bar{(t)}}$ plus a linear combination of functions $\boldsymbol{\xi_k(t)}$ (3.3).

$$\boldsymbol{f_i(t) = f\bar{(t)} + \sum_{k=1}^{K} \alpha_{ik}\xi_k(t)} \tag{3.3}$$

The functions $\boldsymbol{\xi_k(t)}$ are the eigenfunctions that we have to calculate. The objective of FPCA is to explain as much of the variance as possible with as few eigenfunctions as possible. There exist multiple criteria to determine the number of eigenfunctions to retain. The criterion we use is that the number of eigenfunctions has to explain at least 70% of the variance.

3.5.1 Modeling Chest Movement

For the chest data of the MTr group, one eigenfunction already accounts for 73% of the variability (Fig. 3.29), which means that the variance among subjects can to a large extent be captured by one single dimension. On this figure the blue line represents the mean logSpeed and an eigenfunction is displayed as a positive and a negative offset to the mean. The offset used here is plus or minus the square root of the corresponding eigenvalue. This allows to visually compare the contributions coming from different eigenfunctions.

Given these results, we can express the performance of every subject (index i) as in equation (3.3). For the chest, the logSpeed exists out of an average function (common to all subjects=commonality) plus a factor (an individual attribute=individuality) times one eigenfunction (common to all subjects=commonality). The individual performance is characterized by a single factor. The factor is calculated as the functional principal component score for the first eigenfunction. Note that the retained eigenfunction has the same sign over the whole time range. This always makes subjects with a positive factor perform higher than average regardless of the time.

This simple model backs up the idea that chest movement (in terms of speed) is determined by simple musical characteristics. From Section 3.4.2.2 we know already that chest movements reflect the tempo of the music.

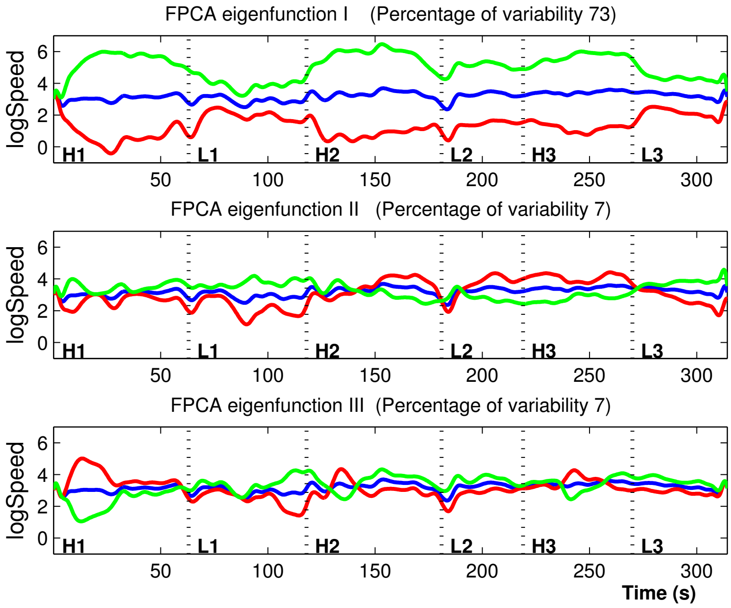


Figure 3.29: Three eigenfunctions for the Chest-Marker. The blue line represents the mean logSpeed and every eigenfunction is displayed twice, once as a positive (green) and once as a negative offset (red) to the mean. The used offset is plus or minus the square root of the corresponding eigenvalue.

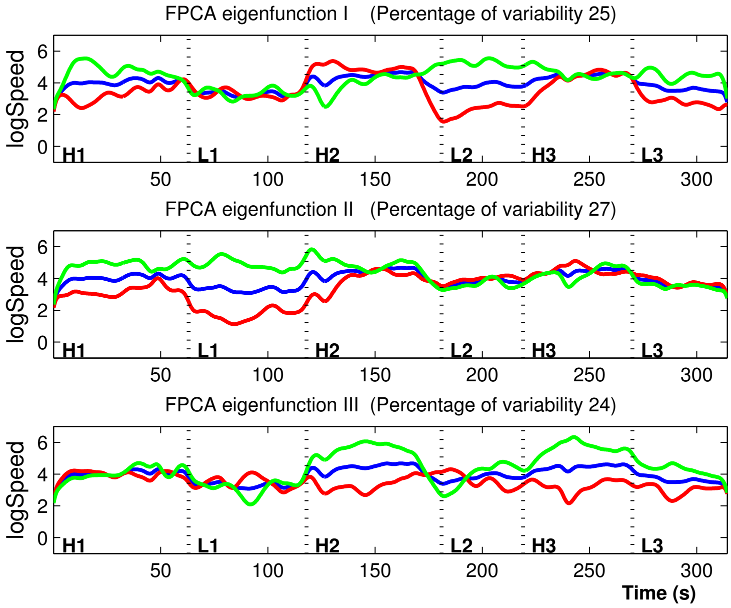


Figure 3.30: Three eigenfunctions for the dominant hand after varimax rotation. The blue line represents the mean logSpeed and every eigenfunction is displayed twice, once as a positive (green) and once as a negative (red) offset to the mean. The used offset is proportional to the amount of variance explained. Note that when an eigenfunction is zero in an interval it will coincide with the mean.

3.5.2 Modeling Dominant Hand Movement

FPCA conducted for the dominant hand reveals that we need three eigenfunctions to cover more than 70% of the variance (Picture not shown). Hands have more degrees of freedom (DOF) than a torso and consequently, the eigenfunctions will be more complex.

Similar to factor analysis for multivariate datasets, we might consider rotation of the functional principal component axes to reveal an underlying structure. One method that does so is the Varimax rotation, although other rotation techniques are possible. Varimax seeks a basis of eigenfunctions that most economically represents each individual in a way that each individual can be well described by a linear combination of only a few basis functions. If we apply a Varimax rotation to our data we get three eigenfunctions explaining 25%, 27% and 24% respectively or in total 76% of the total variance (Fig. 3.30).

We can proceed the same way as in traditional factor analysis and assign a label to every eigenfunction. The first eigenfunction (top in Fig. 3.30) shows four time intervals where the eigenfunction values deviate from zero. Three intervals with a positive deviation coincide with the first heroic part and the second and third lyric part respectively. The negative deviation coincides with the second heroic part. Subjects that score high on this eigenfunction will have higher logSpeed levels than average in the positive deviation parts and lower than average in the negative. 25% of the variance can be explained by this phenomenon and it is interesting to note that there is a relationship in expressiveness between the first heroic part and some lyric parts (part 2 and part 3). The second eigenfunction is positive in the first half of the musical excerpt, covering the first and second heroic fragment together with the first lyric fragment. This accounts for 27% of the variance. The third eigenfunction has a dominant positive contribution in the second and third heroic part. It has a mixed contribution to the second lyric part: first a negative contribution and afterwards a positive contribution when the music goes crescendo. It is remarkable that the eigenfunctions relate so well to the style intervals.

We calculate, for every subject, a component score (individuality) per eigenfunction. Three values suffice to model a subject's performance (dominant hand). The scores for the subjects in the MTr group are made visible in a 3D plot on Fig. 3.31. This representation offers new opportunities for interpreting dance performances : Dance ensembles or choreographers can use this representation to select dancers for a performance by picking out dancers that have equal scores. The tool offers also possibilities for remedial actions in the sense that out-of-line scores on a principal component can be related to a particular time interval. Eventually clustering algorithms can help to categorize the dancers.

As example we show the results of a K-means clustering with K set to three clusters (see Fig. 3.31). A first cluster groups subject 2,8,12 and 16 having high positive scores on all components. Let's label this the highly energetic group. A second group is formed by subjects 1,4,5,11,13,14 and 17. They score high on at least two components and neutral on a third. Let us call this the energetic group. The third group (low-energy) comprises the other subjects and is characterized by low logspeed levels. Interesting information is that subjects numbered from 1 to 10 are all male dancers and that they make up the majority (83%) of the low-energy group. We come back on this finding in Section 3.6.

3.5.3 Summarized

With a multivariate approach a subject's performance is represented by 31,500 variables. Using FDA and FPCA we can reduce this to 3 variables

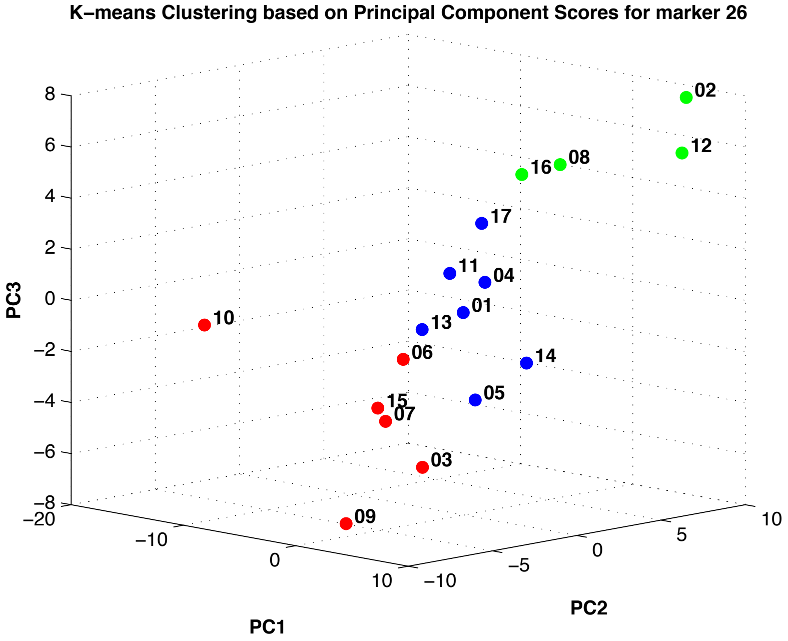


Figure 3.31: Representation of subjects by their Principal Component Scores (Individualities for the Dominant Hand). Clusters as assigned by k-means clustering (3 clusters) : cluster 1 (2, 8, 12 en 16), cluster 2 (1, 4, 5, 11, 13, 14, 17), cluster 3 (3, 6, 7, 9, 10, 15).

for the dominant hand and to 1 variable for the chest. These variables are the principal component scores, which we can also interpret as weights of eigenfunctions. The mean function and eigenfunctions are identical to all subjects and this is what defines the commonality of this population of subjects. The weights are the variables that define the individuality of the subjects.

There is an interesting relationship between the coherence/consistency analysis and the FPCA. FPCA uses the correlation matrix as input and the strong dimension reduction is only possible because of high correlations due to coherence and consistency (see section 3.4.4). However, it is not always that obvious to understand coherence and consistency in terms of eigenfunctions. In simple cases coherent intervals show up as time intervals of equal signs in eigenfunctions. Consistency appears when one eigenfunction has multiple of these coherent intervals. A simple case is when a time interval is dominated by only one eigenfunction.

3.6 Functional ANOVA to model Gender effects

We noticed in section 3.5.2 that the low-energy cluster for the dominant hand had 83% of its subjects being male. A logical question that arises from this finding is: "Do we see here a gender effect and if yes is it significant?". We use Functional Analysis of Variance or FANOVA to investigate this. FANOVA partitions the functional response according to the main effects and interactions of factors. In our example the functional response is the logspeed signal and we have only one factor (gender) with two levels (male-female). All this can be written down in a functional linear model for subject $\boldsymbol{i}$ as in equation (3.4).

$$\boldsymbol{logspeed_{ig}(t) = \mu(t) + \alpha_g(t) + \epsilon_{ig}(t);} \tag{3.4}$$

The function $\boldsymbol{\mu(t)}$ is the grand mean function of the log(speed) across all male and female dancers. The term $\boldsymbol{\alpha_g}$ refers to a same effect for all male dancers (in which case g=male) or a same effect for all female dancers (in which case g=female). Note that these effects are now functions as well. To uniquely identify these functions they are required to satisfy the sum-to-zero constraint (3.5).

$$\boldsymbol{\alpha_{male}(t) + \alpha_{female}(t) = 0 \quad , \; \forall \; t} \tag{3.5}$$

The posed problem can be solved similar to the multivariate case as a least squares problem with the parameters ($\boldsymbol{\mu(t)}$,$\boldsymbol{\alpha_g(t)}$) to estimate now being function of time. Because the parameters are functions we decompose them in basis functions using a smoothing penalty. The smoothing penalty parameter is determined by cross validation.

The analysis was run for all subjects irrespective of the group they belong to. The results were not significant neither for the dominant hand neither for the chest but nevertheless they are worthwhile discussing. First thing to know is that there is an effect when a parameter function $\boldsymbol{\alpha_g(t)}$ has values different from zero. To understand if the effect is point-wise significant we add confidence intervals to express uncertainty. For the chest (Fig. 3.32) the contribution of $\boldsymbol{\alpha_{male}(t)}$ to the logspeed signal is for men positive over the whole time-interval but it is significantly different from zero in only a very small area (around time=175 s). What we conclude is that the logspeed (and consequently also the speed) for moving the torso tends to be higher for men than for women. For the dominant hand (Fig. 3.33) the trend is opposite. Men tend to move their hands with less speed ($\boldsymbol{\alpha_{male}(t) < 0}$) than women but again this is not point-wise significant. The above findings suggest a gender trend and even more interesting is that the found trend works in opposite directions for Chest compared to Dominant Hand.

A consequence is that if we have a mixed gender group and we order the subjects in terms of their chest logspeed then the ordering will most probably be different from their ordering in terms of the dominant hand logspeed. In other words there is no correlation between chest logspeed and dominant hand logspeed. This can easily be verified from a Cross Correlation diagram (Fig. 3.34) where we plot the correlation between chest and dominant hand at all timestamps. Striking is the absence of a ridge along the positive diagonal, confirming that there is no correlation over subjects between these two body parts. This is most apparent in the lyric parts.

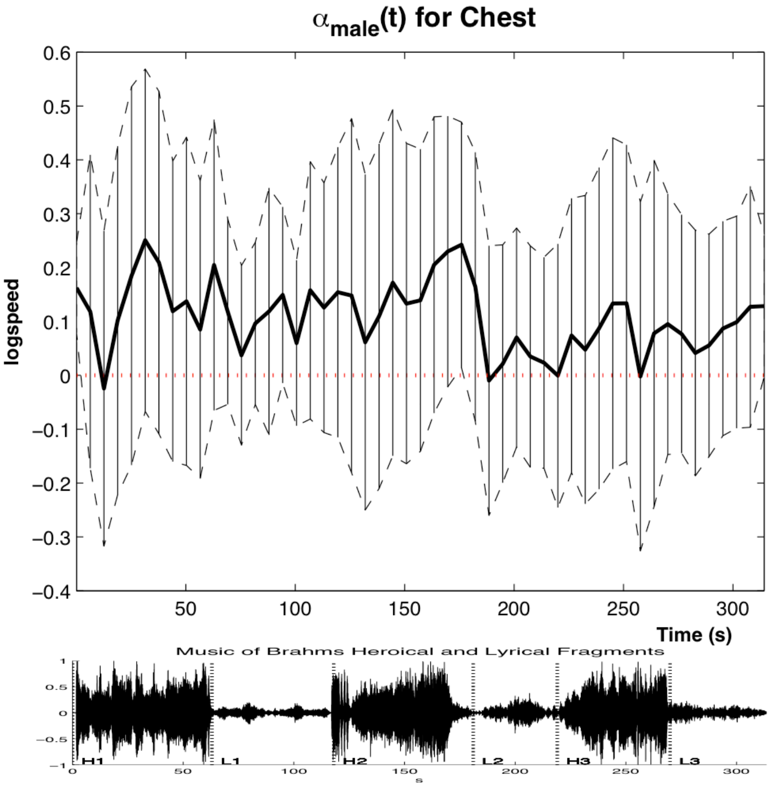


Figure 3.32: Regression coefficient function for Gender(male) in a model predicting logspeed of Chest. The cross-hatched area is the point-wise 95% confidence region for the function.

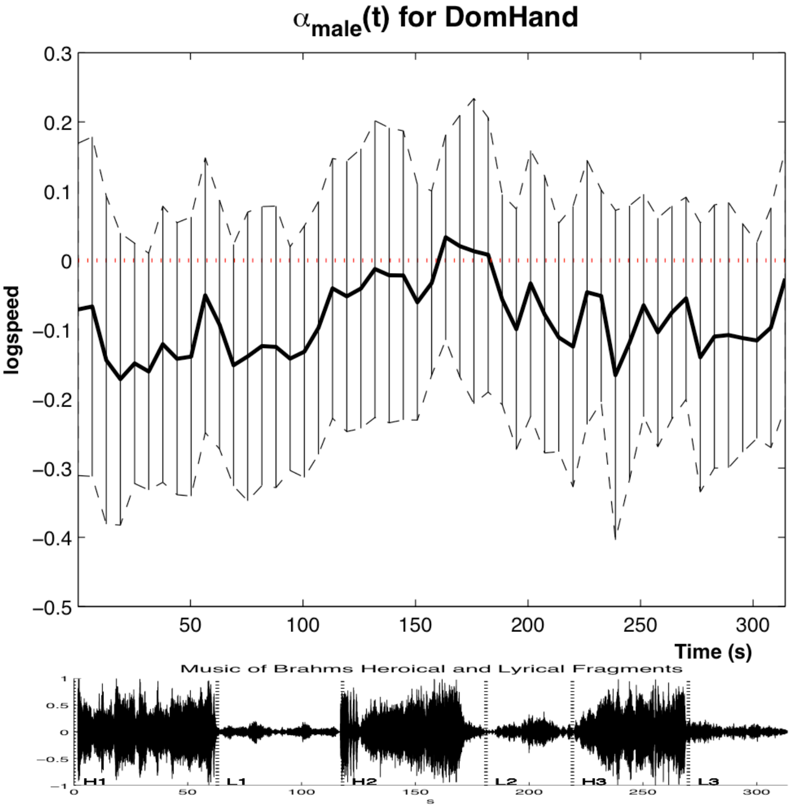


Figure 3.33: Regression coefficient function for Gender(male) in a model predicting logspeed of the Dominant Hand. The cross-hatched area is the point-wise 95% confidence region for the function.

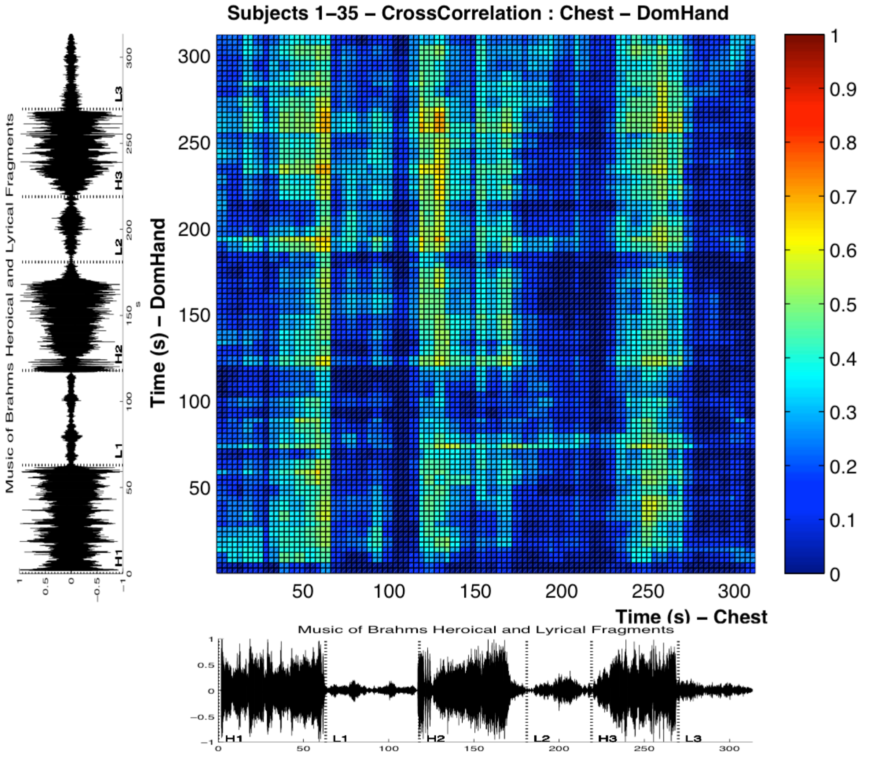


Figure 3.34: Cross Correlation density plot between Chest and Dominant Hand

3.7 Discussion

The aim of this study was to develop a dynamic approach to analyze coherence, consistency, commonality and individuality in how people mirror musical expression in their free and spontaneous body movement responses to music.

With Functional Principal Component Analysis (FPCA) we modeled every subject's bodily movement as a linear combination of a group average and a number of eigenfunctions. The group average and the eigenfunctions are common to all subjects and make up what we call the commonalities. An individual performance is then modeled by a set of scores (the individualities), one score per eigenfunction.

The model consist of one eigenfunction for chest and three eigenfunctions for the dominant hand, which covers more than 70% of the variance. Therefore the models require only few individualities which means that music is perceived and embodied in similar ways and this facilitates interaction and social effects.

FPCA uses the correlation matrix (correlating movement of subjects at distinct timestamps) as input. The low dimensionality of our models is due to the high correlations found in this correlation matrix. These high correlations are described by the concepts coherence and consistency. Coherence and consistency are interrelated and therefore we grouped them into one combined measure.

Using this combined measure we investigated differences in body parts (chest and dominant hand) and in musical background (the musically untrained (MunTr) group and the musically trained (MTr) group). The MunTr group has for the chest the highest value for this combined measure. This explains that chest movement is mainly driven by a simple musical characteristic. Our assumption is that this must be the tempo of the music and this is supported by Fourier analysis. Thus, the MunTr group focuses on periodic movements in the tempo of the music (metronomic movement).

As we move further in the kinematic chain, additional DOFs are used to express higher hierarchical structural elements in the music. The case of the dominant hand illustrates this. It still tracks the tempo as indicated by Fourier analysis, but in addition we discover especially for the MTr group a high correlation with the amplitude of the music. The MTr group focuses on additional musical characteristics (like musical amplitude) and uses the dominant hand for this.

Summarized, this confirms earlier findings [26, 27] that humans are capable of interpreting musical syntax and semantics in their movement responses to music, using different body parts.

3.8 Conclusions and Future Work

In this experiment, we analyzed two groups of subjects moving spontaneously and individually to music. One group had received formal musical education and the other group had not. The results of the study show that differences between the two groups can be largely quantified by the terms coherence and consistency. The existence of coherence and consistency leads to low dimensional models for expressiveness.

Using a combined measure for coherence and consistency, we conclude that the musically untrained group focuses on torso movement expressing the tempo of the music and that the musically trained group focuses on the dominant hand expressing additional structural elements such as musical amplitude.

Eventually, these models could also be directly applicable to analyze group movement in a diverse set of human activities, such as ensemble music playing, group dancing. Dance ensembles or choreographers can use this representation to select dancers with similar individuality scores for a collective performance. The models could also be applied in the rehabilitation of movement deficiencies such as for example, gait analysis in Parkinson patients.

References

[1] M. Leman. *Embodied music cognition and mediation technology.* The MIT Press, 2008.

[2] A. Gabrielsson and P.N. Juslin. *Emotional expression in music performance: between the performer's intention and the listener's experience.* Psychology of music, 24(1):68–91, 1996.

[3] N.L. Wallin and B. Merker. *The origins of music.* The MIT Press, 2001.

[4] A.P. Merriam. *The anthropology of music*, volume 11. Northwestern Univ Pr, 1964.

[5] R.I. Godøy and Marc Leman. *Musical gestures: Sound, movement, and meaning.* Routledge, 2009.

[6] M. Argyle. *Bodily communication (2nd ed.).* Madison, WI: International Universities Press, 1988.

[7] M. Goldstein. *Gestural coherence and musical interaction design.* In Systems, Man, and Cybernetics, 1998. 1998 IEEE International Conference on, volume 2, pages 1076–1079. IEEE, 1998.

[8] J.W. Davidson. *Qualitative insights into the use of expressive body movement in solo piano performance: a case study approach.* Psychology of Music, 35(3):381–401, 2007.

[9] F. Desmet, L. Nijs, M. Demey, M. Lesaffre, J.P. Martens, and M. Leman. *Assessing a Clarinet Player's Performer Gestures in Relation to Locally Intended Musical Targets.* Journal of New Music Research, 41(1):31–48, 2012.

[10] M. R. Thompson and G. Luck. *Exploring relationships between pianists' body movements, their expressive intentions, and structural elements of the music.* Musicae Scientiae, 16(1):19–40, 2012.

[11] A. Kirke and E. R. Miranda. *Guide to computing for expressive music performance.* Springer, 2013.

[12] B. H. Repp. *Musical motion: Some historical and contemporary perspectives.* In A. Friberg, J. Iwarsson, E Jansson, and J. Sundberg, editors, Proc. Stockholm Music Acoustics Conference (SMAC), pages 128–135, 1993.

[13] P. Shove and B. H. Repp. *Musical motion and performance: Theoretical and empirical perspectives.* In J. Rink, editor, The practice of performance, pages 55–83. Cambridge, UK: Cambridge University Press, 1995.

[14] M. L. Johnson. *Embodied musical meaning.* Theory and Practice, 22:95–102, 1997.

[15] A. Cox. *Embodying music: Principles of the mimetic hypothesis.* Music Theory Online, 17(2):1–24, 2011.

[16] A. Tierney and N. Kraus. *The ability to move to a beat is linked to the consistency of neural responses to sound.* The Journal of Neuroscience, 33(38):14981–14988, 2013.

[17] R.I. Godøy. *Motor-mimetic music cognition.* Leonardo, 36(4):317–319, 2003.

[18] M. Leman, F. Desmet, F. Styns, L. Van Noorden, and D. Moelants. *Sharing musical expression through embodied listening: A case study based on Chinese Guqin music.* Music Perception, 26(3):263–278, 2009.

[19] R. I. Godøy. *Gestural affordances of musical sound.* In R. I. Godøy and M. Leman, editors, Musical gestures: Sound, movement, and meaning. New York, NY: Routledge, 2010.

[20] F. Desmet, M. Leman, M. Lesaffre, and L. Bruyn. *Statistical analysis of human body movement and group interactions in response to music.* Advances in Data Analysis, Data Handling and Business Intelligence, pages 399–408, 2010.

[21] G. Varni, M. Mancini, G. Volpe, and A. Camurri. *Sync'n'Move: social interaction based on music and gesture.* User Centric Media, pages 31–38, 2010.

[22] T. Eerola, G. Luck, and P. Toiviainen. *An investigation of preschoolers' corporeal synchronization with music.* In Proceedings of the 9th International Conference on Music Perception & Cognition (ICMPC9). Bologna, Italy: ICMPC and ESCOM, pages 472–6, 2006.

[23] F. Ofli, E. Erzin, Y. Yemez, and A.M. Tekalp. *Learn2Dance: Learning Statistical Music-to-Dance Mappings for Choreography Synthesis.* Multimedia, IEEE Transactions on, 14(3):747–759, 2012.

[24] M. Leman and L. Naveda. *Basic gestures as spatiotemporal reference frames for repetitive dance/music patterns in Samba and Charleston.* Music Perception, 28(1):71–91, 2010.

[25] R. Fan, S. Xu, and W. Geng. *Example-Based Automatic Music-Driven Conventional Dance Motion Synthesis.* Visualization and Computer Graphics, IEEE Transactions on, 18(3):501–515, 2012.

[26] S. Dahl and A. Friberg. *Expressiveness of musician's body movements in performances on marimba.* Gesture-based communication in human-computer interaction, pages 361–362, 2004.

[27] P. Toiviainen, G. Luck, and M.R. Thompson. *Embodied meter: Hierarchical eigenmodes in music-induced movement.* Music Perception, 28(1):59–70, 2010.

[28] B. Tversky. *Functional significance of visuospatial representations.* Handbook of higher-level visuospatial thinking, pages 1–34, 2005.

[29] P.L. Jackson, A.N. Meltzoff, and J. Decety. *Neural circuits involved in imitation and perspective-taking.* Neuroimage, 31(1):429–439, 2006.

[30] A. Billard. *Learning motor skills by imitation: a biologically inspired robotic model.* Cybernetics & Systems, 32(1-2):155–193, 2001.

[31] G. Goldenberg and H.O. Karnath. *The neural basis of imitation is body part specific.* The Journal of neuroscience, 26(23):6282–6287, 2006.

[32] T.M. Nakra. *Inside the Conductor's Jacket: Analysis, interpretation and musical synthesis of expressive gesture.* PhD thesis, Massachusetts Institute of Technology, Department of Media Arts and Sciences, 1999.

[33] S. Goldin-Meadow. *Hearing gesture: How our hands help us think.* Belknap Press, 2005.

[34] S. Goldin-Meadow and M. W. Alibali. *Gesture's role in speaking, learning, and creating language.* Annual review of psychology, 64:257–283, 2013.

[35] E. T. Hall, R. L. Birdwhistell, B. Bock, P. Bohannan, A. R. Diebold Jr, M. Durbin, M. S. Edmonson, JL. Fischer, D. Hymes, S. T. Kimball, et al. *Proxemics [and Comments and Replies].* Current anthropology, pages 83–108, 1968.

[36] S. Gibet. *Sensorimotor Control of Sound-producing Gestures.* Musical Gestures: Sound, Movement, and Meaning, page 212, 2009.

[37] J. JA. Denissen, R. Geenen, M. AG. Van Aken, S. D Gosling, and J. Potter. *Development and validation of a Dutch translation of the Big Five Inventory (BFI).* Journal of Personality Assessment, 90(2):152–157, 2008.

[38] C. E. Osgood. *The measurement of meaning*, volume 47. University of Illinois Press, 1957.

[39] W. F. White and J. H. Butler. *Classifying meaning in contemporary music.* The Journal of Psychology, 70(2):261–266, 1968.

[40] K. Swanwick. *Musical cognition and aesthetic response.* Bulletin of the British Psychological Society, 26(93):285–289, Oct 1973.

[41] S. Nielzén and Z. Cesarec. *On the perception of emotional meaning in music.* Psychology of music, 9(2):17–31, Oct 1981.

[42] T. Fujihara and N. Tagashira. *A multidimensional scaling of classical music perception.* Japanese Journal of Psychology, 55(2):75–79, Jun. 1984.

[43] M. Senju and K. Ohgushi. *How are the player's ideas conveyed to the audience?* Music Perception, pages 311–323, 1987.

[44] T. Murakami and P.M. Kroonenberg. *Three-mode models and individual differences in semantic differential data.* Multivariate Behavioral Research, 38(2):247–283, 2003.

[45] E. Bigand and B. Poulin-Charronnat. *Are we "experienced listeners"? A review of the musical capacities that do not depend on formal musical training.* Cognition, 100(1):100–130, 2006.

[46] E. Bigand, S. Vieillard, F. Madurell, J. Marozeau, and A. Dacquet. *Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts.* Cognition and emotion, 19(8):1113–1139, 2005.

[47] P.-J. Maes, E. Van Dyck, M. Lesaffre, P.M. Kroonenberg, and M. Leman. *A dimensional model for the study of the coupling of action and perception in musical meaning formation: A case study with Brahms' First Piano Concerto.* Music Perception, in-press.

[48] E. Carl. *Seashore, The Psychology of Music*, 1938.

[49] A. Camurri, B. Mazzarino, M. Ricchetti, R. Timmers, and G. Volpe. *Multimodal analysis of expressive gesture in music and dance performances.* Gesture-based communication in human-computer interaction, pages 357–358, 2004.

[50] W.T. Dempster and G.R.L. Gaughran. *Properties of body segments based on size and weight.* American Journal of Anatomy, 120(1):33–54, 2005.

[51] E. Pöppel. *The Measurement of Music and the Cerebral Clock: A New Theory.* Leonardo, 22(1):pp. 83–89, 1989.

[52] A. Savitzky and M. JE. Golay. *Smoothing and differentiation of data by simplified least squares procedures.* Anal. Chem., 36(8):1627–1639, 1964.

[53] D. Mazzoni and R. Dannenberg. *Audacity [software]. Pittsburg*, 2000.

[54] L. van Noorden and D. Moelants. *Resonance in the perception of musical pulse.* Journal of New Music Research, 28(1):43–66, 1999.

[55] RR Sokal and FJ Rohlf. *Biometry (3rd edn).* WH Freman and company: New York, 1995.

[56] L.H. Koopmans, D.B. Owen, and JI Rosenblatt. *Confidence intervals for the coefficient of variation for the normal and log normal distributions.* Biometrika, 51(1/2):25–32, 1964.

[57] E. Diletti, D. Hauschke, and VW Steinijans. *Sample size determination for bioequivalence assessment by means of confidence intervals.* International journal of clinical pharmacology, therapy, and toxicology, 29(1):1, 1991.

[58] S.A. Julious and C.A.M. Debarnot. *Why are pharmacokinetic data summarized by arithmetic means?* Journal of biopharmaceutical statistics, 10(1):55–71, 2000.

[59] F. Desmet. *A statistical framework for embodied music cognition.* PhD thesis, University of Ghent, Department of Arts, 2011.

[60] B. Dodson. *The Weibull analysis handbook.* ASQ Quality Press, 2006.

[61] J.O. Ramsay. *Functional data analysis.* Wiley Online Library, 111 River Street Hoboken NJ 07030-5774 USA, 2006.

[62] F.B. Butar and J.W. Park. *Permutation tests for comparing two populations.* Journal of Mathematical Science & Mathematics Education V, 3:19–30, 2008.

[63] I. DJ. Bross. *Critical levels, statistical language and scientific inference.* Foundations of Statistical Inference, pages 500–513, 1971.

[64] D. Kilshaw and M. Annett. *Right-and left-hand skill I: Effects of age, sex and hand preference showing superior skill in left-handers.* British Journal of Psychology, 74(2):253–268, 1983.

[65] CA Armstrong and JA Oldham. *A comparison of dominant and non-dominant hand strengths.* Journal of Hand Surgery (British and European Volume), 24(4):421–425, 1999.

Statistics may be defined as a body of methods for making wise decisions in the face of uncertainty.

W.A. Wallis, 1912-1998

4

Beating-Time Gestures: Imitation Learning for Humanoid Robots

Abstract

Beating-Time Gestures are movement patterns of the hand that sway along with the music, thereby indicating musical pulses in time while performing a particular spatial conducting pattern. The spatiotemporal configuration of these patterns makes it difficult to analyze and model them. In this paper we present a modeling approach that is based upon imitation learning or programming by demonstration (PbD). PbD derives a generalized trajectory from a set of demonstrations to use as target for humanoid robots. Our procedure uses a Dirichlet Process Mixture Model as front end for a continuous Hidden Markov Model to characterize every beating-time gesture by a set of non-equidistant key points. Dynamic Time Warping is our solution for handling the temporal variation of these key points. Eventually, we produce a smooth generalized trajectory by means of non-uniform cubic spline regression. The regression step accounts for the spatial variation in the set of demonstrations. The parametric form of the generalized trajectory makes it suitable for use with any musical fragment.

4.1 Introduction

Body movements having a particular goal-directed component in relation to music are called musical gestures [1]. For example, musicians may move their fingers on a string in order to play notes with a particular expressive quality. In a similar way, listeners may perform learned repetitive movement patterns while listening to music. In both cases, the movements follow some intended spatial trajectory within timely boundaries, which is a sufficient reason to call them gestures.

Most studies of musical gestures reduce the broad multifaceted aspect of musical gestures by focusing on particular gestures, or particular movement tasks that constrain the gestures. In this paper, we constrain the multifaceted nature of musical gestures by defining a movement task related to conducting, that is, movement patterns of the hand that sway along with the music, thereby indicating musical pulses while performing a particular spatial conducting pattern. These gestures are what we call beating-time gestures.

Conducting movements have been studied from different perspectives, including recognition for conducting systems (See [2] for an overview), expressiveness [3], synchronization with musicians [4]. However, the spatiotemporal account has not often been explored in this context. Systems are typically restricted to either the temporal domain or the spatial domain. One exception is [5], where spatiotemporal motion templates are used.

Dealing with variance in the spatial and in the temporal domain is challenging and it constitutes the main part of this chapter. Our approach starts from a set of demonstrations and as such our technique is called PbD (also known as learning by imitation). This technique is well known in the robotics industry. It allows a robot to learn a skill through demonstrations and thus without explicitly programming each detail. Our work is inspired by [6] but adapted to the specificities of beating-time gestures.

The paper is structured as follows. Section 4.2 provides the necessary definitions and gives an overview and motivation of the methodology used. Section 4.3 describes the experiment that lays at the basis of our research. Data processing makes up the core of our research and is handled in section 4.4. The results are then benchmarked against other methods in section 4.5. Eventually an application for music is introduced in section 4.6. This is followed by a discussion in section 4.7 and conclusions are drawn in section 4.8.

4.2 Background

4.2.1 Definitions

Beating-time gestures indicate both the musical beats and the higher level metrical-structure (here the 4/4 meter) that is defined by the accentuation of these beats. The obvious method for describing these gestures is by looking at the shape of their trajectories. The classical trajectory shows a movement where the right hand goes down to reach the first beat, left to reach the second beat, right to the third beat and up for the forth beat, shown as model 1 in Fig. 4.1. Obviously, other patterns exist and in this paper we present different four-beat patterns as the ones shown in Fig. 4.1.

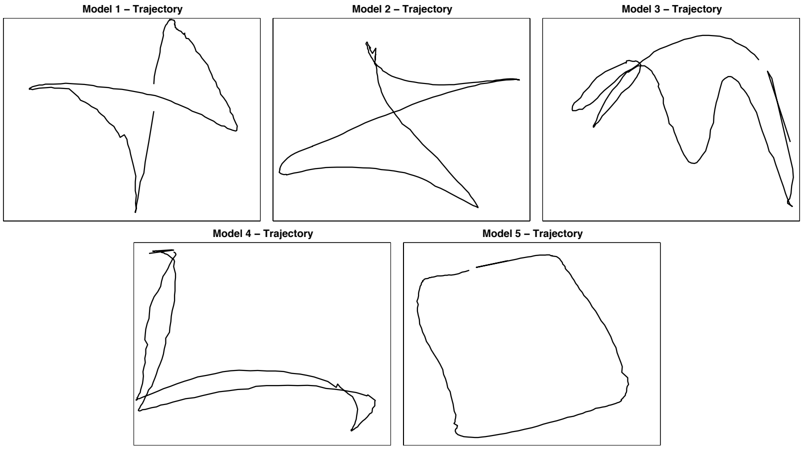


Figure 4.1: Spatial configuration of beating time gestures with time-structure (4/4 meter)

Fig. 4.1 shows for every model a, what we call, basic gesture. A basic gesture characterizes one single accentuated beat pattern that is extracted from a single performance containing recurrent movements [7]. A basic gesture is performed in exactly one measure (bar).

4.2.2 Goal and methodology

The goal of the study is to create a template for a robot to allow a performance of subsequent basic beating-time gestures in a continuous manner.

There exist many implementations (hardware and software) for robots but a common principle is that force is applied to accomplish a positional and/or velocity target [6]. Force is normally expressed as an acceleration command (4.1) and it is used to track the desired velocity and position using a proportional-derivative (PD) controller.

$$\ddot{\boldsymbol{x}} = \boldsymbol{\kappa_v}(\hat{\dot{\boldsymbol{x}}} - \dot{\boldsymbol{x}}) + \boldsymbol{\kappa_p}(\hat{\boldsymbol{x}} - \boldsymbol{x}) \tag{4.1}$$

$\boldsymbol{\kappa_v}$ and $\boldsymbol{\kappa_p}$ are gain parameters similar to damping and stiffness factors. $\boldsymbol{x}$ is a vector representing positional information in line with the degrees of freedom (DOF) of a robot. $\boldsymbol{x}$ can hold Cartesian coordinates as well as angle coordinates. We follow here the conventional notation for derivatives being $\dot{\boldsymbol{x}}$ for speed and $\ddot{\boldsymbol{x}}$ for acceleration. The hat-symbol is used for indicating the target values: $\hat{\boldsymbol{x}}$ stands for the target position and $\hat{\dot{\boldsymbol{x}}}$ for the target velocity.

Equation (4.1) explains the main interest of this paper. We assume that a robot can determine its current position ($\boldsymbol{x}$) and its current velocity ($\dot{\boldsymbol{x}}$). We do not discuss the details of tuning a robot ($\boldsymbol{\kappa_v}, \boldsymbol{\kappa_p}$). The focus of this paper lays completely on the calculation of $\hat{\boldsymbol{x}}$ and $\hat{\dot{\boldsymbol{x}}}$ or, in other words on calculating a target trajectory for a beating-time gesture.

Our solution proposes PbD for generating the target trajectory. PbD calculates a generalized gesture from a set of demonstrations. Obviously, this could be done by selecting one of the demonstrated basic gestures using some criterion. This solution however, although tempting, does not take into consideration the spatial and temporal variation that exists in *all* demonstrations.

Our solution handles the spatial variation by cubic spline regression. This has as additional advantage that it deals well with periodic boundaries. A beating-time gesture is part of a continuously repeated sequence, and so we want the beginning and the end of the generalized gesture to coincide. Cubic spline regression is often done with a set of equidistant knots (uniform splines). Then, extrema in the trajectory can or can not coincide with the knots. If they do not coincide, the consequence is that the extrema of the trajectory are flattened out resulting in a compressed shape. Because beating-time gestures use the extrema to convey beat information, we do not go that path and we choose for non-uniform splines instead.

We handle the temporal variation by adding a dynamical time warping (DTW) step. This is achieved by warping all demonstrations non-linearly in the time dimension to a reference signal. Here, the challenge comes from the calculation of a reference signal.

We propose to handle the remaining issues by fitting an HMM. The average timestamps of where the HMM state transitions happen are then used (i) for setting the non-equidistant knots for cubic spline regression and (ii) for the creation of a reference signal for DTW.

As we prefer to keep the set of demonstrations low we need a simple model, in our case a HMM with few parameters. The number of HMM states and the initial values for Baum-Welch training of the HMM parameters follow from a Dirichlet Process Gaussian Mixture Model (DPGMM) that we fit to the data. DPGMM is a Bayesian method using a Dirichlet process as prior. The prior acts as a regularizer preventing overfitting and resulting in models that usually generalize better. This is an asset, as in our case we have few data and model fitting with few data is prone to overfitting. For more information on DPGMM we refer to existing literature (e.g. Teh [8] and El-Arini [9]). Fig. 4.2 gives an overview of the complete PbD procedure we propose.

The resulting template has a number of features that make it easy for further applications. The form of the template is parametric what makes it easily adjustable in amplitude as well as in time. In this way it can be easily adjusted to match the beats of any musical fragment. Additionally, because of the spline representation, it is also easy to calculate the derivatives.

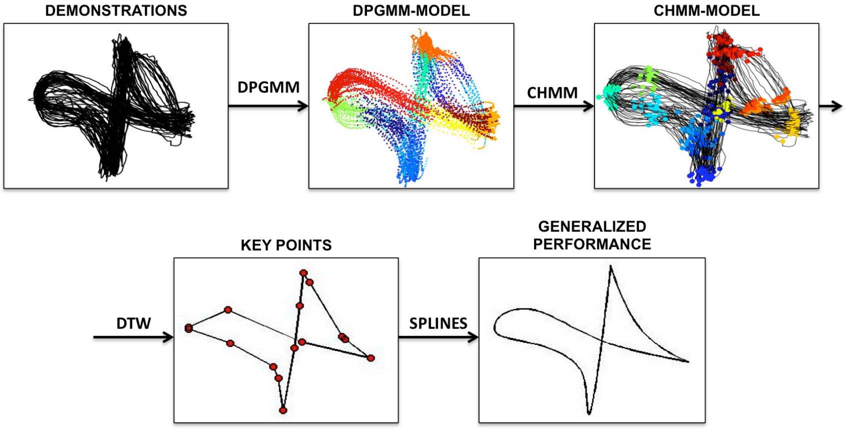


Figure 4.2: Flow Chart of the followed procedure for PbD. (i) DPGMM, as front-end for a continuous HMM, is used to calculate its number of hidden states and to set its initial emission values. (ii) a cHMM defines the key points which are used to create a reference signal for (iii) DTW. Eventually (iv) non-uniform cubic spline regression on the warped gestures produces a smooth generalized gesture.

4.3 Experimental Set-Up

Subjects + Task. Four subjects having no musical background and aged between 18 and 20 were asked to perform repetitive cycles of beating-time gestures. The gestures were defined upfront and were selected out of five different conducting models as depicted in Figure 4.1. Every subject performed 40 basic gestures of a particular model but did not perform on all five models. The assignment of conducting models to subjects was random with the restriction that every subject had to perform on the commonly known conducting model (labeled as model 1) and two more conducting models. More particularly, Subject 1 performed on conducting models 1, 2 and 3. Subject 2 performed on conducting models 1,2 and 4. Subject 3 and 4 performed on conducting models 1,3 and 5. This meant that in total 480 basic beating-time gestures (120 per subject) were generated.

Stimuli. The stimuli consisted of 40 bars of a repetitive metrical pattern exhibited by metronome ticks at a tempo of 120 BPM using a 4/4 time signature. A basic conducting gesture lasted 2 s.

Data. The data from the movements of the hand was collected by an OPTITRACK infrared optical system consisting of 12 synchronized cameras with related ARENA motion capture software (`http://www.naturalpoint.com`). Recordings were made at a sample rate of 100 Hz. Participants were asked to put on two sets of three infrared reflecting markers, each set defining a rigid body that can be easily identified by the motion capture software. One set was placed at the hand and one set at the chest. The set at the chest was meant for positional reference.

4.4 Data Processing

Initial inspection of the data showed that in a series of performances usually the first ones and the last ones were outliers comparable to a warming-up and cooling-down effect. These were excluded from the analysis.

Beating-time gestures are simple geometric movements and most of them can be studied by projection of the positional coordinates onto the frontal (coronal) plane. In our set-up the coronal plane is defined by the recorded chest markers. The coordinates of the hand markers, making up the conducting gesture, were then orthogonally projected onto this coronal plane. This three to two dimensional reduction permits a better visualization. For the envisioned application (robot) full dimensional data should be used instead.

In addition to the positional coordinates we calculated the velocity as the derivative of the positional data. A local (linear) regression filter was applied to calculate smooth derivatives. The size of the regression window was set to 0.100 s corresponding with a linear frequency response of the derivation filter in the useful frequency band of 0-6 Hz. The 0-6 Hz range was derived from spectrograms. This regression filter was applied to all coordinates.

In the course of a demonstration, we spotted that position and scaling changed from measure to measure. To overcome this issue we implement a normalization step. We consider two different methods for normalization giving slightly different results (Fig. 4.3). One method interprets the entire set of demonstrations as one long lasting gesture. Normalization is then equal to high pass filtering (detrending) followed by scaling. Another method views the entire set as a sequence of separate individual basic gestures and normalizes per basic gesture by subtracting the basic gesture's average values. Visual inspection learns that the latter fits better reality. That is particularly visible at sample 7200-7400 where we see that an entire basic gesture shifts up. So we choose for the second method by normalizing the individual basic gestures.

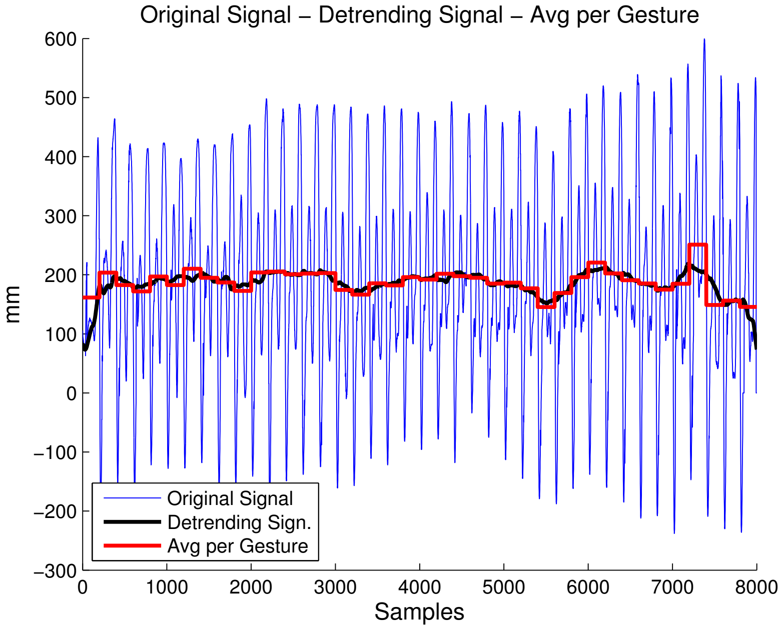


Figure 4.3: Differences between 2 normalization methods. Method 1 uses the detrending signal (black) for normalization. Method 2 uses the average per basic gesture (red) for normalization .

The normalized variables are then stored in a four-dimensional data vector: $\boldsymbol{X_{m,n}} = [\boldsymbol{posx_{m,n}}\ \boldsymbol{posy_{m,n}}\ \boldsymbol{velx_{m,n}}\ \boldsymbol{vely_{m,n}}]$ where $\boldsymbol{m}$ is the basic gesture index (in our case a value from 1 to 40) and $\boldsymbol{n}$ the sample index in our basic gesture (in our case n ranges from 1 to 200, the number of samples per gesture). We then use the notation $\boldsymbol{X_m}$ to refer to all samples from one basic gesture. In Fig. 4.4 we display the variables $\boldsymbol{posx}, \boldsymbol{posy}, \boldsymbol{velx}, \boldsymbol{vely}$ representing the normalized versions of the horizontal position, respectively the vertical position, the horizontal velocity and the vertical velocity.

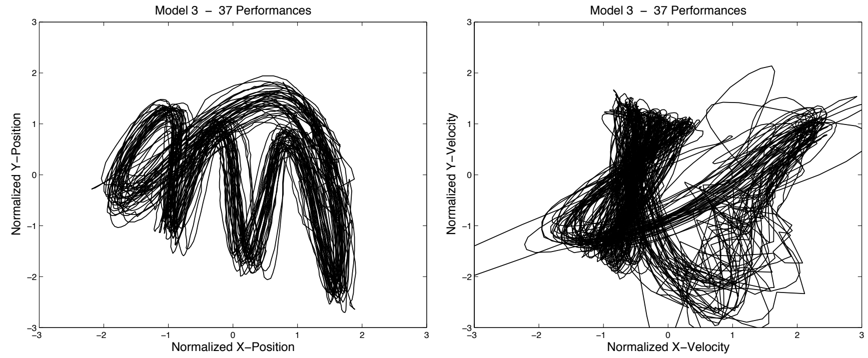


Figure 4.4: Normalized trajectories. Left, positional coordinates, right, velocity coordinates. The trajectories for 37 basic gestures are shown.

The challenge now, is to handle two sources of variation that are correlated namely, the spatial and the temporal variation. To that end we extract information from all the demonstrated basic gestures, rather than just smoothing an average basic gesture. In this way we follow the methods explored by Vakanski et al. [10] and Aleotti et al. [11] but with adaptations to our specific needs. Our gestures are beating-time gestures and as such they are subject to temporal constraints.

The procedure breaks down in three major steps: First (i) we do key point extraction using HMMs and (ii) secondly we apply DTW followed by (iii) the generation of a generalized trajectory via non-uniform B-Spline regression.

4.4.1 Key point extraction.

The key points constitute the fingerprint of a gesture, the minimum amount of information to reconstruct a trajectory. Our approach places the key points at the hidden state transitions of a continuous HMM (cHMM).

To fit an HMM we consider our movement trajectories in an augmented feature space of four dimensions (4D), having two-dimensional (2D) position variables, and 2D velocity variables. Remember from (4.1) that we need a target trajectory for position and velocity.

The number of internal HMM states and the initial values of the HMM parameters are calculated from a DPGMM. The DPGMM is similar to a Gaussian Mixture Model (GMM) except that the number of clusters is determined directly from the data and not from an additional data validation step. The DPGMM clusters are shown in Fig. 4.5 using two separate 2D representations, one for the positional coordinates and one for the velocity coordinates. The cluster assignment reveals that the performance can be understood as a chain of single Gaussians. We therefore propose as model a Bakis left-to-right HMM with single Gaussian emissions.

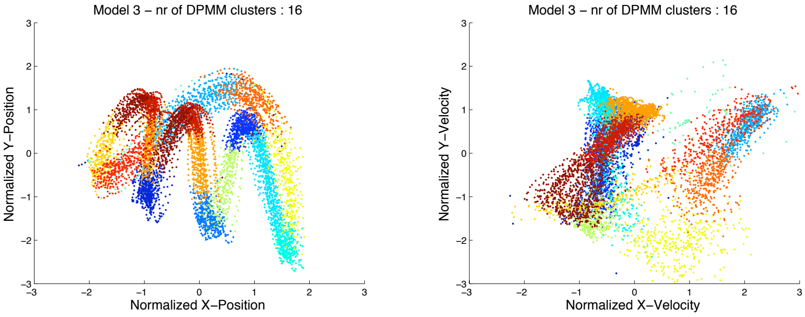


Figure 4.5: Cluster assignment (based upon positional coordinates and velocity coordinates).

The exact number of states is derived from a single basic gesture $\boldsymbol{m^*}$ which is selected via some criterion. Our criterion is the maximum log-likelihood of the gesture given the DPGMM (4.2-4.3).

$$\boldsymbol{p(x|\theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)} \tag{4.2}$$

$$\begin{aligned} \boldsymbol{m^*} &= \boldsymbol{argmax_m(log(P(X_m|\theta)))} \\ &= \boldsymbol{argmax_m(\sum_{n=1}^{200} log(\sum_{k=1}^{K} \pi_k \mathcal{N}(x_{m,n}|\mu_k, \Sigma_k)))} \end{aligned} \tag{4.3}$$

Note that $\boldsymbol{x}$ stands here for a 4D data vector (including position and velocity), $\boldsymbol{K}$ for the number of clusters, $\boldsymbol{\pi_k}$ for the mixing weight of cluster $\boldsymbol{k}$ and ($\boldsymbol{\mu_k}$,$\boldsymbol{\Sigma_k}$) are the mean and the covariance matrix of cluster $\boldsymbol{k}$. $\boldsymbol{X_m}$ stands for the $\boldsymbol{x}$-vectors of basic gesture $\boldsymbol{m}$. The number of states is set equal to the number of segments (vectors with the same winning Gaussian) in the best basic gesture $\boldsymbol{m^*}$.

Besides the number of hidden states we need to learn the other HMM parameters as well. HMM parameters are usually denoted as $\boldsymbol{\lambda = (\pi, A, E)}$ with $\boldsymbol{\pi}$ the vector of initial state probabilities, $\boldsymbol{A}$ the matrix of the transition probabilities and $\boldsymbol{E}$ representing the emission probabilities.

For a cHMM, $\boldsymbol{E}$ consists of a set of parameters describing a density. Our choice for Gaussian densities was based on the observation that one cluster dominates $\boldsymbol{p(x|\theta)}$ at all times. This choice also helps to reduce the number of HMM parameters.

Without loss of generality we can set $\boldsymbol{\pi} = [\mathbf{1}\ \mathbf{0}\ \mathbf{...}\ \mathbf{0}]$ meaning that we always start at hidden state 1. The other parameters ($\boldsymbol{A}$,$\boldsymbol{E}$) are learned from the data by means of the Baum-Welch algorithm. The Baum-Welch algorithm needs initial values for the transition probability matrix ($\boldsymbol{A}$) and for the emission densities ($\boldsymbol{E}$). The transition probabilities are set to allow only self-transitions and forward transitions to the next state and to the second next state. Their initial settings are calculated from the state assignments implied by the best basic gesture $\boldsymbol{m^*}$ found before. Here $\boldsymbol{\tau_i}$ represents the duration of the segment corresponding to state $\boldsymbol{i}$. According to the recommendations of [10], the transition probabilities are set to

$$\begin{aligned} \boldsymbol{A_{i,i}} &= \mathbf{1} - \frac{\mathbf{1}}{\boldsymbol{\tau_i}} \\ \boldsymbol{A_{i,i+1}} &= \frac{\mathbf{1}}{\boldsymbol{\tau_i}} \\ \boldsymbol{A_{i,i+2}} &= \frac{\mathbf{1}}{\mathbf{4}\boldsymbol{\tau_i}} \end{aligned} \tag{4.4}$$

and eventually normalized so that $\sum_j \boldsymbol{A(i,j)} = \mathbf{1}$. As explained earlier on, every cluster (segment) of $\boldsymbol{m^*}$ corresponds with one hidden state. In the initial emission structure we store the mean and the covariance matrix of the corresponding Gaussian cluster.

All initial parameters are set now and the HMM is ready for training using the Baum-Welch algorithm. Once the HMM is trained, an obvious solution [12] would be to select the basic gesture with the highest log-likelihood given the HMM. This straightforward solution might look attractive at first sight but it fails to handle the temporal variation in an appropriate way. This is because HMM's exhibit some degree of invariance to local warping of the time-axis [13].

We propose to calculate and to define for every basic gesture the most likely hidden state sequence using the Viterbi algorithm and to define the HMM key points where the hidden state transitions occur. However, as shown in Fig. 4.6, they suffer from positional and temporal variation and in order to solve that problem we apply DTW.

4.4.2 DTW

Our DTW approach consists of two steps. In the first step we calculate a reference signal, in the second step we align each basic gesture with that reference signal.

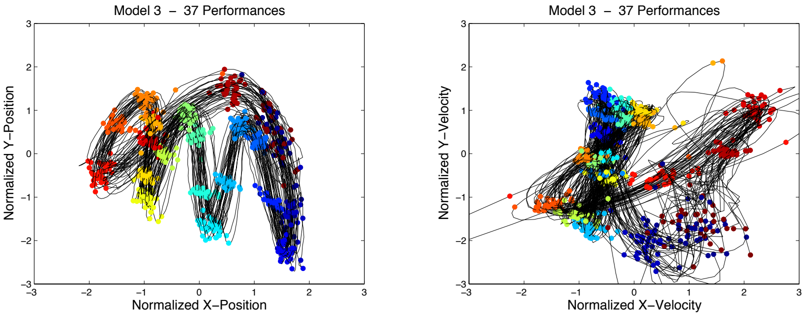


Figure 4.6: HMM Keypoints for all basic gestures and their positional variation

The procedure followed is basically the procedure from [10] but adapted to accommodate the temporal constraint that all gestures must complete in one measure. To incorporate temporal information from all gestures a new time vector is calculated: Firstly, we determine the average duration of every hidden state using key point information from all basic gestures and secondly, we use that information to fraction the period of one measure. It are these fractions that make-up the new time vector, which is referred to as the set of DTW key points timestamps.

Now, we use the basic gesture with the highest log-likelihood given the HMM and align this gesture to the previously produced time vector by linear temporal interpolation. This means linear stretching or shrinking of the corresponding state intervals. The resulting signal is the reference signal used for DTW.

Next, we warp all other basic gestures to this reference signal. DTW is preferred here over linear temporal interpolation as it handles the spatial distortion of the signals more efficiently [10].

The DTW procedure requires for every basic gesture a (dis)similarity matrix ($\boldsymbol{D}$). The task of DTW is to find herein an optimal path. Every element of the dissimilarity matrix ($\boldsymbol{D_{i,j}}$) is calculated as the Euclidean $\boldsymbol{l_2}$-norm between a 4D sample $\boldsymbol{i}$ of the reference signal $\boldsymbol{s_{ref}}$ and a 4D sample $\boldsymbol{j}$ of the basic gesture $\boldsymbol{s_{bas}}$ (See equation 4.5). Note that some authors recommend a shape preserving time constraint while calculating the optimal path [14].

$$\boldsymbol{D_{i,j} = \|s_{ref}(i) - s_{bas}(j)\|_2} \tag{4.5}$$

The similarity matrix is then used to find the sequence of pairs $\boldsymbol{(i,j)}$ forming a path along which the sum of distances $\boldsymbol{D(i,j)}$ is minimal. This path represents a time warping. A comparison of the original gestures and

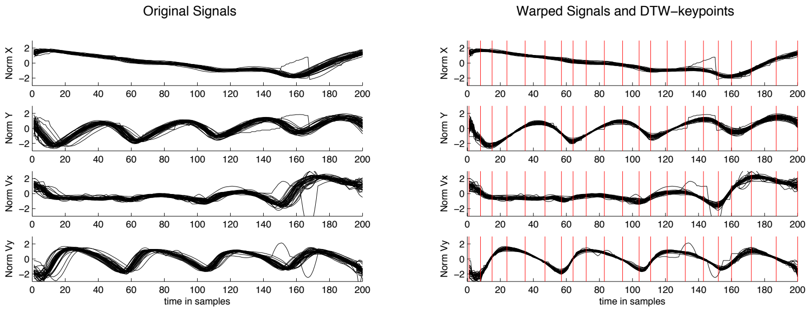


Figure 4.7: The original signals are shown in the left figure. The warped signals together with the time stamps of the DTW key points (red lines) are shown in the right figure (for conducting model 3).

the time warped gestures is displayed in Fig. 4.7. It is clear that the bundles from the warped signals are more compact what confirms that our procedure takes some of the variance away. The actual DTW implementation was done via a Matlab program from D. Ellis [15].

4.4.3 Generalized trajectory

As a result from DTW we have a set of time warped basic gestures and we have their values (DTW key points) at the newly created time vector. The time vector defines the (non-equidistant) knots for cubic spline regression and the DTW key points are input to the regression. The whole procedure is visualized in Fig. 4.8. Here the non-equidistant knots (time vector) are symbolized by a red line and the DTW key points are represented by blue dots. The resulting regression line is shown in black.

The regression lines for all coordinates make up the generalized trajectory of a beating-time gesture. This is presented for model 3 in Fig. 4.9. The red dots correspond here with the calculated time vector used for the DTW key points. The generalized trajectories of the other models can be found in Fig. 4.10.

4.5 Benchmarking

We benchmark the results of our method against two more methods. A first other method is where we produce a generalized trajectory directly from all basic gestures in the demonstration. For this *uniform* cubic splines (having equidistant knots) are used. A second other method uses a Gaussian Mixture Regression (GMR) [16]. We set the number of Gaussian compo-

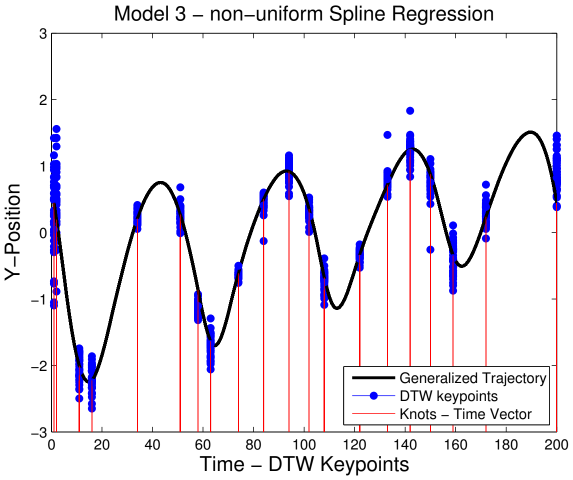


Figure 4.8: Spline Regression: knots are set at the calculated time vector. The DTW key points are the values from the warped basic gestures at this time vector. The knots and the DTW key points are input to cubic spline regression. The example shown here is for the y-coordinate and is for the generalized trajectory of subject 1 - model 3.

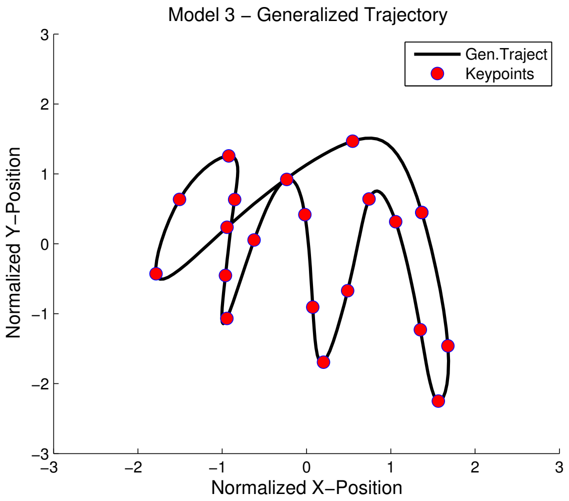


Figure 4.9: Generalized trajectory and its key points for conducting model 3.

nents in this method equal to the number of components discovered by our DPGMM. Eventually, we compare our proposed solution of a key point-based generalized trajectory with the two other methods in Fig. 4.11.

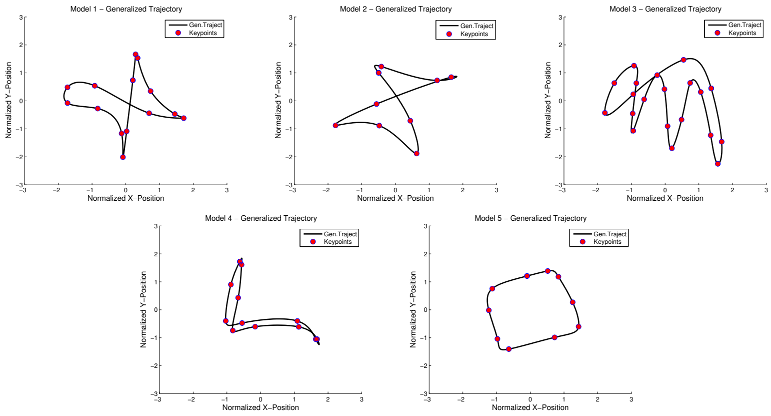


Figure 4.10: Generalized trajectory and key points of all conducting models.

The main difference is that the extrema are more pronounced for our keypoint-based method compared to the two other methods. The preservation of the extrema is due to the removal of the temporal variance by using DTW. This step should therefore be part of best practice [6].

A next topic is whether we can define quantitative performance indicators to benchmark these various solutions. This proves to be a difficult point.

In literature we often find the Root Mean Square Error (RMSE) as metric for benchmarking [6, 10]. RMSE evaluates how well a gesture $\boldsymbol{x}$ matches another gesture $\boldsymbol{y}$ using equation 4.6.

$$\boldsymbol{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} \|\boldsymbol{x}(\boldsymbol{i}) - \boldsymbol{y}(\boldsymbol{i})\|^{2}}{\boldsymbol{N}}} \tag{4.6}$$

The major concern with this metric is that it completely overlooks the temporal variation and that makes it useless for our application. An improvement, namely time warped RMSE, calculates the RMSE values not for the original but for the time warped basic gestures. Although this handles temporal variation there is now the additional question of what signal should be used as reference for DTW alignment. Selecting one or another reference strongly biases the RMSE results.

These issues make us consider other performance indicators that relate to the ultimate application. Beating-time gestures use the extrema to convey beat information, so preserving the extrema is an important performance indicator. Other indicators we propose, measure how suitable a

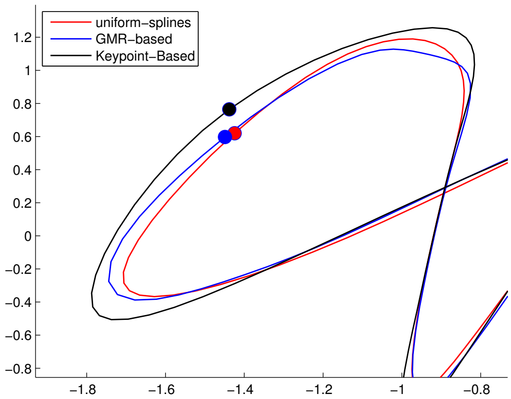


Figure 4.11: Benchmark of our proposed solution (Keypoint-Based) against an uniform-splines solution and a GMR solution for model 3. For the uniform-splines solution the knots are set equidistant, this opposed to our Keypoint-Based solution. The GMR-solution uses the same number of Gaussian clusters as discovered by our DPGMM method. For visibility reasons this figure zooms in on the top left part of the gesture. For convenience of the reader we added solid circles to all solutions indicating the position of the trajectories at the fourth beat. We notice that the Keypoint-Based solution excels in handling the temporal variation as it is better in preserving the extrema.

target trajectory is for a robot. Candidate indicators are jerk (derivative of acceleration) of robot movement and also the required on-line computation time.

4.6 An Application

Beating-time gestures indicate the musical beat, meaning that these gestures have temporal targets. This is a peculiarity not found in many day-to-day gestures. Therefore, to explain our application we firstly introduce the theory of goal points from Godøy. Godøy defines goal points as certain salient events in the music such as downbeats, or various accent types, or melodic peaks where sound-producing and sound-accompanying movements are centered [17]. Goal points link gestures with time and for beating-time gestures, the goal points of interest are the beat times. Note that the goal points are different from the previously discussed key points. Goal points relate to timing whereas key points reflect the shape of a trajectory.

The concept of goal points is useful for our application where we want to generate a sequence of beating-time gestures that fit to music. Fitting to music means adapting the gestures in terms of musical tempo and musical amplitude.

For the metronome performance the goal points are known as they coincide with the timestamps of the metronome ticks. For music the goal points must coincide with the beat points in the music. These beat points can for example be retrieved by some beat tracker program like BeatRoot [21]. Check McKinney [22] for an overview of beat tracker programs.

To adapt a generalized gesture to music we map the intervals between the goal points from the generalized gesture to the intervals made up by the beat points of the music (Fig. 4.12). This can be done by stretching and shrinking of the time intervals and the easiest way to achieve this is by linear temporal interpolation as is shown in Fig. 4.13.

This works well for positional data but for velocity data an additional step is required. Remember from (4.1) that a robot needs a target for position and velocity. For velocity data we do linear temporal interpolation as well but in addition all velocity values have to be changed proportionally to the stretch of the time interval. If the time interval doubles (i.e. music has a slower tempo than the metronome), the velocity should be set to half.

We recall that the generalized trajectory for our conducting gesture is made from a set of normalized performances. That makes the generalized trajectory also normalized and easily scalable. Scales can be chosen in accordance with musical amplitude.

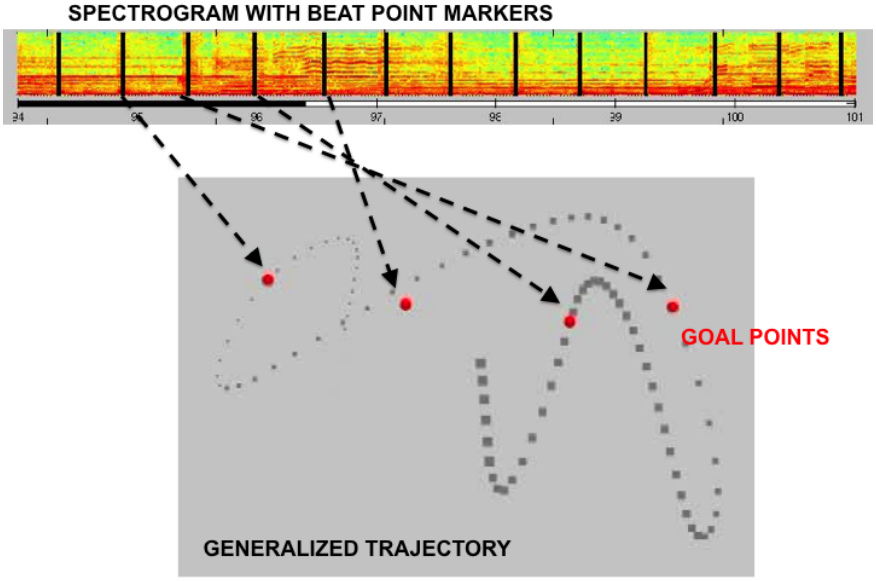


Figure 4.12: An Application: A beating-time gesture for music is made by mapping the goal points (beat times) of a generalized trajectory to the beat points of the music. This is achieved by linear temporal interpolation. The time progress bar shows the actual time stamp of the music and the actual position of the gesture (between beat two and three). This operation changes the run-through speed of the generalized gesture. Additionally the amplitude of the generalized gesture can be changed in accordance with the musical amplitude.

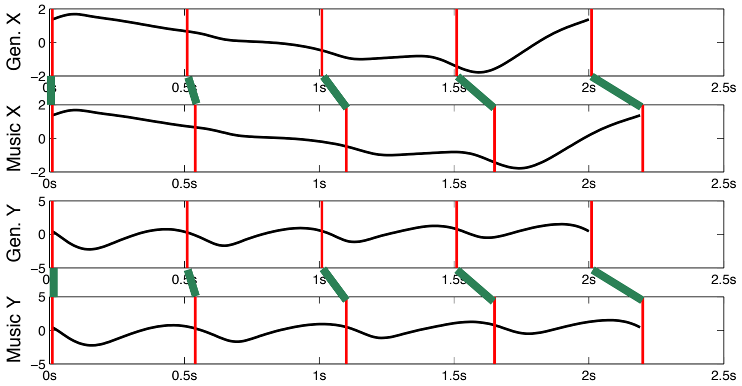


Figure 4.13: Making a beating-time gesture for music (Music X,Music Y) from a generalized trajectory (Gen X,Gen Y). The general trajectory is labeled Gen X for its X-coordinate and Gen Y for its Y coordinate, the synthesized gesture Music X and Music Y. All horizontal-axes express time in seconds. The procedure maps the goal points (metronomic ticks) of the generalized trajectory onto the goal points (beat points) of the music. The goal points are visualized by vertical red lines, the mapping by green lines. The synthesized beating-time gesture (Music X,Music Y) is calculated by linear temporal interpolation.

4.7 Discussion

Our solution is in essence a dynamic time warping solution that focuses on the construction of a reference signal for time warping. The reference signal holds temporal information coming from all basic gestures and is calculated from the hidden state transitions of a fitted HMM. This calculation involves an averaging step and as such it is outlier sensitive. Care should be taken to remove outlying basic gestures prior to the analysis.

The whole procedure has quite some similarities to methods used for speech recognition and speech synthesis. As such it follows the ideas of a vocoder being an analysis/synthesis system, used to reproduce human speech.

For analyzing speech, HMMs are the de-facto standard. Using HMMs might look like an interesting alternative to our solution but in our setting it has also some disadvantages. Most HMM based applications require large data sets for training. For speech recognition this is no problem but for robot PbD it renders the solution infeasible. We want to work there with a low number of demonstrations. Hence for the modeling phase, we we inserted prior knowledge in the HMM [19] choosing a topology with few parameters.

Additionally, an HMM does not perform well for synthesis since the duration model (hidden state self transitions) is rather simplistic [18]. In this case a better duration model is required and we do no longer talk about a HMM but about a HSMM, a Hidden Semi-Markov Model.

The location of the goal points for our calculated generalized trajectories contradicts the intuitive understanding of a conducting gesture that most of us have. Most people anticipate the beats to occur at the extremities of the conducting gesture movement. For example for model 1 this is at the top, bottom, left and right position. Our research learns however that there is a lag of approximately 0.25s (compared to the 2s for the bar) between these positions and the actual beat points. Although our study was limited to four subjects and generalization is impossible, this result is in line with a previous study from Luck and Toiviainen [4]. In their study Luck and Toiviainen found that an ensemble's performance, executed in an ecological setting with a conductor, tended to be most highly synchronized with periods of maximal deceleration along the trajectory, in second place followed by periods of high vertical velocity.

Main criticism of our method is that the synthesized beating-time gesture is not human. During the production process of a generalized trajectory we focused on timing, resulting in an artificial trajectory, rather than on the human factor, what would mean selecting one performance out of a set. Our artificial gesture was eventually *humanized* by making it smooth through cubic spline regression.

4.8 Conclusion and Future Work

The present study provided a method to produce a synthesized beating-time gesture for use with a humanoid robot. From a set of demonstrated beating-time gestures on metronome ticks a generalized trajectory in parametric form was calculated. As such the followed method is called programming by demonstration (PbD) or imitation learning. The calculation used two probabilistic models namely a DPGMM and a HMM together with a DTW algorithm to cope with the spatial and temporal variation of the demonstrated conducting gestures. Using the concept of goal points (temporal targets) it was easy to adapt the generalized gesture, to make it suitable for music.

Our work is an initial but important step towards a fully automated conducting system. Our present implementation is now limited to beating-time gestures. A next step could be to move to a more extended set of gestures. Our system is still off-line: We extract the beat points off-line and up-front and we use them to generate a synthesized beating-time gesture also off-line and up-front. Moving from an off-line beat detection algorithm to an on-line beat detection algorithm would make it possible for a conductor to adapt the timing of his gestures to what the orchestra is actual playing. We suggest using an adaptive learning approach, based on a maximum a posteriori (MAP) estimation, and integrating the propagated knowledge from previous time intervals.

References

[1] M. Leman and R.I. Godøy. *Why Study Musical Gestures?* In Musical gestures: Sound, movement, and meaning, chapter 1, page 5. Routledge, 2009.

[2] G. Johannsen and T.M. Nakra. *Conductors' Gestures and Their Mapping to Sound Synthesis.* Musical Gestures: Sound, Movement, and Meaning, page 264, 2009.

[3] G. Luck, P. Toiviainen, and M.R. Thompson. *Perception of Expression in Conductors' Gestures: A Continuous Response Study.* Music Perception, 28(1):47–57, 2010.

[4] G. Luck and P. Toiviainen. *Ensemble musicians' synchronization with conductors' gestures: An automated feature-extraction analysis.* Music Perception, 24(2):189–200, 2006.

[5] P.-J. Maes, D. Amelynck, M. Lesaffre, M. Leman, and DK Arvind. *The "Conducting Master": An interactive, real-time gesture monitoring system based on spatiotemporal motion templates.* International Journal of Human-Computer Interaction, 2012.

[6] S. Calinon, F. D'halluin, E. L. Sauser, D. G. Caldwell, and A. G. Billard. *Learning and reproduction of gestures by imitation.* Robotics & Automation Magazine, IEEE, 17(2):44–54, 2010.

[7] M. Leman and L. Naveda. *Basic gestures as spatiotemporal reference frames for repetitive dance/music patterns in Samba and Charleston.* Music Perception, 28(1):71–91, 2010.

[8] Y.W. Teh. *Dirichlet Process.* Submitted to Encyclopedia of Machine Learning, 2007.

[9] K. El-Arini. *Dirichlet Process : A gentle tutorial.* Select Lab Meeting, 10 2008.

[10] A. Vakanski, I. Mantegh, A. Irish, and F. Janabi-Sharifi. *Trajectory Learning for Robot Programming by Demonstration Using Hidden Markov Model and Dynamic Time Warping.* Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 42(4):1039–1052, 2012.

[11] J. Aleotti and S. Caselli. *Robust trajectory learning and approximation for robot programming by demonstration.* Robotics and Autonomous Systems, 54(5):409–413, 2006.

[12] S. K. Tso and K. P. Liu. *Demonstrated trajectory selection by hidden Markov model.* In Robotics and Automation, 1997. Proceedings., 1997 IEEE International Conference on, volume 3, pages 2713–2718. IEEE, 1997.

[13] C. M. Bishop et al. *Pattern recognition and machine learning*, volume 4. springer New York, 2006.

[14] H. Sakoe and S. Chiba. *Dynamic programming algorithm optimization for spoken word recognition.* Acoustics, Speech and Signal Processing, IEEE Transactions on, 26(1):43–49, 1978.

[15] D. Ellis. *Dynamic Time Warp (DTW) in Matlab.*

[16] S. Calinon, Guenter F., and Billard A. *On Learning, Representing and Generalizing a Task in a Humanoid Robot.* IEEE Transactions on Systems, Man and Cybernetics, Part B, 37(2):286–298, 2007.

[17] R.I. Godøy, A.R. Jensenius, and K. Nymoen. *Production and perception of goal-points and coarticulations in music.* Journal of the Acoustical Society of America, 123(5):3657, 2008.

[18] S. King. *An introduction to statistical parametric speech synthesis.* Sadhana, 36(5):837–852, 2011.

[19] S. Calinon and A. Billard. *Stochastic gesture production and recognition model for a humanoid robot.* In Intelligent Robots and Systems, 2004.(IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on, volume 3, pages 2769–2774. IEEE, 2004.

Without music, life would be a mistake.

F. Nietzsche

5

The Surprising Character of Music.

A search for sparsity in music evoked body movements.

Abstract

The high dimensionality of music evoked movement data makes it difficult to uncover the fundamental aspects of human music-movement associations. However, modeling these data via Dirichlet Process Mixture (DPM) Models facilitates this task considerably. In this paper we present DPM models to investigate positional and directional aspects of music evoked bodily movement. In an experimental study subjects were moving spontaneously on a musical piece that was characterized by passages of extreme contrasts in physical acoustic energy. The contrasts in acoustic energy caused surprise and triggered new gestural behavior. We used sparsity as key indicator for surprise and made it visible in two ways. Firstly as the result of a positional analysis using a Dirichlet Process Gaussian Mixture Model (DPGMM). Secondly as the result of a directional analysis, where a Dirichlet Process Multinomial Mixture Model (DPMMM) unveiled a dominant direction mix for the low energetic acoustic parts but random directional behavior in the high energetic acoustic parts. The results show that gestural response follows the surprising or unpredictable character of the music.

5.1 Introduction

Several authors have suggested that humans perceive something as aesthetically interesting when there is a balanced mixture between recognition and surprise [1]. In 1933, Birkhoff was one of the first to present a mathematical theory for aesthetic measures, which he defined as the ratio of O (order) to complexity (C) [2]. The idea that surprise is related to aesthetic feeling fully resonates with known theories of music processing and emotional arousal [3–5].

Surprise is often intended and in music it has a strong power to arouse listeners. Mayer [6] drew an analogy between musical structures and recurrence structures in chaotic systems. He stated that : "Perceived order and disorder, recurrence and complexity are common features observed in both chaos and music. These features can be perceived in music because the music has been intentionally designed to reveal them." An extreme example is the famous Symphony No. 94 in G major (Hoboken 1/94) written by J. Haydn, also known as the Surprise Symphony. Haydn was reputed for this type of surprises, and the Surprise Symphony is exemplary in that it contains a sudden fortissimo chord at the end of a piano opening theme in the variation-form second movement. The music then returns to normal and subsequent movements do not repeat the surprise. And this brings us to a key indicator of surprise and that is sparsity. Sparsity is a major attribute of many descriptions of surprise (e.g. [5, 7–9]).

Based on the key insights that cognition is necessarily situated and embodied [10, 11] we assume that the surprising character of the music gets embodied in the movement idiosyncrasies of subjects. A cognitive system, such as the human mind, is always interacting with its environment via its sensors that perceive, and effectors that produce actions. The complexity of the real world is dealt with not by manipulating abstract internal representations, but by interacting with the world itself, i.e. by performing actions and monitoring their results via perceptions. There are theories that state that music perception is built on a bidirectional action-perception coupling [12]: "In one direction, incoming sensory information is transformed into corresponding motor representations on the basis of a direct-matching or mirroring [13]. It explains why so many people tend to move along with the expressive patterns they hear in music. In the other direction, sensory outcomes are predicted based on planned or executed actions [14]. This explains why the perception of ambiguous musical patterns can be influenced by movements, as movements prompt people to impose - at least temporarily - certain anticipated structures (e.g., rhythm, melody, dynamics, phrasing etc.) or affective qualities onto the music. The two directions are coupled in the sense that the mere activation of one representation (action or perception) results in the activation of the other (perception, action)."

Given the tight coupling between perception and action, musical surprises can be called gestural affordances [15]. For listeners and dancers, these surprises, or failures to anticipate, afford new opportunities to move along with the music [15, 16]. The movement idiosyncrasies are assumed to result from two control mechanisms which we call the action-perception loop and the sensory-motor loop [17]. Fig. 5.1 shows the overview taken from [17] page 27. The action-perception loop is rooted in an action repertoire that contains previously learned associations between perception and action. For example, it contains representations that associate perceived sounds with gestures, and it also contains representations that allow the generation of actions in function of desired perceptive outcomes in response to music. The sensory-motor loop is rooted in the environment and it allows corrections of executed actions in response to input from the environment. The hypothesis is that music offers affordances with which listeners interact on the basis of their action repertoire and interaction with the environment.

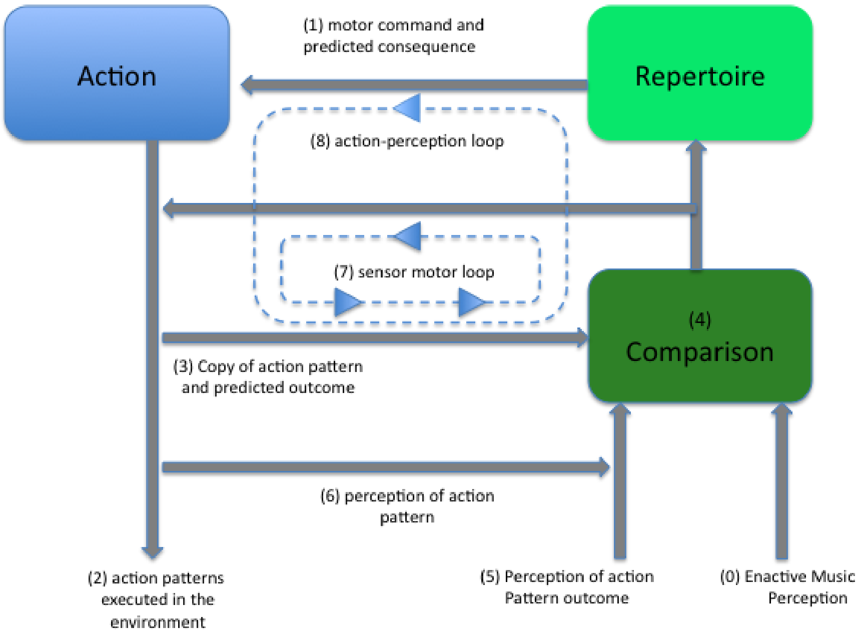

Action
(1) motor command and predicted consequence
Repertoire
(8) action-perception loop
(7) sensor motor loop
(4) Comparison
(3) Copy of action pattern and predicted outcome
(6) perception of action pattern
(2) action patterns executed in the environment
(5) Perception of action Pattern outcome
(0) Enactive Music Perception

Figure 5.1: Schema for an action-perception coupling system.

The goal of the present paper is to provide a set of methods to identify sparse expressive behavior, hence the link with the unpredictable and surprising character of music. The provided methods are objective, analytically tractable and keep a direct link with time as music is a function over time. Just like a Short-Time Fourier Transform allows to investigate a basic movement characteristic like periodicity, our methods also focus on basic movement characteristics, being position and direction. Position and direction are known to have strong relationship with musical valence and arousal (e.g. [18] [19]). Position is the coordinate of an optical marker attached to the body of a subject and is measured either in an absolute or in a relative reference coordinate system. Direction reflects the direction this optical marker is moving (left/right and up/down) and is here expressed as a mixture over a time interval (viz. it quantifies concepts like more up/down than left/right movement).

The paper is organized as follows. In section 5.2 we describe the experiment that is at the basis of our research. Section 5.3 describes the methods of analysis used: how to preprocess the data and what feature space to use. The results are in section 5.4, a section that is split-up in two parts: the first part handles the results for positional analyisis, the second part concerns directional analysis. Eventually a discussion can be found in section 5.5.

5.2 Experimental set-up

At the basis of our research lays an experiment in which subjects moved spontaneously on music. The music for the experiment was selected for its extreme contrasts in physical acoustic energy, symbol for the surprising character of the music.

- *Subjects + Task.*

  Thirty-six subjects participated in a music evoked body movement experiment. They were chosen from a pool of students enrolled in various academic disciplines and they volunteered freely. The group was composed of 20 males and 16 females with a mean age of 24,2 year (SD=4,2). The experiment was set-up on a per individual basis, having one subject performing at a time. Before the actual execution of the experiment, the participants received the task of moving spontaneously to the music. This was formulated as: “Translate your experience of the music into free full-body movement. Try to become absorbed by the music that is presented and express your feelings into body movement. There is no good or wrong way of doing it. Just perform what comes up in you.” The actual *motor-attuning* experiment took place in a motion capture space: an octagonal space enclosed by black curtains in order to separate the participants from the experimenters. The participants could thereby use the space indicated by a white, round carpet with a diameter of 4 meter. Furthermore, we made the room completely dark, as a pilot study had indicated that this made the participants more comfortable and less constrained to execute their task. The music was played through a stereo setup formed with two Behringer B2031A Truth Active Studio Monitors at a predefined volume which was esteemed as agreeable by the experimenters.

- *Stimuli.*

  The music was part of Johannes Brahms' *First Piano Concerto*, Opus 15 in D minor from 1858 (in a recording by Krystian Zimmerman and the Berlin Philharmonic Orchestra, conducted by Simon Rattle). The musical piece is characterized by passages articulating extreme contrasts in physical acoustic energy. Based on this, we define two contrasting musical style categories which structure the main outline of the composition, namely a Heroic and Lyric style category. In the stimulus three Heroic passages are presented in alternation with three Lyric passages. Because the first Lyric passage is relatively long in comparison with the other we deleted some portion of that passage (1 min 56 s - 2 min 46 s of the recording) in a way it was not audible for people that do not know the musical piece well. The remaining musical stimulus had a duration of 5 minutes and 12 seconds (Fig. 5.2).

- *Data recording.*

  Registration of movement data for the complete upper body was realized with an OPTITRACK infrared optical system consisting of 12 synchronized cameras with related ARENA motion capture software (`http://www.naturalpoint.com`). Participants were asked to put on a special jacket and cap on which markers were attached with Velcro. A default human upper body skeleton model provided in the ARENA software was constructed from the 22 infrared reflecting markers that were attached to jacket and cap in a predefined manner: four markers for hip, three markers each for head, chest, upper arms, and hands. Afterwards, the performances of all participants were exported into BioVision Hierarchy (BVH) files. With the help of the MATLAB motion capture toolbox (`http://www.cs.man.ac.uk/~neill/mocap`), we calculated the three-dimensional position (at a sample rate of 100 Hz) for all "joints" making up the upper body and head. Although we collected data from multiple markers, we focused for this analysis on the data of the right hand.

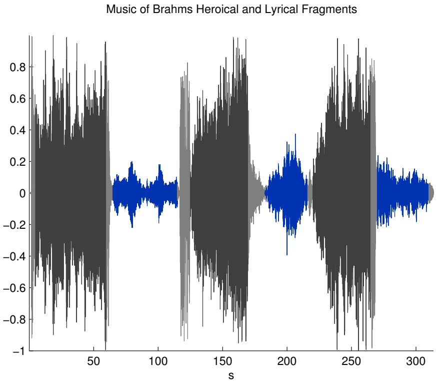


Figure 5.2: Johannes Brahms Opus 15 in D Minor - subdivided in Heroic Parts with high acoustic energy and lyric Parts with low acoustic energy.
(Dark Grey = Heroical Parts. Dark Blue = Lyric Parts. Light Grey = Transitional or Not classified)

5.3 Analysis Method

The analysis employs clustering techniques. For the positional analysis we fit a Dirichlet Process Gaussian Mixture Model (DPGMM) to the data. This model divides the feature space in a number of Gaussian clusters with the particularity that the number of clusters is learned from the data. For the directional analysis a Dirichlet Process Multinomial Mixture Model (DPMMM) is applied. This models the feature space as a mixture of (directional) mixtures. For readers not familiar with Dirichlet Process Models we refer to the existing literature (see e.g. Teh [20] and El-Arini [21]) or for a brief introduction to appendix A.1. Let us first explain how we preprocessed the data and how we set-up the respective feature spaces.

5.3.1 Pre-processing of the data

The data from one subject were discarded due to technical problems during the recording.

The main focus of the experiment was on the movement data from the hand as it is a body part that experiences the highest degree of freedom (DOF). To eliminate the influences from other body parts (like translations and rotations of torso and/or shoulders) a new three dimensional axis system was defined as in Fig. 5.3: Axis 1 was defined as the line going through the clavicle (shoulder-neck). Axis 2 was defined by the projection of the upposition (chest-neck) onto a plane perpendicular to axis 1 and eventually axis 3 was determined as the cross product of axis 1 and axis 2. All calculations and findings in this paper are based on this new relative coordinate system.

Eventually the positional data was translated to a new origin being the mean of the data points of one subject. This step served no specific goal except that it was required for the chosen implementation of the DPGMM algorithm.

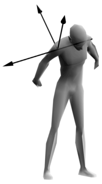


Figure 5.3: Relative Axis System used for hand representation.

The data for the directional analysis is based upon the velocity signals, which were calculated as derivatives from the positional data. To calculate these derivatives a local (linear) *derivation* filter was applied to the positional data. The size of the filter window was set to a value of 0.175 s corresponding with a linear frequency response of the derivation filter in the useful frequency band of 0-4 Hz . The 0-4 Hz range is in line with the information coming from the spectrogram in Fig. 5.4. This derivation filter was applied to all coordinates to calculate the first derivative. The sole purpose of this filter was to remove noise while calculating the derivatives.

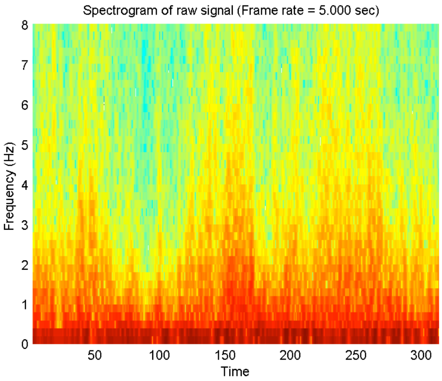


Figure 5.4: Spectrogram of hand positional data

5.3.2 Feature Space

The feature space for positional analysis consists of the positional coordinates (Cartesian coordinates x-y-z). The feature space for directional analysis is calculated from the velocity vectors (section 5.3.1). Velocity vectors were first converted to spherical coordinates (radius, elevation and azimuth). Afterwards, the spherical coordinates were used to categorize directional information. Categorization was done with the help of the following two indicators: A first indicator came from the elevation ($[\frac{-\pi}{2}, \frac{\pi}{2}]$) and divided the elevation range in four quadrants of $\frac{\pi}{4}$. A second indicator was derived from the azimuth ($[-\pi, \pi]$) dividing its range in eight octants of again $\frac{\pi}{4}$. The combination of these two indicators resulted in total in 32 (4 times 8) categories. In addition, we created one extra category labeled "lack of movement". The criterion for lack of movement was based on low speed as indicated by the radius of the velocity vector. The decision border for low speed values was set per subject in such a way that 5% of the values having the lowest radius values would always be categorized as lack of movement. Eventually every velocity sample got assigned to a category. The category information made up the feature space for the directional analysis.

Because of the degree of randomness or should we say chaos [22] in music evoked body movement, we do not look at directional data at distinct time stamps but at mixtures of directional categories over a limited time interval. We are rather interested in the amount of up-down movements or left-right movements in a particular time interval.

To calculate the mixtures we use time intervals of three seconds conform Pöppels' theory of the 3 s window of temporal integration [23]. At a sample rate of 100 Hz this means that we have 300 samples (3*100) of the velocity vector for every 3 s time interval. Every single sample of these 300 samples is then classified into one of the 33 classes as we explained earlier on. The entire 3 s time interval is then described by its mix of classes. This mix is modeled as a sample of a multinomial distribution with N=300 and 33 categories (classes). Our musical fragment lasts approximately 5 minutes and 12 seconds and can be divided in 104 of these 3 s time intervals. To avoid artifacts we work with an overlap window of 50% to end with 208 samples of multinomial distributions. DPMMM clusters these multinomial samples by assigning samples that have a large probability to originate from the same multinomial distribution to the same cluster. Eventually, the result of the DPMMM analysis is a mixture (clusters) of mixtures (directional classes).

5.3.3 Dirichlet Process Models

For the analysis we use Dirichlet Process Mixture (DPM) models (See also Appendix A.1). By using a Dirichlet Process mixture (DPM) model we target at modeling the dataset with a limited number of clusters and still preserving its most important features. DPMs have the advantage that they learn the number of clusters from the data. This is in contrast with algorithms like K-means clustering or Gaussian Mixture Models (GMMs) where the number of clusters has to be specified upfront or has to be determined by additional validation steps. The practical implementation was done in Matlab with the help of the demo programs from Yee Whye Teh (http://www.stats.ox.ac.uk/∼teh/) .

5.4 Results

In what follows, we first discuss the results for the positional analysis using DPGMMs. Then we discuss the results for the directional analysis using DPMMMs.

5.4.1 DPGMM for Positional Analysis

The musical excerpt of Brahms was split-up in six fragments, namely, three heroic style fragments alternating with three lyric fragments as illustrated in Fig. 5.2. The DPGMMM analysis was performed for every combination of subject and fragment. By way of example Fig. 5.5 shows the data points collected for subject 2 fragment 2 (a lyric fragment) on the left hand side. The result of the DPGMM clustering (same subject, same lyric part) is displayed on the right hand side. We see that for this particular case the movement in terms of position can be described by a three clusters system. Every cluster stands for a three dimensional Gaussian and is uniquely defined by its mean and its 3x3 covariance matrix. In other words, the model allows to describe the dataset for subject 2 fragment 2 containing 5000 data samples by only 26 parameters (three means, three covariance matrices and two parameters for the cluster mixture). All subsequent analyses are based on these model parameters.

5.4.1.1 Analysis of small data clusters

Small clusters stand for sparse movement and therefore they might link to surprising, salient events in the music. For our analysis we defined a small cluster as a cluster having a maximum of 300 data points. This corresponds to a three second time interval if all the data points in that cluster are

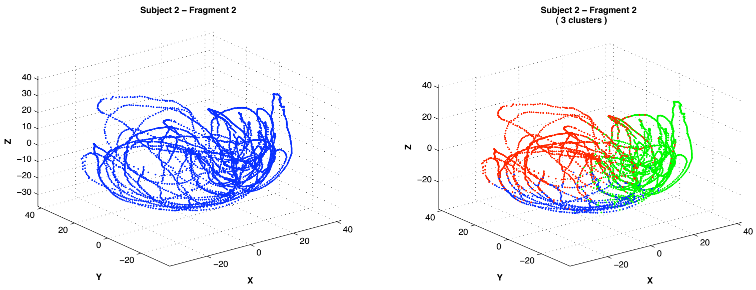


Figure 5.5: Hand Movement : a. Raw Data and b. Clustered Data

adjacent in time. Our assumption is that because these points are close to each other in space (belonging to the same cluster) they are also close to each other in time, as human movement is continuous and smooth. That means that most of these small clusters represent small abnormal moves.

An interesting question is whether subjects did make of these abnormal movements at the same time. That would point to some effect in the music that triggers these sudden (surprising) events. Fig. 5.6 summarizes the findings and depending on the threshold used we see that there are four moments in time where five subjects or more concurrently made such short abnormal moves. These moves took place at 5.2 s - 57.1 s - 102.4 s - 300.6 s. For 5.2 s (warm-up?) and 57.1 s we find no obvious explanation in the music but intriguing is that we notice a similar event at 102.4 s and 300.6 s. There we localize a change in the harmonic structure of the music with a major cord (happy) changing into a minor chord (sad). This happens at time stamp 102.4 s in the orchestra part and at 300.6 s in the solo part for piano.

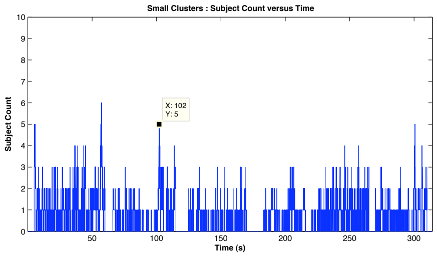


Figure 5.6: Small Clusters : Subject Count. The bars represent how many subjects had their positional coordinates in a small cluster at a particular time.

5.4.1.2 Analysis of large data clusters

To understand sparsity we must also understand what is common. Therefor it is instructive to study the large clusters as well. We defined clusters as large clusters if they contained at least 10 % of the data points of a particular fragment. Let us first study the positions of the means of these large clusters, and next we investigate their covariance matrices.

In our experiment, the means (positions) of these clusters can be influenced by two conditions (heroic or lyric). A first analysis checks the position of these clusters with respect to the planes of motion namely the sagittal plane (medial-lateral), the frontal plane (anterior-posterior), and the transverse or horizontal plane (superior-inferior). For every subject we calculate the average cluster location in the lyric and heroic condition. Paired T-tests reveal then the following:

- On average the cluster locations in lyric fragments (M=0.95, SE=0.40) are significantly more medial compared to the clusters in heroic fragments (M=-1.93, SE=0.53) t(34)=-4.48, p<.05, r=.6. The paired difference between lyric and heroic clusters for the average cluster distance to the sagittal plane is significantly normal (D(35)=0.09, p>.05). This means that lyric movements happen more in front of the body (or towards the heart) compared to heroic movements that happen more aside.

- On average heroic clusters (M=5.69, SE=0.57) lay significantly higher than lyric clusters (M=-3.96, SE=0.63) t(34)=9.72, p<.05, r=.9. The paired difference in height between the two cluster types is significantly normal (D(35)=0.10, p>.05).

- On average heroic clusters (M=0.89, SE=0.37) lay significantly closer to the frontal plane than lyric clusters (M=-1.38, SE=0.37) t(34)=3.86, p<.05, r=.6. The paired difference in this position between the two cluster types is however not significantly normal (D(35)=0.18, p<.05). The latter result despite its non-normality may seem surprising at this point. We expect lyric movements closer to our body and heroic movements further away. The main reason why this result seems not to match our intuition is because of the definition of close. We defined “close” as the distance to the frontal plane. If you stretch your arm and raise your arm straight-up, your hand lays in the frontal plane and the distance to the frontal plane is zero. So distance to the frontal plane has nothing to do with the intuitive concept of close to the body.

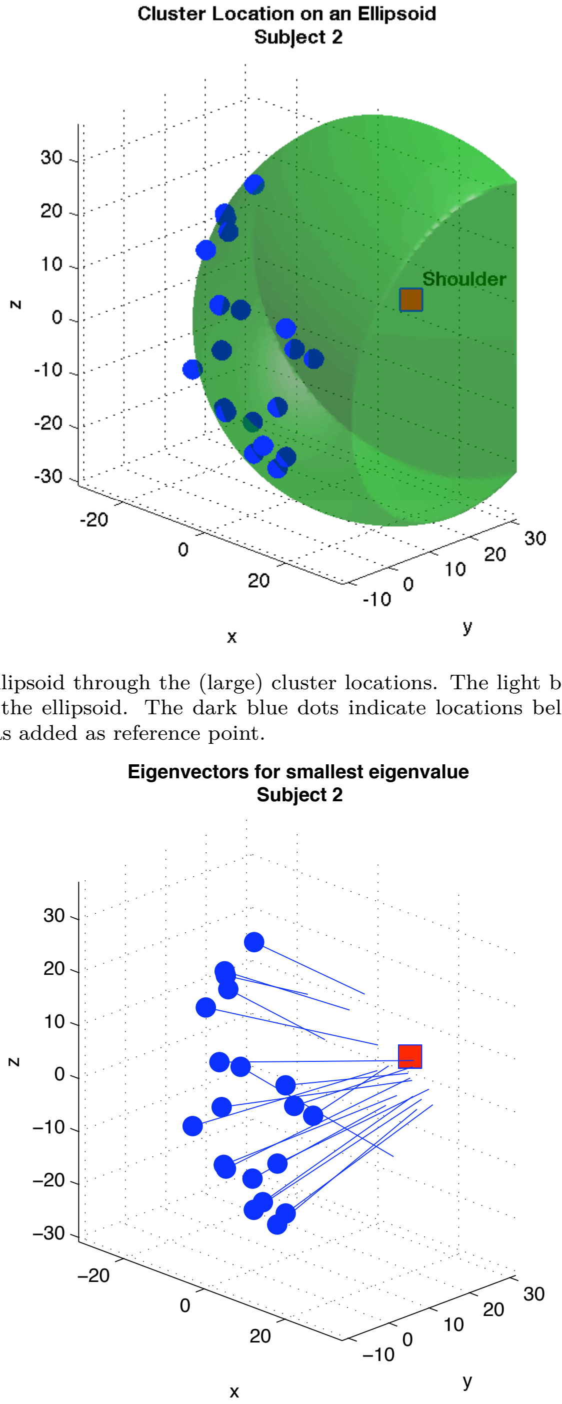
Visual inspection learns us that the movement mainly happens on the surface of an ellipsoid. We used a least square fitting algorithm made available by Dr. Y. Petrov (Northeastern University, Boston, MA) to fit an ellipsoid through the cluster locations. Unfortunately the cluster locations are concentrated on a limited area of the ellipsoid surface what makes the fitting algorithm not always successful for all subjects. Nevertheless for some subjects, like subject 2, the fit is quite good as can be understood from Fig. 5.7(a). If movement happens on the surface, then we need only two coordinates to specify it. According to the mathematical definition of dimension, this movement is then two dimensional and not three dimensional. It is our hypothesis that surprises in music cause movement to suddenly enter a higher dimension. However with the fitting problems of ellipsoids other methods have to be sought to back-up this hypothesis.

After investigating the cluster means, we can have a look now at the cluster covariance matrices. For every covariance matrix we calculate the eigenvalues and the eigenvectors. The size of a cluster can be approximated by the volume of an ellipsoid with semi-principal axes (a,b,c) set equal to the square root of the eigenvalues as shown in Formula (5.1). Note that calculating the volume using eigenvalues makes the volume a measure for the density of a cluster. High volumes are then synonym for low density.

$$\boldsymbol{V = \frac{4}{3} * \pi * a * b * c} \tag{5.1}$$

On average heroic clusters (M=2458.38, SE=189.41) have a significantly higher volume than lyric clusters (M=1589.92, SE=176.24), t(34)=8.49, p<.05, r=.8. The difference in volume of the clusters is significantly normal distributed (Shapiro-Wilk) W(35) = 0.97, p>.05. What this says is that movement in lyric fragments is more concentrated (denser) than movement in heroic fragments. Additionally we see that one eigenvalue is always considerably lower than the two other eigenvalues. Averaged per subject we see that the smallest eigenvalue explains only about 8% of the variance (M = 8.10, SD=1.20). In other words the movement of the hand is locally (centered at the cluster) rather two dimensional rather than three dimensional. The orientation of this two dimensional plane can be visualized by looking at the orientation of the eigenvector that goes together with the smallest eigenvalue as the two dimensional plane is perpendicular to this eigenvector. In Fig. 5.7(b) we see that the direction of these eigenvectors points to a central point near the shoulder. Variance in that direction corresponds with punching movements (from the body away and back). As this is the direction with the lowest variance (smallest eigenvalue) we can say that this type of movement was almost absent in our experiment.

Summarized, large cluster research revealed that hand movement mainly happened on two dimensional manifolds and that these manifolds are tan-



(a) Fitting an ellipsoid through the (large) cluster locations. The light blue dots indicate locations above the ellipsoid. The dark blue dots indicate locations below the ellipsoid. The shoulder was added as reference point.

(b) Direction of the Eigenvectors corresponding with the smallest eigenvalues of the (large) clusters. The shoulder was added as reference point.

Figure 5.7: Both figures explain that hand movement mainly happens on two dimensional manifolds and that these manifolds are tangent to the surface of an ellipsoid.

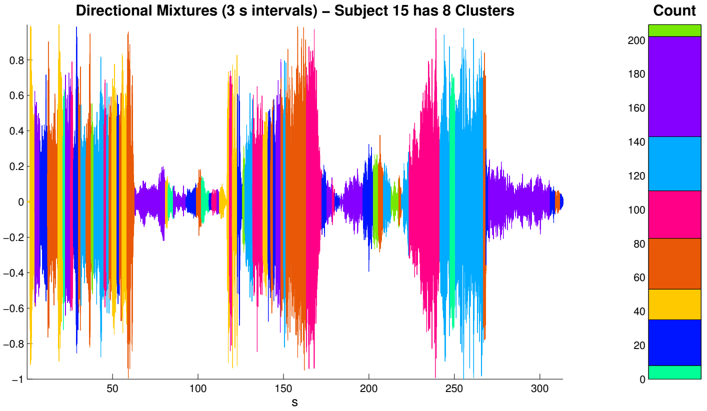


Figure 5.8: Cluster assignment for subject 15, visualized on top of the musical amplitude. Each cluster stands for movement with the same directional mix.

gent to the surface of an ellipsoid. Any movement that violates these "rules" can be classified as sparse and is a possible indicator of surprise.

5.4.2 DPMMM for Directional Analysis

In section 5.3.1 we explained how to calculate the directional mix (viz. the amount of left/right and up/down movement) for every three second interval. We did the calculation for in total 208 of these three second intervals allowing them to overlap by 50%. All the calculated directional mixtures can be understood as samples coming from many multinomial distributions. The task of a DPMMM is to cluster mixtures that are likely to come from the same multinomial distribution. Fig. 5.8 shows the clustering result for subject 15. As our interest lays in the relationship with music, the cluster assignment is displayed with the help of the musical amplitude.

For this particular subject, DPMMM assigned the intervals to in total 8 different clusters. This means that the subject's movement style can be reduced to 8 different ways of moving (direction-wise). We further notice that the lyric style fragments are dominated by the same cluster what basically tells us that lyric style fragments have the same directional mix. Heroic intervals on the contrary have a less distinct dominant cluster and show also different cluster patterns. In particular, the first and second heroic fragment show what we call sparse behavior. The clusters alternate there in a fast sequence. The third heroic interval however does not show this behavior and gives the impression that the subject is not anymore surprised by the music.

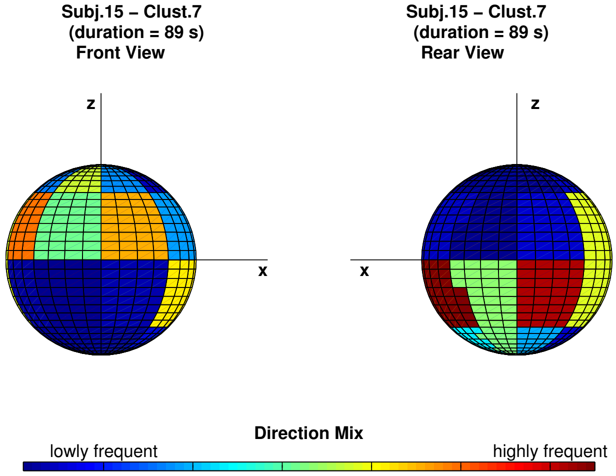


Figure 5.9: Visualization of the directional mix for Subject 15 Cluster 7 (the dominant cluster of Fig 5.8). The prevailing direction is horizontal.

We visualize now for subject 15 the directional mix of his/her dominant cluster in Fig. 5.9. There the red color stands for dominant directions and the blue color for directions that are not present. From this figure we learn that (for this subject) the lyric intervals are dominated by a prevailing horizontal (left/right) movement.

The above findings were based upon the results for a single subject. Can we generalize some of the results? A paired T-test confirms that the time fraction of the dominant cluster is indeed higher for lyric intervals ($\boldsymbol{M = 0.56, SE = 0.03}$) than for heroic intervals ($\boldsymbol{M = 0.42, SE = 0.02}$) $\boldsymbol{t(34) = -5.23, p < .05, r = .7}$. The paired difference in time fraction between the two styles is significantly normal ($\boldsymbol{D(35) = 0.19, p > .05}$).

Another way of consolidating is to bundle the results of all subjects in a single diagram, in a what we call a *directogram*. A *directogram* is a visual representation of the gestural affordances (direction-wise) in a musical excerpt. It represents a square matrix calculated as follows: If for example interval 17 and interval 24 belong for one subject to the same cluster we increase the value of element (17,24) of the square matrix by one. We loop then over all subjects and display the resulting matrix in a kind of density plot (Fig. 5.10). This plot is what we define as a "*directogram*".

Warmer colors indicate that more subjects were moving their hand similarly (intra-subject!) at the timestamps given by the horizontal and vertical index. The directogram learns that the musical excerpt is splitting into five

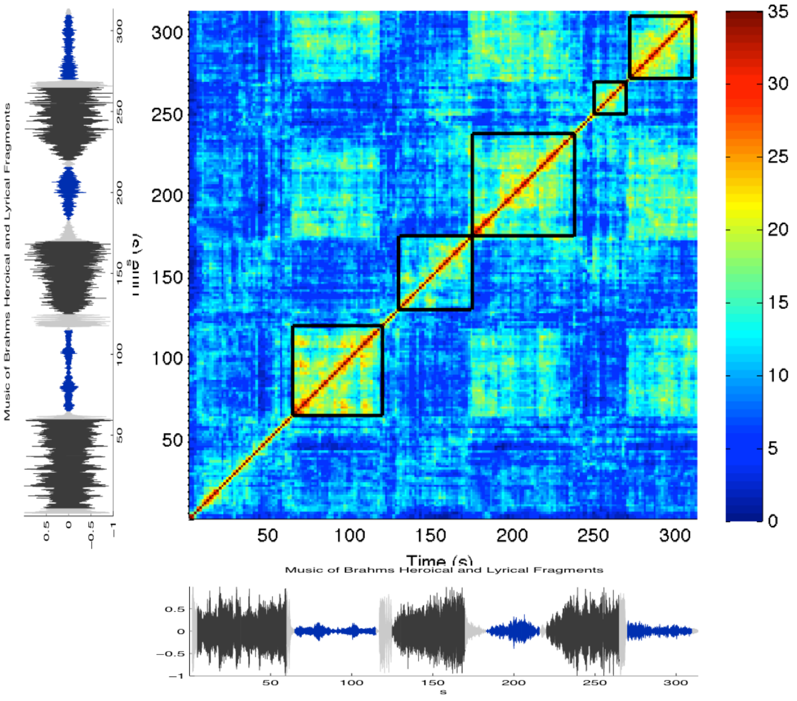


Figure 5.10: Directogram: reveals 'directional movement' characteristics of a musical excerpt. Persistency, along the diagonal, answers questions like how long do we move similarly (in terms of direction). Consistency, off-diagonal, compares remote intervals: do we move the same way (direction wise) in remote lyrical intervals?

time intervals as indicated by the black rectangles. This subdivision resembles the subdivision of the music into heroic and lyric fragments but it is not exactly identical in terms of duration of the segments. Major differences come from the absence of the first heroic interval, the second lyric interval that seems to last longer and from the third heroic interval that seems to start later. If we look to the music (Fig. 5.8) for a possible explanation we see that the build up for the third heroic part is gentle and rather an extension of the previous lyric part.

The subdivision in intervals is based on a characteristic that we define as persistency. Persistency refers here to a continuous time interval where for every subject a particular (direction) cluster dominates. This causes the high values near the diagonal in the density plot. The persistency effect is clearly highest for lyric time intervals and corresponds with what we anticipated from the individual analyses as made in Fig. 5.8, namely, that lyric parts have a highly dominant cluster.

Further we discover also something what we define as consistency. Consistency is visible as off-diagonal high density areas and appears only at time intervals corresponding with the lyric time intervals. This reveals that all lyric time intervals are not only dominated by one single cluster but that this cluster is also identical to all lyric intervals. In other words the dominating cluster of lyric time interval 1 is also the dominating cluster of lyric time intervals 2 and 3. Consequently, as far as direction is concerned, subjects move the same way in all lyric fragments. The lack of consistency for the heroic intervals is clearly visible by the lack of off-diagonal high density areas at the heroic time intervals. Subjects move (direction-wise) differently in the three heroic time intervals and this supports the viewpoint of linking surprise in music to new gestural affordances as surprise is dominantly present in the heroic intervals.

5.5 Discussion

The methods we propose in this paper identify sparsity in the positional and directional data of music evoked body movement. As there is a correlation between sparsity and surprise we expect to obtain a better understanding of the unpredictable character of music. Our findings confirm that sparse movement arises for some subjects at identical time intervals pointing to a single cause, being music. Sparse movement emerges in outlying positional data and also in movement direction data.

To come to the findings for positional data we subdivided the movement space in areas of high density by modeling this space as a mixture of Gaussians. Using the Bayesian statistical framework, a Dirichlet Process prior is used to learn the number of clusters directly from the data. Small clusters stand for sparse movement and are correlated with the sparse moments of surprise in the music. Large clusters reveal other interesting effects. Although more research is required, in our case (with music from Brahms) the large clusters appear to lay on the surface of an ellipsoid. The surface of an ellipsoid can be described by two parameters giving it a dimension of two. Leaving this ellipsoid or otherwise stated changing the dimensionality of the movement could also be an indicator of surprise in a musical excerpt.

The directional analysis used a multinomial mixture model with a Dirichlet Process as prior. This analysis clustered a set of fixed time intervals (having a duration of 3 s in our case) with respect to similar directional content. This led to the discovery of percentage wise larger clusters in the lyric style intervals. Eventually, the results from multiple subjects were consolidated in a *directogram* used to visualize the directional content of the gestures over the whole musical excerpt. This directogram is a powerful visualization tool as it allows to derive the directional information in terms of persistency (directional content stays unchanged) and consistency (same directional movement at different times in the music).

Future work could avoid the categorization process during directional analysis and work directly with the variables azimuth and elevation. Note that these variables should be handled by circular statistics and that should result in clusters of bivariate von Mises distributions. We refer for example to work of Lennox et al.[24] where this approach was followed, although in a completely different field.

From the directogram we understand that the directional content of movement follows the different musical styles: In the lyrical style fragments (low acoustic energy) it was highly persistent, meaning it was direction-wise similar for a long period and it showed also consistency between remote lyric fragments. In the heroic style fragments (high acoustic energy) the picture was different. Persistency was lower, meaning more novice gestures and hence the link with surprise. Consistency between heroic style intervals was low to completely absent, underlining the new gestural affordances found in each heroic style interval.

A major recurring element in many definitions for surprise is the *failure to anticipate*. To understand how this can be placed in practice, we would like to start this discussion by citing Itti and Baldi [8]: "In the Bayesian framework, we develop the only consistent theory of surprise, in terms of the difference between the posterior and prior distributions of beliefs of an observer over the available class of models or hypotheses about the world. We show that this definition derived from first principles presents key advantages over more ad-hoc formulations, typically relying on detecting outlier stimuli."

Itti and Baldi tested their theory on humans directing their gaze towards surprising items while watching television and video games. Abdallah and Plumbley [25] elaborated on these principles and worked out an advanced model for surprise detection in the perception of music. Fundamental to their theory are the Bayesian approach and measures like entropy and mutual information.

Although we consider the theory from Abdallah and Plumbley as very valuable, we miss the aspect of embodiment. Abdallah and Plumbley consider their disembodied approach even as a feature as it makes their model more generic: "the model operates at a level of abstraction removed from the details of the sensory experience". Another comparable experiment with also a disembodied approach is from Eerola et al. [26]. where listeners had to continuously rate how easy it is to predict (continue) a melody. There, the authors suggest as future work to include an aspect of embodiment by letting subjects to sing along and investigate what happens. Including embodiment is in line with our vision and follows the ideas of Leman [11], who states that music cognition can not be understood loose from an embodied approach. This makes the implementation of a Bayesian model considerably more difficult and that is why we left that path and why we applied a secondary descriptor of surprise, namely sparsity. Note that we prefer to use the term sparsity compared to other terms like outliers or novelties, as often used by other authors (e.g. [8]).

5.6 Conclusion

In this experiment, we analyzed a group of subjects dancing spontaneously and individually to music. The idea was to search for sparsity, sparsity being a secondary indicator of surprise in music. The followed method made use of Dirichlet Process Mixture (DPM) models and allowed to identify sparsity in positional and directional attributes of movement data. The time stamps of sparsity could be linked to suspected moments of surprise in the music.

An additional result was the development of a new type of graph, namely a directogram that can be used as a summary descriptor for a musical excerpt. A directogram explains a musical excerpt in terms of the directional content of evoked body movement and has in some perspective analogies with a correlation diagram.

The present experiment was executed with subjects moving on music of Brahms. It would be interesting to apply our method to other, even modern musical styles.

Our methods are not limited to only music evoked body movement but can in principle be extended to other fields were sparsity (in movement) has to be measured. We think for example of applications in sports analysis and rehabilitation.

References

[1] N. Birbaumer, W. Lutzenberger, H. Rau, C. Braun, and G. Mayer-Kress. *Perception of music and dimensional complexity of brain activity.* International Journal of Bifurcation and Chaos in Applied Sciences and Engineering, 6(2):267–278, 1996.

[2] G. D. Birkhoff. *Aesthetic measure.* Cambridge, Mass., 1933.

[3] L. B. Meyer. *Emotion and Meaning in Music.* The University of Chicago Press, 1956.

[4] D. E. Berlyne. *Aesthetics and psychobiology.* Appleton-Century-Crofts, 1971.

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

[6] G. Mayer-Kress, R. Bargar, and I. Choi. *Musical structures in data from chaotic attractors.* University of Illinois at Urbana-Champaign, 1992.

[7] E. H. Margulis. *Surprise and listening ahead: Analytic engagements with musical tendencies.* Music Theory Spectrum, 29(2):197–217, 2007.

[8] L. Itti and P. Baldi. *Bayesian surprise attracts human attention.* Advances in neural information processing systems, 18:547, 2006.

[9] E. Keogh, S. Lonardi, and B.-C Chiu. *Finding surprising patterns in a time series database in linear time and space.* In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 550–556. ACM, 2002.

[10] A. Clark. *Being there: Putting brain, body, and world together again.* The MIT Press, 1998.

[11] M. Leman. *Embodied music cognition and mediation technology.* The MIT Press, 2008.

[12] M. Leman. *Music, Gesture, and the Formation of Embodied Meaning.* In Musical gestures: Sound, movement, and meaning, chapter 6, pages 126–153. Routledge, 2009.

[13] G. Rizzolatti, L. Fogassi, and V. Gallese. *Neurophysiological mechanisms underlying the understanding and imitation of action.* Nature Reviews Neuroscience, 2(9):661–670, 2001.

[14] F. Waszak, P. Cardoso-Leite, and G. Hughes. *Action effect anticipation: neurophysiological basis and functional consequences.* Neuroscience & Biobehavioral Reviews, 36(2):943–959, 2012.

[15] R.I. Godøy. *Gestural Affordances of Musical Sound.* In Musical gestures: Sound, movement, and meaning, chapter 5, pages 103–104. Routledge, 2009.

[16] F. Heylighen. *Cognitive systems - a cybernetic perspective on the new science of the mind.* ECCO: Evolution, Complexity and Cognition - Vrije Universiteit Brussel, 2009.

[17] M. Leman. *Fundamentals of Embodied Music Cognition: a Basis for Studying the Power of Music.* In The Power of Music, Researching Musical Experiences:a Viewpoint from IPEM, pages 17–34. ACCO Leuven Den Haag, 2013.

[18] A. Camurri, I. Lagerlöf, and G. Volpe. *Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques.* International journal of human-computer studies, 59(1):213–225, 2003.

[19] D. Amelynck, M. Grachten, L. Van Noorden, and M. Leman. *Toward E-Motion-Based Music Retrieval a Study of Affective Gesture Recognition.* Affective Computing, IEEE Transactions on, 3(2):250–259, 2012.

[20] Y.W. Teh. *Dirichlet Process.* Submitted to Encyclopedia of Machine Learning, 2007.

[21] K. El-Arini. *Dirichlet Process : A gentle tutorial.* Select Lab Meeting, 10 2008.

[22] J.C. Sprott. *Chaos and time-series analysis*, volume 69. Oxford University Press Oxford, UK:, 2003.

[23] E. Pöppel. *The Measurement of Music and the Cerebral Clock: A New Theory.* Leonardo, 22(1):pp. 83–89, 1989.

[24] K. P. Lennox, D. B. Dahl, M. Vannucci, and J. W Tsai. *Density estimation for protein conformation angles using a bivariate von Mises distribution and Bayesian nonparametrics.* Journal of the American Statistical Association, 104(486):586–596, 2009.

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

[26] T Eerola, P Toiviainen, and CL Krumhansl. *Real-time prediction of melodies: Continuous predictability judgments and dynamic models.* In Proceedings of the 7th international conference on music perception and cognition, pages 473–476, 2002.

Elephants Don't Play Chess.
Rodney A. Brooks

6 Outlook - A glimpse of the future

The world of machine learning is in full evolution. What is new today might be outdated tomorrow. Nevertheless in this chapter we would like to discuss some of the recent developments and give our opinion on what can be important for future musical gesture research.

Research on musical gestures is hindered by high dimensionality of the data and by complex ecological settings, making it difficult to deal with large groups of participants. Think for example about experiments with motion capture data: Just putting on a mocap suit can easily take five minutes making it impossible to work with a huge number of participants.

Unfortunately large groups are required to estimate probabilities because an obvious method to estimate, is to calculate the frequency of something happening. This brings us to the realm of statistics that is usually said to exist out of two camps : the Frequentists and the Bayesians.

6.1 Are you a Frequentist or a Bayesian?

For a Frequentist, a probability is asymptotically linked with the frequency with which one expects to observe the data, given some hypothesis about the world. It answers the natural question: What is the chance that I see this data given a hypothesis, $\boldsymbol{P(D|H)}$?

Bayesians focus on another quantity $\boldsymbol{P(H|D)}$, namely the probability of the hypothesis given the data. The link between both quantities is made by Bayes' theorem (6.1). The term $\boldsymbol{P(H)}$ in the numerator is referred to as the *prior* and expresses the knowledge before the actual data was collected. This is the main strength of a Bayesian approach as it allows to integrate expert knowledge in the analysis but it is also its weakness as expert knowledge is judged to be subjective. Because of this subjectiveness the term *probability* is often replaced by *belief.*

$$\boldsymbol{P(H|D) = \frac{P(D|H) * P(H)}{P(D)}} \tag{6.1}$$

What is the chance that Belgium wins the world soccer championship in Brasil in 2014? That is a question that a Baysian can answer and a Frequentist not because it clearly concerns a belief here and not a probability. We can't let Belgium play thousands of finals to calculate the rate of success.

This explains also why a Bayesian is called an optimist and a Frequentist a pessimist [1]. A Bayesian is an optimist, not because he believes in Belgium winning the title, but because he can give solutions based (conditioned) upon even little data. For a Bayesian, statistical inference must be understood in the context of the decisions that will be made on the basis of the inferences. Bayesian decision theory is a formal theory of decision making under uncertainty. Frequentists on the contrary are pessimists and will not take any decisions before they have seen large amounts of data, because they do not condition on the data. Their decisions should hold in all circumstances.

What we see now is that, compared to the previous century, Bayesian statistics is experiencing a revival and becomes a standard element in the machine learning toolkit. It lets us make the best possible use of a sophisticated inferential tool [1]. One of the major reasons for the revival of Bayesian statistics is that with the advent of Markov Chain Monte Carlo (MCMC) tools calculations became analytically tractable. The frequentist approach however is still out there and stays also valid: For testing new medication we prefer the pessimistic frequentist approach that guarantees a new medicine will perform well in all circumstances for all people and not just conditioned on a small group of people.

An example of using a Bayesian approach related to our experiments is the prediction of the timestamp of the next beat for making a conducting avatar. Given knowledge (prior) about the beats per minute (BPM's) a Bayesian model can predict the next beat. The more data we collect, the less important the a priori knowledge becomes and the more we will rely on the data. A frequentist would handle the same problem by expressing the information held in a Bayesian prior into a constraint. This could be achieved by limiting the beat time intervals to a certain fixed range.

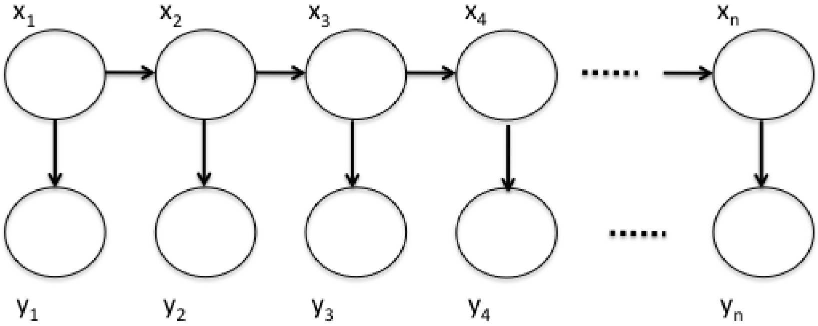


Figure 6.1: HMM visualized as a Bayesian Network. Variables (nodes) are presented as circles, conditional dependencies are displayed as arrows. The $\boldsymbol{X_i}$ *represent the hidden nodes (categorical variables) and the* $\boldsymbol{Y_i}$ *represent the observed nodes (continuous or discrete variables).*

6.2 Probabilistic Independence

Loose from a Bayesian or a Frequentist approach we see that new types of probabilistic models are emerging in music research. In our thesis we used for example GMMs, HMMs , DPMs,... Common to most of these new models is that they factorize the joint probability over a set of variables in a product of probabilities where just a few variables appear per factor. This is achieved by exploiting independencies between variables. There exist basically two ways of doing so: (1) Bayesian Networks a.k.a. Belief Network (visualized by directed acyclic graphs) and (2) Markov Networks or a.k.a. Markov Random Fields (visualized by undirected graphs).

6.2.1 Bayesian Network

A Bayesian Network has nothing to do with Bayes formula or with following a Bayesian approach but is simply a way to express a joint probability as a product of conditional properties. For example an HMM is a Bayesian Network (Fig. 6.1) and its joint probability can be expressed as in (6.2), a product of probabilities. This modularity facilitates model fitting.

$$\begin{aligned}\boldsymbol{P(X1, X2, X3, ...Y1, Y2, Y3) =}\\ \boldsymbol{P(X1) * P(Y1|X1) * P(X2|X1) * ...P(Yn|Xn)}\end{aligned} \tag{6.2}$$

Bayesian networks have an implicit causal interpretation, even if technically spoken this is not always correct. The reason is that the joint probability is factorized in a product of conditional probabilities which are often associated with causality.

Consider the following case where we have three binary variables indicating respectively the condition of the lawn (wet/dry), the status of a sprinkler system (on/off) and the weather conditions (rain/dry). Now our finding is that the lawn is wet because it either rains or because either the sprinklers are on. This points to causal conditioning. Causality is here obvious as no one will assume that the sprinklers will activate because the grass being wet.

6.2.2 Markov Network

There are circumstances where a causal interpretation is not appropriate. Take for example an application in image handling where two variables representing adjacent pixels in an image are correlated but have no causal relationship. In other words it is not the left pixel that triggers the right pixel. In this case we talk about correlations (or associations) and not about causal relationships. Dealing with correlations is the strength of a Markov Network.

For a Markov Network the joint probability of all variables is factorized in a product of joint probabilities of subsets of variables. Variables inside these subsets are probabilistic dependent and describe associations. Handling associations is the strength of a Markov Network. It represents these associations (or correlations) in terms of mutual energy between variables. Low probability corresponds with high mutual energy, high probability corresponds with low mutual energy.

A Markov network is usually represented as an undirected graph. As example we present the so called Ising model in Fig. 6.2. An Ising model is a mathematical model originating from the world of physics (ferromagnetism) where every variable is binary (representing magnetic dipole moments of atomic spins that can be in one of two states (+1 or -1)). In an Ising model every binary variable ($\boldsymbol{Y_{ij}}$) is only influenced by its directly adjacent nodes (called a cliqué $\boldsymbol{C}$). The joint probability of all binary variables $\boldsymbol{p(y)}$ can then be expressed as in (6.3). Here $\boldsymbol{Z}$ is the partition function that ensures the distribution sums to 1 and $\boldsymbol{E_C}$ is the energy function of a cliqué $\boldsymbol{C}$. High energy values correspond with low probability states.

$$\boldsymbol{p(y) = \frac{1}{Z} \exp^{-\sum_C E_C(Y_C)}} \tag{6.3}$$

The difficulty with Markov networks lays in inference because of the partition function $\boldsymbol{Z}$. The partition function $\boldsymbol{Z}$ requires a summing over a

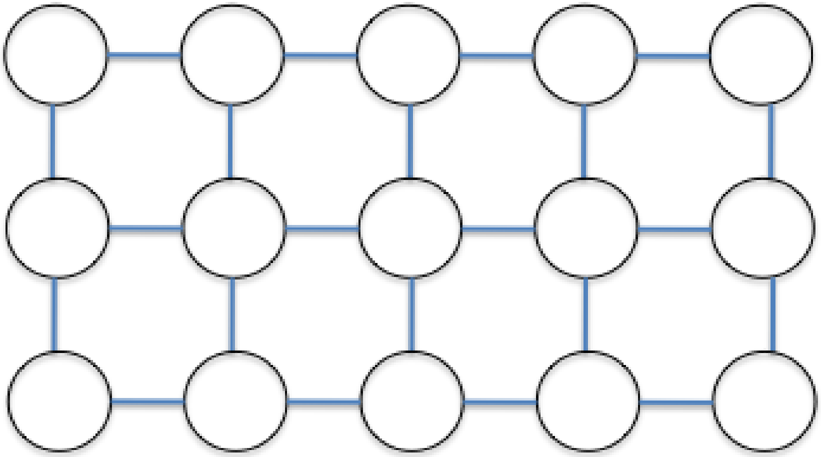


Figure 6.2: Ising model as example for a Markov Network. Every (binary) node (represented by a circle) depends only on the adjacent nodes (being the node above, the node under, the node left and the node right). The node and its adjacent nodes form a cliqué ($\boldsymbol{C}$). The cliqué makes by definition a node independent of other nodes not belonging to the cliqué. The dependency between nodes in a cliqué is described by an energy function $\boldsymbol{E_C}$. Dependency is shown by connecting lines without arrows.

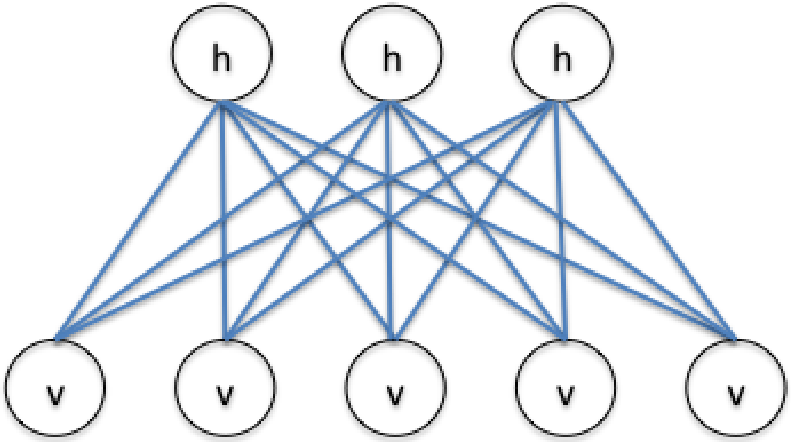


Figure 6.3: RBM : Restricted Boltzmann Machine. ***v*** *nodes are the visible nodes and refer to observed data. The visible nodes do not depend on other visible nodes.* ***h*** *nodes are the hidden nodes and refer to an an underlying bot not observed structure. The hidden nodes are also independent of other hidden nodes.*

potentially high number of variable states. Various solutions exist and usually they handle this complexity by limiting the general architecture of the Markov Networks. Particular successful is the Restricted Boltzmann Machine (RBM) (Fig. 6.3). The RBM is a generative stochastic neural network with an architecture that exists out of a bipartite graph : having no connections between hidden nodes and having no connections between visible nodes. RBM's can be learned using contrastive divergence algorithm [2].

An example of an application for RBM's in a musical context can be found in a collaborative filtering system, a technique used by some recommender systems. Here, a visible (observed) variable corresponds to users liking or disliking a particular song. But in reality beneath the visible variables lives a lower dimensional space of hidden variables. Users liking song A will usually also like song B, but dislike song C. A hidden variable could then represent 'liking song A and B, but disliking C'. Determination of the hidden variables can help to predict a user's appreciation over songs he never has heard of. If a new user likes A and dislikes C, we predict he will probably also like song B.

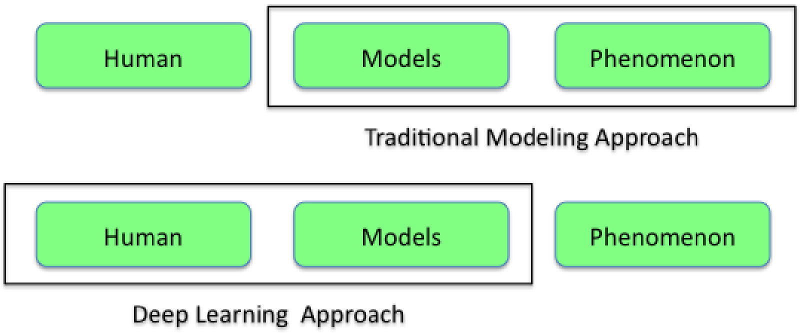


Figure 6.4: Deep Learning shifts the focus from modeling phenomena to modeling the human (brain).

6.3 Deep Learning

So far we focused on probabilistic models for various phenomena: We built models for understanding the directional content of musical gestures. We built models for understanding the emotional content of musical gestures. We built models for conducting gestures. For every phenomenon we built or could build models. Most of these models emulate how a human being makes sense of a phenomenon. Dimension reduction, feature extraction, hierarchical clustering, correlation analysis, ... are all amongst the techniques used by a human to get an understanding of a natural phenomenon. So an obvious question is whether is not feasible to model the 'reasoning' capabilities of a human in order to supersede all individual phenomenon dependent models (Fig. 6.4). That is the approach of "Deep Learning".

Deep Learning is according to *MIT Technology Review* one of the 10 breakthrough technologies for 2013 : "With massive amounts of computational power, machines can now recognize objects and translate speech in real time. Artificial intelligence is finally getting smart."

Humans interfere with the real world using associations and predicting (causal) outcome. Outcome is something inherent to Bayesian Networks (conditional probabilities) and modeling associations is the main strength of a Markov network (joint probabilities). The idea of Deep Learning is to make up a layered structure with sigmoid belief nets (a type of Bayesian Network) at the bottom and a Restricted Boltzmann Machine (a type of Markov Network) at the top (Fig. 6.5). The layers in such models correspond to distinct (hidden) levels of concepts, where higher-level concepts are defined from lower-level ones, and the same lower-level concepts can help to define many higher-level concepts [3].

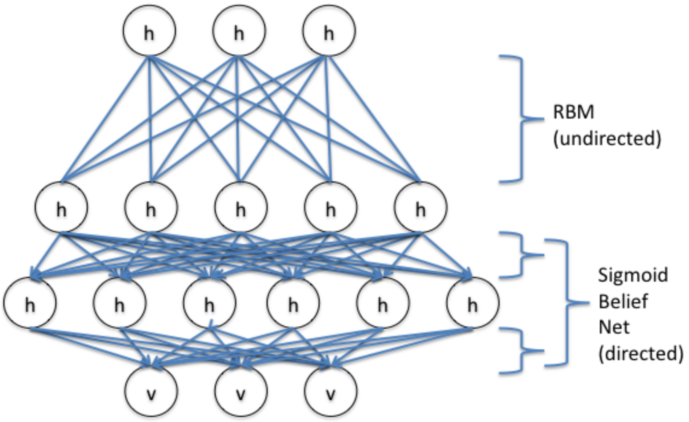


Figure 6.5: Deep Belief Network. h *stands for hidden nodes.* v *for visible nodes. (Figure based upon[4])*

One of the reasons for the impressive results of Deep Learning obtained in several areas is that the supervised learning task (e.g. digit recognition) involves an unsupervised learning component, usually in an unsupervised pre-training phase. [5].

In mathematical terms: Determining the parameters of a model requires to solve an optimization function over a high dimensional feature space. Usually this is solved by some gradient descent approach to find the optimum. This procedure starts with a search from some random defined point and continues until a optimum is found. Usually this is a local optimum. To improve the algorithm an unsupervised pre-training can define a better area to start. The descent is then so to speak to fine-tune. That is the mathematical explanation for the success of this method.

A more intuitive explanation is illustrated in Fig. 6.6 for a musical gesture labeling task. Unsupervised pre-training (Fig. 6.6.b) means that first we try to rebuild music from gestures. This is equivalent to an unsupervised feature detection step. The found features will eventually help to determine the labels. This procedure is closer to how a human brain works with collected information. If a human has to identify a dog on a picture, he does not look at every pixel but first he tries to identify features of a dog (like legs, ears, ..). This is analog to rebuilding music from gestures, namely the reconstruction of the originating source (here a dog) from some data (here a picture).

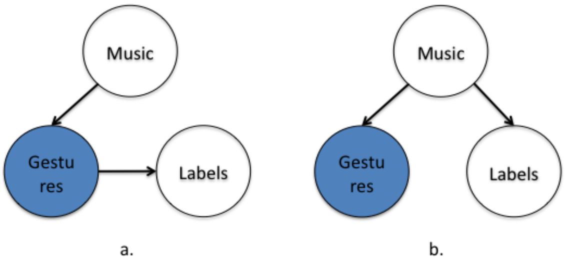


Figure 6.6:

Shaded node (Gestures) is the observed data. Unshaded nodes are not observed. The illustration is for a classification (supervised) problem.

a. $\boldsymbol{P(Label|Gestures)}$ *is independent from Music given the Gestures. Labels are learned directly from the gestures without considering their cause.*

b. $\boldsymbol{P(Label|Gestures)}$ *requires information over Music : We have first to rebuild Music from the observed data (Gestures). This step is what we call unsupervised pre-training.*

(The above figure is a customized figure. Original figure was shown at Machine Learning Summer School (MLSS), Cambridge 2009 and is from author: Geoffrey E. Hinton, Department of Computer Science, University of Toronto)

6.4 A Critical Note

Deep Learning triggered a revival of Artificial Intelligence. From the past we learned that despite some initial successes artificial intelligence so to speak stagnated. Will it be different this time? Anyway it is a good moment to refresh our memory by citing Rodney A. Brooks [6]: ”The traditional A.I. approach has emphasized the abstract manipulation of symbols, whose grounding, in physical reality has rarely been achieved.”

Therefore he suggests a research methodology which emphasizes ongoing physical interaction with the environment as the primary source of constraint on the design of intelligent systems. As such Brooks stresses the importance of embodiment in an approach to artificial intelligence.

My personal opinion is that the success of ”Deep Learning” will greatly depend on the presence of an embodiment component as ... elephants don’t play chess [6] .

Figure 6.7: Elephants don’t play chess.

References

[1] M. I. Jordan. *Are You a Bayesian or a Frequentist?* Summer School Lecture, Cambridge, 2009.

[2] G. E. Hinton. *Training products of experts by minimizing contrastive divergence.* Neural computation, 14(8):1771–1800, 2002.

[3] Y. Bengio. *Learning deep architectures for AI.* Foundations and Trends® in Machine Learning, 2(1):1–127, 2009.

[4] K. P. Murphy. *Machine learning: a probabilistic perspective, pg 996.* The MIT Press, 2012.

[5] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio. *Why does unsupervised pre-training help deep learning?* The Journal of Machine Learning Research, 11:625–660, 2010.

[6] R. A. Brooks. *Elephants don't play chess.* Robotics and autonomous systems, 6(1):3–15, 1990.

A
Appendix - Technical Background

A.1 Dirichlet Process Mixture model

A.1.1 Introduction

A Dirichlet process is just like the more familiar Gaussian process a stochastic process. Unlike a Gaussian process where every sample path is a function, for a Dirichlet process every sample path is a distribution function. A Dirichlet process can therefore be understood as a distribution over distributions. For a random distribution $\boldsymbol{G}$ to be distributed according to a Dirichlet Process, its marginal distributions have to be Dirichlet distributed [1]. Let $\boldsymbol{G_0}$ be a distribution over $\boldsymbol{\Theta}$ and $\boldsymbol{\alpha}$ be a positive number. If for any measurable partition $\boldsymbol{A_1}$, $\boldsymbol{A_2}$... $\boldsymbol{A_n}$ of $\boldsymbol{\Theta}$ the relation

$$\boldsymbol{G(A_1), G(A_2), ..., G(A_n) \sim Dir(\alpha G_0(A_1), \alpha G_0(A_2), ..\alpha G_0(A_n))} \tag{A.1}$$

holds, we say $\boldsymbol{G}$ is Dirichlet distributed with base distribution $\boldsymbol{G_0}$ and concentration parameter $\boldsymbol{\alpha}$, written G $\sim$ DP($\boldsymbol{\alpha}$,$\boldsymbol{G_0}$) [2]. The parameter $\boldsymbol{\alpha}$ is the concentration parameter and tells how concentrated the distribution is around the Base distribution $\boldsymbol{G_0}$. $\boldsymbol{G_0}$ can also be understood as the average distribution for a Dirichlet process.

This is symbolically illustrated in Fig. A.1 where we have a Gaussian unimodal distribution $\boldsymbol{G_0}$ as basis. $\boldsymbol{G}$ is then a random sample (i.e. a distribution function) from the Dirichlet Process with $\boldsymbol{G_0}$ as basis (as explained

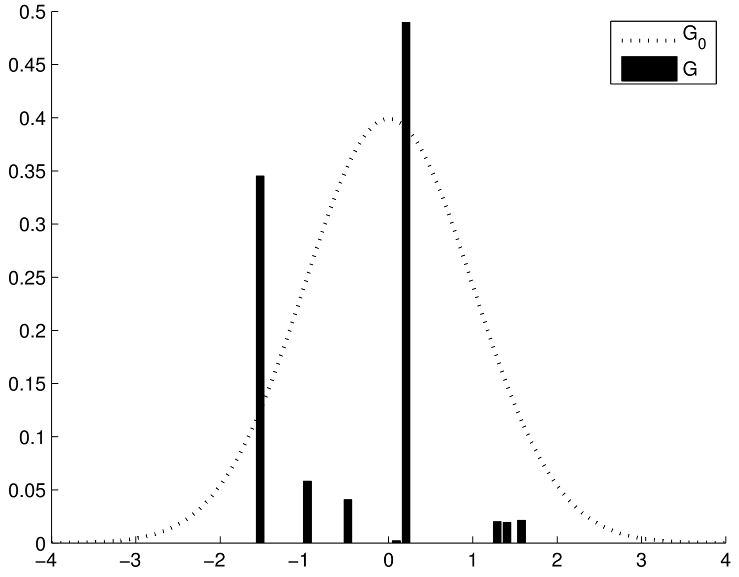


Figure A.1

by equation A.1). Note that although $\boldsymbol{G_O}$ is a pdf (probability density function), $\boldsymbol{G}$ is not. $\boldsymbol{G}$ is a pmf (probability mass function) having all its probability mass concentrated in a discrete number of parameter values. In the presented example $\boldsymbol{G}$ was generated by the so called stick-breaking method, a method that is also used as proof for the existence of a Dirichlet Process [3].

The most common application of the Dirichlet process is in clustering data using mixture models [4] [5]. Here the nonparametric nature of the Dirichlet process translates to mixture models with a countably infinite number of components. Let us model a set of n observations $\{\boldsymbol{y_1, ..., y_n}\}$ using a set of latent parameters $\{\boldsymbol{\theta_1, ..., \theta_n}\}$. Each $\boldsymbol{\theta_i}$ is drawn independently and identically from $\boldsymbol{G}$, while each $\boldsymbol{y_i}$ has distribution F($\boldsymbol{\theta_i}$) parameterized by $\boldsymbol{\theta_i}$:

$$\begin{aligned} \boldsymbol{y_i|\theta_i} &\sim \boldsymbol{F(\theta_i)} \\ \boldsymbol{\theta_i|G} &\sim \boldsymbol{G} \\ \boldsymbol{G|\alpha, G_0} &\sim \boldsymbol{DP(\alpha, G_0)} \end{aligned} \tag{A.2}$$

The predictive distribution for $\boldsymbol{\theta_{n+1}|\theta_1, ..., \theta_n}$ is :

$$\boldsymbol{\theta_{n+1}|\theta_1, ..., \theta_n \sim \frac{1}{\alpha + n}\left(\alpha G_0 + \sum_{i=1}^{n} \delta_{\theta_i}\right)} \tag{A.3}$$

The sequence of predictive distributions (A.3) for $\boldsymbol{\theta_{n+1}|\theta_1, ..., \theta_n}$ is called the Blackwell-MacQueen urn scheme [6]. It tells that a new sample for $\boldsymbol{\theta}$ comes from a new random sample of the base distribution $\boldsymbol{G_0}$ or that it is equal to an existing $\boldsymbol{\theta}$ and this proportional to the point mass that exists in that $\boldsymbol{\theta}$ (empirical distribution). The unique values of $\boldsymbol{\theta_1, ..., \theta_n}$ induce a partitioning of the set [n] = 1,..., n into clusters such that within each cluster, say cluster k, the $\boldsymbol{\theta}$'s take on the same value $\boldsymbol{\theta_k}$. The distribution over partitions is called the Chinese Restaurant Process (CRP) due to a different metaphor [2].

A.1.2 Practical Implementation

The most direct approach to sampling for model (A.3) is to repeatedly draw values for each $\boldsymbol{\theta_i}$ from its conditional distribution given both the data and the $\boldsymbol{\theta_j}$ for $\boldsymbol{j \neq i}$ (written as $\boldsymbol{\theta_{-i}}$). This conditional distribution is obtained by combining the likelihood for $\boldsymbol{\theta_i}$ that results from $\boldsymbol{y_i}$ having distribution F($\boldsymbol{\theta_i}$), which will be written as F($\boldsymbol{y_i}$, $\boldsymbol{\theta_i}$), and the prior conditional on $\boldsymbol{\theta_i}$, which is given by (A.3). When combined with the likelihood, this yields the following conditional distribution for use in Gibbs sampling [7]:

$$\boldsymbol{\theta_i|\theta_{-i}, y_i \sim \sum_{i \neq j} q_{i,j}\delta_{\theta_i} + r_i H_i.} \tag{A.4}$$

Here, $\boldsymbol{H_i}$ is the posterior distribution for $\boldsymbol{\theta}$ based on the prior $\boldsymbol{G_0}$ and the single observation $\boldsymbol{y_i}$ with likelihood F($\boldsymbol{y_i, \theta}$). The values of the $\boldsymbol{q_{i,j}}$ and of $\boldsymbol{r_i}$ are defined by

$$\begin{aligned} \boldsymbol{q_{i,j}} &= \boldsymbol{bF(y_i, \theta_j)} \\ \boldsymbol{r_i} &= \boldsymbol{b\alpha \int F(y_i, \theta) dG_0(\theta)} \end{aligned} \tag{A.5}$$

where b is such that $\boldsymbol{\sum_{j \neq i} q_{i,j} + r_i = 1}$. This algorithm (A.4)(A.5) is known as algorithm 1 for Dirichlet Process Models. For this Gibbs sampling method to be feasible, computing the integral defining $\boldsymbol{r_i}$ and sampling from $\boldsymbol{H_i}$ must be feasible operations. This will generally be so when $\boldsymbol{G_0}$ is the conjugate prior for the likelihood given by F. The conjugate prior distribution for a multivariate Gaussian in which both the mean $\boldsymbol{\mu}$ and the precision $\mathbf{\Sigma^{-1}}$ are unknown is the Gaussian-Inverse Wishart distribution. The conjugate prior for a multinomial distribution is a Dirichlet distribution. Since the algorithm (A.5) cannot change the $\boldsymbol{\theta}$ for more than one observation simultaneously, a change to the $\boldsymbol{\theta}$ values for observations in such a group can occur only rarely, as such a change requires passage through a low-probability intermediate state in which observations in the group do not all have the same $\boldsymbol{\theta}$ value [7].

This problem is avoided [7] if Gibbs sampling is instead applied to the model formulated as in (A.6), with the mixing proportions p integrated out. When K goes to infinity, we cannot, of course, explicitly represent the infinite number of $\boldsymbol{\phi_c}$. We instead represent, and do Gibbs sampling for, only those $\boldsymbol{\phi_c}$ that are currently associated with some observation. Gibbs sampling for the $\boldsymbol{c_i}$ is based on the following conditional probabilities (with $\boldsymbol{\phi_{c_i}}$ here being the set of $\boldsymbol{\phi_c}$ currently associated with at least one observation).

$$\begin{aligned} y_i|c_i,\phi &\sim F(\phi_{c_i}) \\ c_i|p &\sim Multinomial(p_1,..,p_K) \\ \phi_{c_i} &\sim G_0 \\ p &\sim Dirichlet(\alpha/K,...,\alpha/K); \end{aligned} \tag{A.6}$$

In a conjugate context, we can often integrate analytically over the $\boldsymbol{\phi_c}$, eliminating them from the algorithm. The state of the Markov chain then consists only of the $\boldsymbol{c_i}$, which we update by Gibbs sampling using the following conditional probabilities :

$$\begin{aligned} &If\ c = c_j\ for\ some\ j\ \neq i: \\ P(c_i = c|c_{-i}, y_i, \phi) &= b\frac{n_{-i,c}}{n-1+\alpha}\int F(y_i,\phi_c)dH_{-i,c}(\phi) \\ P(c_i \neq c_j\ for\ all\ j \neq i|c_{-i}, y_i, \phi) &= b\frac{\alpha}{n-1+\alpha}\int F(y_i,\phi)dG_0(\phi) \end{aligned} \tag{A.7}$$

This algorithm (known as algorithm 3) is presented by MacEachem for mixtures of normals [8] and by Neal for models of categorical data [7].

References

[1] T.S. Ferguson. *A Bayesian analysis of some nonparametric problems.* The annals of statistics, 1(2):209–230, 1973.

[2] Y.W. Teh. *Dirichlet Process.* Submitted to Encyclopedia of Machine Learning, 2007.

[3] J. Sethuraman. *A constructive definition of Dirichlet priors.* Technical report, DTIC Document, 1991.

[4] M.D. Escobar and M. West. *Bayesian Density Estimation and Inference Using Mixtures.* Journal of the american statistical association, 90(430), 1995.

[5] R.M. Neal. *Bayesian mixture modeling.* In Maximum Entropy and Bayesian Methods: Proceedings of the 11th International Workshop on Maximum Entropy and Bayesian Methods of Statistical Analysis, Seattle, pages 197–211, 1991.

[6] D. Blackwell and J.B. MacQueen. *Ferguson distributions via Pólya urn schemes.* The annals of statistics, 1(2):353–355, 1973.

[7] R.M. Neal. *Markov chain sampling methods for Dirichlet process mixture models.* Journal of computational and graphical statistics, 9(2):249–265, 2000.

[8] S.N. MacEachern and P. Müller. *Estimating mixture of Dirichlet process models.* Journal of Computational and Graphical Statistics, 7(2):223–238, 1998.