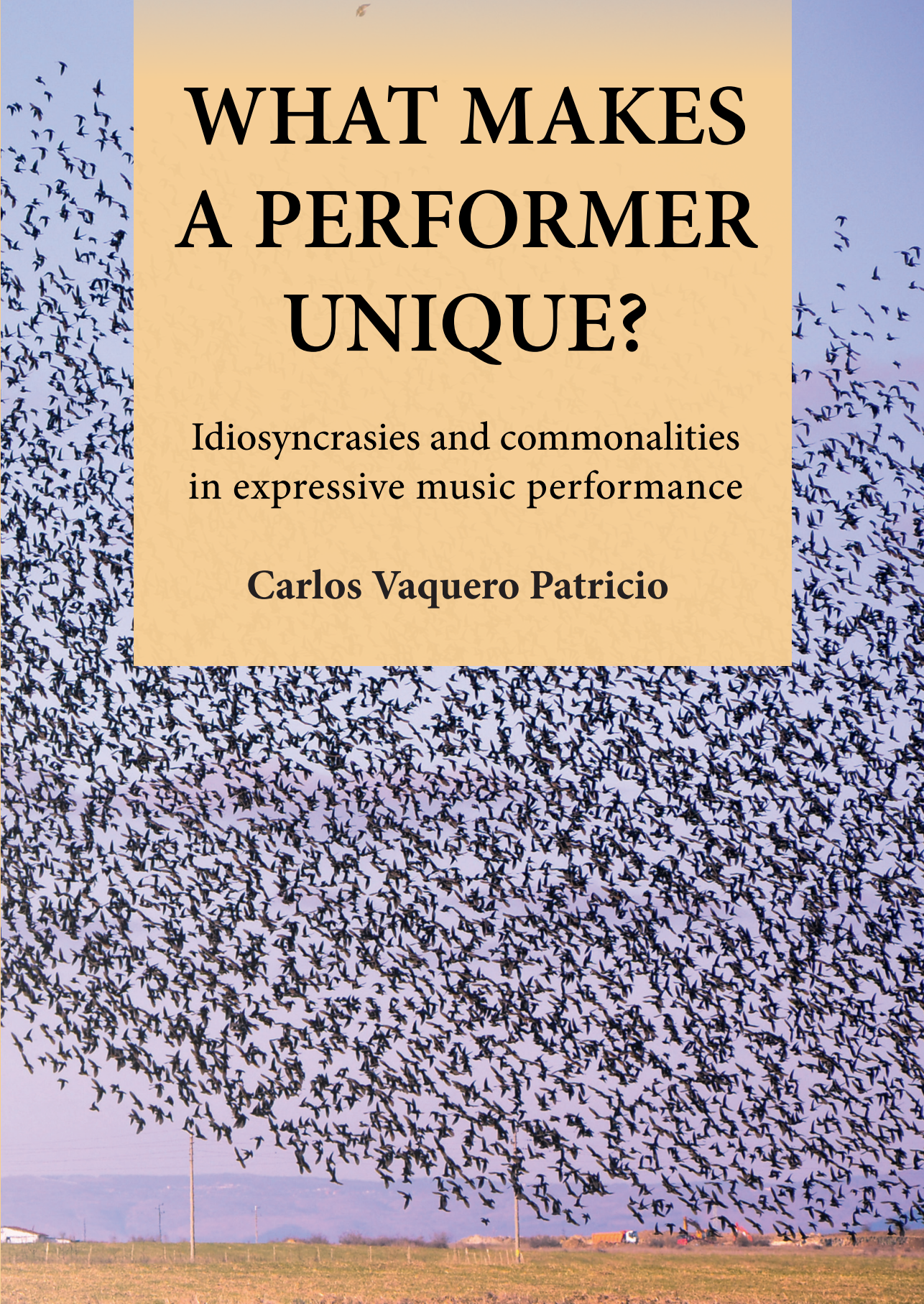


WHAT MAKES A PERFORMER UNIQUE?

Idiosyncrasies and commonalities in expressive music performance

Carlos Vaquero Patricio

WHAT MAKES A PERFORMER UNIQUE?

Idiosyncrasies and commonalities in expressive music performance

Carlos Vaquero Patricio

ILLC Dissertation Series DS-2019-01

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

For further information about ILLC-publications, please contact

Institute for Logic, Language and Computation
Universiteit van Amsterdam
Science Park 107
1098 XG Amsterdam
phone: +31-20-525 6051
e-mail: illc@uva.nl
homepage: http://www.illc.uva.nl/

This document was typeset using the LaTeX template classicthesis developed by André Miede (http://code.google.com/p/classicthesis/).

Cover image by Hristo Rusev / Shutterstock.com
Printed and bound by Off Page.
ISBN: 978-94-6182-950-4

WHAT MAKES A PERFORMER UNIQUE?

Idiosyncrasies and commonalities in expressive music performance

Academisch Proefschrift

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. K.I.J. Maex
ten overstaan van een door het College voor Promoties ingestelde
commissie, in het openbaar te verdedigen in de Agnietenkapel
op donderdag 23 mei, te 14:00 uur

door

Carlos Vaquero Patricio

geboren te Madrid, Spanje

Promotiecommisie

Promotor:	Prof. dr. H.J. Honing	Universiteit van Amsterdam
Copromotor:	Dr. I.A. Titov	Universiteit van Amsterdam
Overige leden:	Prof. dr. E. Chew	Queen Mary University of London
	Prof. dr. J.J.E Kursell	Universiteit van Amsterdam
	Prof. dr. L.W.M. Bod	Universiteit van Amsterdam
	Dr. R.S. Ahrendt	Universiteit Utrecht
	Dr. M. Sadakata	Universiteit van Amsterdam

Faculteit der Geesteswetenschappen

"Momo listened to everyone and everything - even to the rain and the wind and the pine trees - and all of them spoke to her after their own fashion."
Michael Ende, "Momo"

CONTRIBUTIONS

Chapter 1,
Carlos Vaquero (CV) wrote the chapter, with contributions by Henkjan Honing (HH).

Chapter 2,
The content of this chapter is based on Vaquero (2015), published in Early Music Journal.
CV collected the dataset, performed the musical and statistical analyses, created the figures and wrote the manuscript, with contributions by HH. An anonymous reviewer gave feedback on the manuscript. CV adapted the journal publication to the dissertation.

Chapter 3,
CV wrote the chapter and created the figures, with contributions by Ivan Titov (IT) and HH.

Chapter 4,
The content of this chapter was presented at the 2017 ESCOM (European Society for the Cognitive Sciences of Music) conference.
CV designed the study, adapted the dataset, programmed the experiments, performed the analysis, created the figures and wrote the manuscript, with contributions by IT and HH.

Chapter 5,
CV designed the study, adapted the dataset, programmed the experiments, performed the analysis, created the figures and wrote the manuscript, with contributions by IT.

Chapter 6,
CV designed the study, collected the stimuli and experimental data, performed the statistical analysis, created the figures and wrote the manuscript. Yasmin Mzayek contributed to the study design and the experimental data collection. Bruno Gingras gave feedback on an early draft of the manuscript. HH advised on the design and contributed to the manuscript editing.

Chapter 7,
CV wrote the chapter, with contributions by HH.

Appendix 2,
The content of this appendix includes an abstract presented during the 2016 ORBEL Conference.
CV collected the dataset, programmed the experiments, performed the statistical analysis and wrote the abstract, with contributions by Elaine Chew.

Appendix 3,
The content of this appendix is based on Vaquero and Honing (2014), presented during the 2013 Symposium on Logic, Music and Quantum Information and the 2015 AISB (Artificial Intelligence and Simulation of Behaviour) conference.
CV wrote and revised the manuscript, with contributions by HH. Stelios Manousakis and an anonymous reviewer gave feedback on an early draft of the AISB proceedings publication. CV extended and adapted the proceedings publication to the dissertation.

CONTENTS

ACKNOWLEDGMENTS xi

1 INTRODUCTION AND OUTLINE 1
- 1.1 Definitions of expressiveness in music performance 2
- 1.2 Expression and communication 5
- 1.3 Expression and structure 6
- 1.4 The role of surprise in performance expressiveness 6
- 1.5 Machine learning and pattern recognition of performance expressiveness 7
- 1.6 Outline of the thesis 8

2 CHALLENGES OF A COMPARATIVE ANALYSIS ON INDIVIDUAL EXPRESSIVE PERFORMANCES 11
- 2.1 Introduction 11
- 2.2 Measuring tools for the analysis of expressivity in music performance 12
  - 2.2.1 Timescapes 12
  - 2.2.2 Kendall's Tau rank correlation coefficients, means and standard deviations 14
  - 2.2.3 Local Maxima Phrase Detection (LMPD) 15
- 2.3 Dataset selection 16
- 2.4 Detection and annotation of note onsets and loudness 16
  - 2.4.1 Timing annotations 16
  - 2.4.2 Loudness extraction 17
- 2.5 Data analysis and results 18
  - 2.5.1 Timescapes analysis 18
  - 2.5.2 Means and Standard Deviations analysis 18
  - 2.5.3 Rank Correlations 20
  - 2.5.4 Local Maxima Phrase Detection (LMPD) 22
- 2.6 Conclusion and discussion 24

3 AN INTRODUCTION TO COMPUTATIONAL MODELING OF PERFORMANCE EXPRESSIVENESS 27
- 3.1 Introduction 27
- 3.2 Machine learning basics 28
  - 3.2.1 Linear regression 30
  - 3.2.2 Gradient descent 32
  - 3.2.3 Parameters and hyperparameters 35
  - 3.2.4 Model selection 35
- 3.3 Neural Networks 37
  - 3.3.1 Feed-forward neural network 38
- 3.4 Machine learning as a performance modeling strategy 41

3.4.1 Previous uses of machine learning to model performers idiosyncrasy 42
3.4.2 Neural networks in expressive performance modeling 43
4 MODELING TEMPO AND LOUDNESS EXPRESSIVENESS AT SCORE MARKINGS 45
4.1 Introduction 45
4.1.1 Hypotheses and Experiments 48
4.2 Material and Methods 49
4.2.1 Dataset 49
4.2.2 Score-based features 51
4.2.3 Performance-based features 52
4.2.4 Models 56
4.2.5 Model selection and evaluation 57
4.3 Experiments and results 58
4.3.1 Experiment 1. Score-based models: Predictions of tempo and loudness at score markings 58
4.3.2 Experiment 2. Performer-based models: Predictions of tempo and loudness at score markings 60
4.4 Discussion 62
5 SEQUENTIAL MODELING OF EXPRESSIVE PERFORMANCES 67
5.1 Introduction 67
5.2 Using sequential models to account for structure in expressiveness 69
5.3 Recurrent Neural Networks and temporal patterns 72
5.4 Long Short-Term Memory networks 74
5.5 Experiment description and hypothesis 77
5.5.1 Experiment description 77
5.5.2 Hypotheses 78
5.6 Dataset and feature preparation 79
5.6.1 Performance features 79
5.6.2 Score features 79
5.7 Model architecture 81
5.8 Experiments 82
5.8.1 Persistence algorithm baseline 83
5.8.2 Experiment 1. Performer-based models 83
5.8.3 Experiment 2. Score-based models 85
5.8.4 Results analysis 88
5.9 Visualization of performers expressive idiosyncratic deviations 90
5.10 Conclusions and discussion 93
6 A PERCEPTUAL STUDY ON THE ROLE OF EXPRESSIVE LOUDNESS AND TIMING 105
6.1 Introduction 105
6.1.1 A review on expressiveness recognition experiments 105
6.1.2 Challenges in diagnosing expressiveness perception 108

6.1.3 Experimental design and hypothesis 110
6.2 Material and methods 110
6.2.1 Excerpts selection 110
6.2.2 Stimuli recording procedure 111
6.2.3 Stimuli collected 112
6.2.4 Experiment design and stimuli manipulation 113
6.2.5 Procedure 114
6.2.6 Participants 115
6.3 Analysis and results 116
6.4 Discussion 119
6.5 Excerpts used as stimuli for experiments 1 and 2 122
7 RECAPITULATION 125
A APPENDIX 1 129
B APPENDIX 2 131
C APPENDIX 3: USING RHYTHMIC CATEGORIES TO GENERATE EXPRESSIVENESS 133
C.1 Introduction 133
C.2 Definition and visualization of rhythmic categories 134
C.3 Lindenmayer systems 136
C.4 Using L-systems to generate expressiveness 139
C.4.1 Implementation 139
C.4.2 Evaluation 142
C.4.3 Practical and conceptual challenges in the implementation of the model proposed 142
C.5 Conclusions 144
REFERENCES 145
SAMENVATTING 157
SUMMARY 159
SUMARIO 161
TITLES IN THE ILLC DISSERTATION SERIES 163

ACKNOWLEDGMENTS

I am deeply grateful to my supervisors, Henkjan Honing and Ivan Titov, for their time, valuable guidance, and patience. For teaching me how to carry scientific research and to enjoy facing the challenges I encountered when delving into the fields of machine learning and expressive performance modeling. And mostly, for accepting me with my virtues and weaknesses and helping me confronting them in a constructive manner.

I am very thankful to the members of my thesis committee, Rebekah Ahrendt, Rens Bod, Elaine Chew, Julia Kursell and Makiko Sadakata, for taking the time to read and evaluate this dissertation and to be present at my defense ceremony.

I thank the University of Amsterdam, for the financial support offered during four years of my Ph.D. through the Research Priority Area Brain & Cognition, and the Institute for Logic, Language and Computation (ILLC), for letting me be part of one of the most intellectual environments I will ever encounter.

Among the many great people I met at the ILLC, I would like to thank my colleagues at the Music Cognition Group for picking my brain on so many different interesting topics in music science. Among these, special thanks to Berit Janssen, Joey Weidema, Bastiaan van der Weij, Ashley Burgoyne, Fleur Bouwer, Myrthe Knetemann, Yasmin Mzayek, and Paula Roncaglia-Denissen, for the feedback and discussions during the group meetings.

From the ILLC, I also wish to thank Aline Honingh, for inviting me to be her teaching assistant in the Computational Musicology course, and Jenny Batson, Karine Gigengack, Tanja Kassenaar and Hotze Mulder (from FGW), for easing the path on many administrative, housing and human aspects. My gratitude goes as well to Inés Crespo, Jordy Jouby, Shengyang Zhong, Raquel Fernandez, Paolo Galeazzi, Dieuwke Hupkes, and Raquel G. Alhama, for the insights or coffees around Science Park.

Another institution to which I am very grateful is the Centre for Digital Music, at Queen Mary University of London, that I had the pleasure to join during the spring/summer of 2015 as a research visitor. I would like to thank everybody I met there for making of this visit a very inspiring time. In particular, Elaine Chew, for welcoming and guiding me during those months, and Katerina Kosta, for her generosity as a researcher, and for encouraging me to keep working on this topic.

In Madrid, Amsterdam, Den Haag, and London I (re-)encountered many good friends. Among these, I am indebted to Jamie McLaren, Facundo Carreiro, Stelios Manousakis, Reinier de Valk, Ana Arribas, Thomas Brochhagen, Iris Pleijsier, Wouter Pleijsier, and Po Heng Chan, for proofreading parts of my thesis in moments in which I needed it much.

For the dinners, music making, and inspiration, I would like to thank Alberto Bernal, Kristian Holsheimer, Flora Lysen, Bogdan Vera, Rui Silva, Qiong He, Pedro Barbadillo,

Juan Carlos Blancas, Bruno Rocha, Maria Panteli, Jan van Balen, Srikanth Cherla, and the Gruys family.

Lastly, I want to thank the most important people to me, my family. Especially, my sisters, Cristina and Elena, and my nieces and nephew, for sending me through skype some light (and the most beautiful smiles) all the way up here. My partner, Sonja Gruys, for her love and compassion. For reminding me that music exists beyond its analysis and modeling, and for keeping it close to me. And my parents, María Pilar Patricio Gude and Carlos Vaquero Tapia, for all their love, dedication, and support at every moment, without which this dissertation would not have been possible. Queridos padres, el mérito de esta tesis es tan mío como vuestro. Os la dedico con todo mi amor.

Amsterdam
December, 2018

Carlos Vaquero Patricio

1

INTRODUCTION AND OUTLINE

The identification and categorization of acoustic cues is a complex cognitive ability and an essential survival mechanism in several animal species (McComb, Shannon, Sayialel & Moss, 2014; Podos, 2010). Specifically, in humans, it has been argued that distinct complex processes in common cortical mechanisms may be involved in the categorization of sounds in music and language (Patel, 2012). These mechanisms allow for diverse tasks such as differentiating between performers when playing the same music piece or distinguishing a particular performer when playing two different pieces of music.

The study of expressiveness in music performance concerns diagnosing how performers interpret and play music, as well as how listeners perceive their performances. Ultimately, it aims to explain what aspects of the performance are communicated, what are the physical and perceptual constraints that affect them, and how these may have an impact on the overall development of music performance aesthetics. Assuming all people, regardless of their musical training, share the grounds of musicality to be able to perceive and appreciate music (Honing, 2018), we may wonder what aspects of music expressiveness contribute to the recognition of performers as individuals. And in this respect, what are the constraints that affect individuality in performers. With such goal in mind, this thesis investigates the production and perception of idiosyncratic expressiveness in music performance. In particular, it studies how performers expressiveness is constrained by their own idiosyncratic style or by the score. Thus, having a systematic musicological approach to the study of expressiveness, the dissertation uses computational modeling to analyze performers individuality.

At least with regards to Western classical music, the search for systematic individual and shared approaches across performances goes back to the beginning of the XIXth century, when Mathis Lussy (1828 - 1910) manually annotated scores (based on his perception) of how performers used timing, dynamics, and phrasing (Dogantan-Dack, 2014). Lussy's work aimed to find systematic patterns of behavior underlying the sound produced by performers in relation to the score with the goal of better understanding whether the score served as a constraint to expressiveness. This probably represents the first systematic work in the history of expressive performance analysis, and, moreover, a very relevant methodological step, since it constrains the expressive communication between performer and audience to an aural dimension. That is, no other elements were considered in the expressiveness communication, probably in order to ease the study and prevent complicated correlations between acoustic and no-acoustic aspects of performance.

Seashore (1866 - 1949) and his team furthered this more objective approach by registering and applying statistical methods to analyze, for the first time, performers expressiveness (Seashore & Metfessel, 1925). This enabled assessment of the relation

between performers and audience entirely in terms of sound. According to Seashore, *"everything that the singer or player conveys to the listener is conveyed through sound waves or in terms of these. This conception simplifies our approach immensely in that it frees us from confusion with unnecessary accessories, furnishes us with a basis for classification and terminology, and paves a way for preservation of findings, measurement, and scientific explanation."* (Seashore & Metfessel, 1925).

While Seashore acknowledged that many other variables could play a role in music performance and its perception (e.g., the effect of vision in expressive performance judgments (Tsay, 2013)), his work set the basis to investigate the role of sound in relation to expressiveness between "performer, music and listener" in a systematic way. This was done by registering the different expressive variables using currently available technologies and statistics. Furthermore, they carried out various behavioral experiments to better understand how listeners respond to musical expressiveness

Following on the reductionist approach initiated by Lussy and Seashore, in the work presented in this dissertation, I will consider sound to be the main channel of communication and present some of the studies carried along my doctoral research. But first, in the rest of this chapter, I will introduce several notions necessary to understanding the rationale behind this dissertation as well as its outline.

1.1 DEFINITIONS OF EXPRESSIVENESS IN MUSIC PERFORMANCE

Over the last hundred years, several definitions of expressiveness to analyze music performance have been proposed. As proposed by Dogantan-Dack (2014), our philosophical standpoint on expressiveness may be reflected in the assumptions underlying the research carried and consequently, how it contributes to our understanding regarding how expressiveness is used by performers and perceived by listeners. Within the literature on performance analysis modeling we can find several definitions of expression (Timmers & Honing, 2002):

- Expression as a deviation from a musical score,

  The first definition of expressiveness in music performance is probably the one by Seashore and Metfessel (1925). In their paper, "Deviation from the regular as an art principle", expressiveness is defined as:

  > "The unlimited resources for vocal and instrumental art lie in artistic deviation from the pure, the true, the exact, the perfect, the rigid, the even and the precise. This deviation from the exact is, on the whole, the medium for the creation of the beautiful for the conveying of emotion. That is the secret of the plasticity of art." (Seashore & Metfessel, 1925)

  This definition was further developed in Seashore (1938) and it is often linked to expressiveness as the deviation performers exercise on the mechanic rendition of the score (Timmers & Honing, 2002).

In the case of timing, representing the deviations from what is notated in the score implies "re-scaling" the score durations to the mean tempo of a performance relative to a common duration unit (e.g., a quarter note). That is, calculating the mean duration of the reference value (in this example, a quarter note) and the relative durations of all other figures notated in the score. Once the score durations are calculated, we can observe how much the performer deviates from this mean.

Despite being a common approach found in literature, there are some constraints from a cognitive perspective to be noticed. For example, proportionately equivalent deviations from different time units may be perceived differently: having a deviation from a whole note is perceived differently than from a sixteenth note even when they would deviate by the same ratio (e.g., half of the value of the notated figure).

Elaborations on the definition of expression as a deviation from a musical score can be found in the work by Gabrielsson (1974), in which performances of rhythm patterns are analyzed based on how much they deviated from the norm as given by the musical notation. Also, in the model proposed by Friberg, Bresin and Sundberg (2009), in which the value of different expressive deviations is added to the score notation value. A detailed discussion on the limitations of this definition can be found in Timmers and Honing (2002).

As a consequence of the development of systematic musicology[1], the research agenda in music performance expressiveness has been increasingly interested in the perceived and cognized representation of the music. In this regard, several alternative definitions of expressiveness as a deviation relative to a notated musical score have been proposed. The most common alternative definitions in the current literature are:

- Expression as microstructure,

  Instead of defining expressiveness having the score as a reference, Repp (1992) and Palmer (1996) define it as those variations in any of the acoustic musical features which exist without the explicit necessity of the score (Timmers & Honing, 2002).

- Expression as a deviation from the norm defined within a performance,

  Proposed by Desain and Honing (1992), this is an intrinsic definition of expressiveness which does not refer to the written musical score, but to the cognized structure by the listener. In this definition, the expressive deviations occur over the norm defined within a performance. For example, a certain inégal articulation pattern is suddenly changed through a performance. In this definition, a hierarchical structural description of the music is needed (Timmers & Honing, 2002). For

1 In this dissertation, the term systematic musicology refers to the scientific musicological research which is *"primarily empirical and data-oriented and involves empirical psychology and sociology, acoustics, physiology, neurosciences, cognitive sciences and computing and technology"* (Parncutt, 2007)

instance, the expressive beat durations are expressed as ratios of the bar duration (Timmers & Honing, 2002). Thus, according to this definition, the norm is set by a higher order unit and the expressive deviations relate to such unit and occur within.

- Expression as a deviation from the norm defined within the performance practice,

  Based on the definition by Desain and Honing (1992), Clarke (1995) suggested that the norm is defined by common music practice and how the most frequently heard renditions of performances set the basis over which a new performance may deviate or not.

Each of these definitions, relates to different constraints and choices when building or developing a performance model. For instance, while the definition of Gabrielsson (1974) takes into account the score notation as a norm, the ones by Desain and Honing (1992) or Clarke (1995) attend, respectively, to the cognized music representation of the listener and performer as the norm. As such, each definition assumes a different predisposition of the performer to expressiveness itself and, even more, from the listeners to the recognition of those deviations.

Having the score notation as the norm implies that the deviations of the auditory representation of the music are based on the score rather than on the cognized auditory representation of the music. In this regard, the exposure that listeners and performers have to previous renditions of a piece or music style is fundamental in order to conform the expressiveness norm. On top of the exposure, musical expertise (which is based on musical education and training) from both the listener as the performer has also been shown relevant when perceiving expressiveness and music structure (Sloboda, 2000), and when identifying performers (Koren & Gingras, 2014).

Having access to the different auditory features (loudness, timing, tempo, phrasing, timbre, spectrum-based features, etc) that may conform our mental representation of a performance, is a first approach to recognise patterns in and across performances. This is because, in order to extract the norm as conformed by several performances or performers, we need to extract patterns over such renditions. Having both a set representing performance-based features and another set representing score based features may help to better define which patterns of performance norms may be captured by the listener.

In this dissertation, I use computational modeling methods to investigate performance expressiveness as those deviations from the norm defined by performers. As such, I depart from the definitions of expressiveness by Desain and Honing (1992) and Clarke (1995) to investigate individual and shared constraints that may conform the expressive norm. In particular, I study how individual expressiveness might be constrained by the structural score based approaches shared by a group of performers or by their idiosyncratic style (as individuals). For such purpose, I use several machine learning methods which aim to capture the expressive norm represented as performance patterns. Thus, aiming to explain how those norms are relevant to understanding how listeners may

relate to the deviations of new performances of the same or different pieces or how performers may be characterized by their idiosyncratic expressiveness.

1.2 EXPRESSION AND COMMUNICATION

While the idea of expressiveness as a communication process between composer, performer and listener can be linked to the use of rhetoric in earlier periods in history (such as the baroque), the first complete model formalizing such a relation is, to the best of my knowledge, found in the work by Kendall and Carterette (1990). In their model, Kendall and Carterette (1990) formalize the chain of musical communication as departing from the composer to the performer and, finally, to the listener. Via these three agents, the musical message is subsequently encoded and decoded by each of them according to their shared and unshared implicit and explicit knowledge of the message as well as their contextual environment. This model, therefore, assumes some sort of fixed representation or score notation of the musical material created by the composer which is therefore interpreted by the performer. A relevant aspect of the model by Kendall and Carterette (1990) treatment of implicit and explicit operations as an information processing model allows for accounting for schemas in Long term memory, Working memory and Conscious awareness. That is, their model presumes a rather high level operationalization of the processes involved in the communication of music.

In Kendall's model, the encoding of expressiveness occurs in two steps. In a first step, the encoding depends on the performers understanding of the music to be performed. This understanding will be based on their exposure to other renditions of the same music as on the structural constraints of the piece to be performed. In a second step, the encoding depends on the listeners' expectations on the music and expressiveness.

How consistent a performer will be on their use of expressiveness (or in defining an individual expressive norm or signature) will partly determine the listener's ability to recognize their individual performance style.

The description and differentiation between performances and performers is possible thanks to several representational and control processes. These processes have to be shared between listeners and performers (Sloboda, 2000), in order for music to be communicated between them. Elucidating which mechanisms are used by performers to communicate expressiveness it is, therefore, a multi-dimensional challenge which combines:

- the structural constraints and control processes defined by the piece
- the mental representation a performer has of the piece to be performed
- the shared expressive approaches derived from the cultural context a performer belongs to

All these factors may constrain the performers' expressive style and individual "performance" signature and therefore how the message will be communicated to the listener.

1.3 EXPRESSION AND STRUCTURE

Within the organization of a given piece of music, the literature often distinguishes between two main levels of structure which together constitute such piece (Jackendoff & Lerdahl, 2006): micro-level structure and macro-level structure. In the context of performance, the macro-level structure refers to the piece form and includes those expressive deviations in tempo, rhythm, large scale dynamics, melodic contour, and harmonic relationships. The other main level of structure is the micro-level structure, which includes instead note level (or short groups of notes) deviations in timing, pitch, loudness, timbre or articulation. According to Sloboda (2000), the micro-structure expressive deviations relate to the prosody and error on a note (or few notes grouped) level, while the macro-structure refers to the use of phrasing.

A performer's mental representation of the macro-structural level might be represented by their use of phrasing. Yet, the most characteristic idiosyncratic approaches might be reflected on the expressive deviations exercised across (often) smaller units of expression (e.g., in the "swing" of a jazz performer).

As discussed by Timmers and Honing (2002), the duration or amplitude of the expressive deviations might depend on the hierarchy set by a higher order unit. Therefore, whether the expressive deviations are large or small is related to the different structural levels. For example, small variations in timing might respond to local shortening, while note lengthening variations respond to larger scale trends. Thus, the expressive features used by performers respond to both long and short time scales and structures and ideally should be captured by a model. How a performers expertise might be conditioned by the use of micro and macro-structure is addressed in Clarke (2002), which suggests that the performance of piece structure is likely to be more controlled and reproducible by expert performers.

1.4 THE ROLE OF SURPRISE IN PERFORMANCE EXPRESSIVENESS

Following Meyer's work on information theory and communication (Meyer, 1957), Huron (2006) compiled one of the most comprehensive resumes on the role of expectations in the musical phenomenon.

According to this theory, we may interpret that the communication between performers and listeners depends on the expectations of the listener as well as on the presuppositions the performer has on the listeners' expectations. Thus, how performers create tension and release it through their expressive choices may determine in which manner listeners (conditioned by their previous exposure to similar music) will be surprised to the performers' expressiveness.

Huron (2006) differentiates between four types of surprise in music, which I will explain in the context of expressive performances:

- Schematic surprise, which relates to the norm defined by the listener based on their exposure to different expressive performances. e.g., exposure to a particular school of performance in a determined style
- Dynamic surprise, which relates to the norm defined by the performer (and perceived by the listener) during a particular performance. This kind of surprise is very much linked to the definition of expressiveness as deviation from the norm defined by the unit within a performance from Desain and Honing (1992)
- Veridical Surprise, which relates to the violation of the listener's knowledge of the musical work being listened to. This violation may occur for different reasons such as , the performer effectively making a mistake and playing something for instance *fortissimo*, when in the score is written *piano*, or, in a more perceptual and complex scenario, when the expressive categories of the performer do not correspond to those of the listener (as a consequence of their different previous exposure) and they are violated.
- Conscious surprise, by which the listener that knows the style or piece being played expects an event knowing that it is not going to happen

How listeners may be "surprised" during the music listening is, according to Huron (2006), explained by two factors: predictability and contrastive valence. Predictability relates to the fulfillment on the expectations of the listeners. Contrastive valence is related to how the limbic system is able to turn negative responses into positive or neutral reactions. Thus, contrastive valence can be defined as the emotional valence between the different expectation responses (Huron, 2006). In the context of musical phrasing, a common example encountered in this sense is when an expected ending of a musical phrase is altered by delaying it with *ritardando* (increasingly lengthening the duration of notes). This delay provokes a negatively valenced tension response that enhances the positive effects of the limbic behavior in relation to the prediction response, once the phrase performed finally closes.

1.5 MACHINE LEARNING AND PATTERN RECOGNITION OF PERFORMANCE EXPRESSIVENESS

As a consequence of the advances in the fields of computational modeling, machine learning and signal processing, the field of music performance modeling essentially developed over the last three decades. This development has been characterized by applying the new computational methods to relate performance to music cognition. These methods have allowed for new models of empirical evidence about the relation between production and perception of expressiveness.

In this dissertation, I use different machine learning methods to study and characterize individual (performer based) constraints in the production of expressiveness. Using these methods, I investigate the expressive norm as defined by an individual performers style, by other performers style, or a combination of both. The patterns learned can be associated to some of the characteristics that may be intentionally performed.

In the literature, we can find machine learning methods such as the one proposed by Stamatatos and Widmer (2005), which outperform humans' perception in the recognition of performers. The methods and studies proposed in this dissertation, however, do not aim to match or beat listeners recognition capabilities. Instead, the intention of the work herewith presented is using these machine learning methods to characterize, within the auditory (and musical) features studied, performers expressiveness. Aiming, like this to define individual expressiveness profiles in the use of certain features which may be perceived by listeners.

The research presented in this dissertation focuses on keyboard music and in particular, on the use of tempo, timing and loudness expressiveness. Both the music instrument as the expressive features chosen have been extensively studied in the literature ever since Seashore (Gingras, 2014). Yet, to my knowledge, some of the studies herewith presented are the first ones entailing a balanced dataset of this size (26 pieces played by 11 performers) and focusing on the possible individual patterns and interactions in the use performers make of tempo (timing as well in Chapters 2 and 6) and loudness. With the goal of finding how individual performers are constrained by shared approaches or by their idiosyncratic style in the use of these features different studies and machine learning models have been explored and presented in order to study different hypothesis.

1.6 OUTLINE OF THE THESIS

The outline of this dissertation is as follows:

- Chapter 2 introduces methods to measure and visualize performances of the same piece played by the same or a different performer. A small dataset consisting of performances played by the same performers but recorded on different dates, is collected and presented for illustrative purposes.
- Chapter 3 presents an introduction to machine learning with the aim of providing a basic understanding of the methods used in Chapters 4 and 5.
- Chapter 4 presents a study on the possible interactions between tempo and loudness in relation to performance constraints defined by score markings. A dataset is presented including 11 performers and 26 piano pieces. Two main models are defined based on individual stylistic approaches or on shared ones (when several performers are playing the same piece).
- Chapter 5 discusses the relevance of using sequential models, in particular, Long Short-Term Memory neural networks, when predicting micro and macro struc-

tural expressiveness. It illustrates and discusses how the predictions and interactions of tempo and loudness are affected differently when the models proposed include information of rhythm as defined by the melody or metrical structure. Moreover, it investigates how the individual predictions per performer differ when the models are trained on shared score constraints among a group of performers or on individual models.

- Chapter 6 presents a behavioral experiment of listeners discrimination between performers based on their use of expressive tempo and loudness. Furthermore, it discusses how the discrimination task might be influenced by musical expertise.
- Chapter 7 recapitulates the contributions and main findings of the dissertation.

2

CHALLENGES OF A COMPARATIVE ANALYSIS ON INDIVIDUAL EXPRESSIVE PERFORMANCES

2.1 INTRODUCTION

Among all aspects of performance practice, that of interpretative choice regarding musical expressivity is arguably the most interesting. In the field of early music, the urge to realize historically informed interpretations has led to new perspectives about our musical legacy from scholars and performers alike (Butt, 2002). During the hundred years since Arnold Dolmetsch published *The interpretation of the music of the XVIIth and XVIIIth centuries* (Dolmetsch, 1915), different schools of early music performance practice have developed.

These developments have been reflected not only in the form of publications (e.g. Donington (1963) or Geoffroy-Dechaume (1964)) but mostly in the form of revised performances of practitioners re-interpreting their aesthetic approaches on the basis of surviving historical treatises such as those of Quantz (1752), Bach (1752) or Mozart (1756). Due to the popularization of recording techniques in the last century, the volume of material available for a possible study of different approaches to performance has increased considerably. In *The End of Early Music*, Haynes (2010) uses recordings in order to compare and classify different performance aesthetics. In the same publication, Haynes also quotes Anner Bylsma commenting on the "enormous" different approaches to be heard when listening to different recordings of the same pieces played by Frans Brüggen during different moments along his career.

Thanks to the on-going advancements in music technology and the popularization of digital tools nowadays these differences can not only be "heard" but also objectively quantified. The analysis of the development of performance practice and aesthetics from the perspective of cognitive and computational musicology is relevant since it can potentially provide insights into the interrelations between musicology, performance practice and cognition.

In the field of performance analysis, we may therefore aim to find out which aspects are involved in the categorization of performances in terms of their constituent expressive features or in the identification and understanding of possible performance trends that might change over time. For instance, can a performer be easily characterized by her or his typical use of expressive timing? Are the sudden changes in tempi within the performance of a piece representative of a certain musicologically informed interpretation? Is this approach communicated by the performance itself? Do perception and cognition play a role in the aesthetic choices involved in performance? Much research has been done in recent decades in order to resolve some of these questions (including the work by Clarke (1999), Gabrielsson (1974), Repp (1997), and others).

In the 1930's, music psychologist Carl Seashore had already published results revealing systematic deviations in timing on a number of sequences of different performances (Seashore, 1936). One of the most recent contributions to the field of performance analysis and modeling has been the Mazurka project. Starting in 2004 at the Centre for the History and Analysis of Recorded Music (CHARM), it collected more than 2500 recorded piano performances of 49 Chopin mazurkas. The Mazurkas dataset continues to grow and inspire new analyses, and has been used within many different domains of computational musicology. Extensive overviews on the work done in performance rendering and analysis have been completed by Gabrielsson (2003) and the recent compilations published by Fabian, Timmers and Schubert (2014) and by Miranda, Kirke and Zhang (2010). In addition to this rich literature in comparative performance, much research has been carried out towards automatic classification of performances or performers within the field of music information retrieval. A few illustrative examples in which performers are automatically identified based on their use of timing are those by Grachten and Widmer (2009), and Serrà, Özaslan and Arcos (2013). Other studies such as the ones by Stamatatos and Widmer (2005), or Ramirez, Maestre and Serra (2012) use a broader range of expressive features (such as loudness and timbre) in addition to timing for the automatic identification of performers.

In this study, I present and apply three state-of-the-art quantitative methodologies in expressive performance analysis to elucidate possible relations among musicological and cognitive interpretations. Doing so will allow me to show how a particular methodology may serve (or constrain) the ability to compare different interpretations and define expressiveness. In particular, I will work with two of the definitions presented in section 1.1. These are: "expression as a deviation from a musical score" and "expression represented as a deviation from the norm defined within a performance". In addition, I will show how these methods might be used for other purposes in performance science as well as in pedagogy. For this, seven different performances of an excerpt played by three performers of the Prelude from J. S. Bach's first Suite for solo cello in G major, bwv1007 (Appendix a), are analysed and compared. Furthermore, I will present an interpretation of the quantitative analysis of these performances.

2.2 MEASURING TOOLS FOR THE ANALYSIS OF EXPRESSIVITY IN MUSIC PERFORMANCE

Over the last 20 years, several quantitative methodologies to develop knowledge representation tools have been proposed. To compare the cello performances I have chosen three distinct methods proposed by Sapp (2007), Gingras, Lagrandeur-Ponce, Giordano and McAdams (2011) and Cheng and Chew (2008b):

2.2.1 *Timescapes*

Within the Mazurka project, Craig Sapp did extensive work defining quantitative methodologies and visualization tools. The scape-plot visualizations suggested by Sapp

(2007) use normalized correlations (Pearson correlations) per cell (e.g. isolated notes), or groups of sequential cells (groups of notes), between a reference performance and another performance in order to visualize the correlations in multiple timescales. Each cell in the scape plot represents the correlation between both performances. For example, the bottom row of Figure 2.1 represents the correlations per pairs of elements of each performance compared to the reference performance. Within the scape plot, the window size for correlations among groups of notes increases the closer we get to the upper vertex of the triangle, having at the top of the triangle the correlation of the whole reference piece with the one (or ones) being compared to. When the visualization is plotted in black and white, the correlation scale will represent black as the lowest correlation and white as the highest. Among the plot shapes originally proposed by Sapp, I chose a bell-shaped plot with a logarithmic scale in the vertical axis, to enhance visualization in the lower part of the plot. This is especially relevant when differences in timing occur within groups of only a few notes.

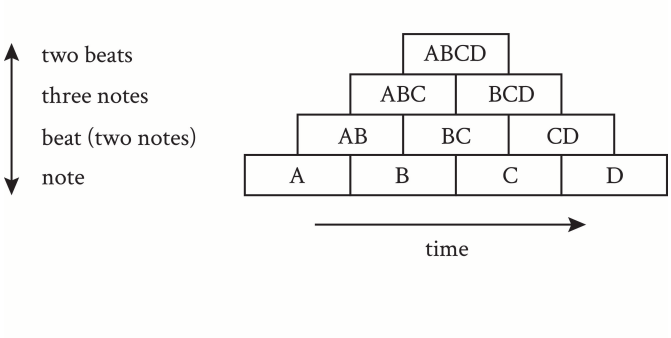


Figure (2.1) Scape plotting based on a hypothetical fragment of four notes (A,B,C,D) and four window lengths. From bottom to top each cell captures the correlation on different levels of the note sequence of the level below. Figure 2.2 shows a bell shape visualization based on such a scheme.

The possible uses of timescapes are diverse. From a musicologist's perspective, they provide a fast overview of similarities among performances but they can also be used for forensic applications. The most illustrative example is the fraud discovered when a few visualizations of recordings credited originally to the pianist Joyce Hatto showed to be identical to the visualizations of a number of recordings previously published by other performers. The reason why this fraud was discovered through these visualizations is because they were not sensitive to the digital manipulations (time-stretches and re-equalizations) that were applied to the original recordings and therefore the similarity shown was immediately evident. The visualization proposed by Sapp (2007) has been particularly useful for tracing similarities in the numerous recordings annotated within the Mazurka dataset. From a cognitive perspective the timescapes may be interpreted as a tool to illustrate the correlations between possible internalized repres-

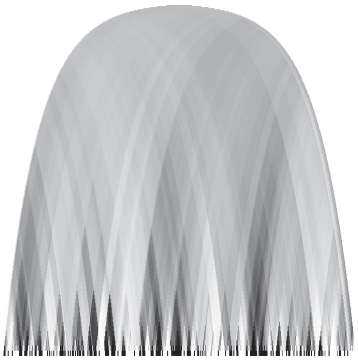


Figure (2.2) Timescape visualization of the correlation between two different performances. The global average correlations in timing between both performances are indicated at the top of the plot, the note level correlations are shown at the bottom of the plot. The color scale from black to white represents the correlation with black (maximal) and white (no correlation).

entations a listener may have against different versions of the same piece. Additionally, timescapes facilitate the identification of consistent deviations at different structural levels revealing aspects not only on the possible characterization of the performance but also on the similarities of structural phrasing. In this chapter, timescapes will be used to represent and compare correlations between different performers who recorded the work more than once.

2.2.2 *Kendall's Tau rank correlation coefficients, means and standard deviations*

A complementary correlation measure to the one used in the visualizations proposed by Sapp is Kendall's Tau. Kendall's Tau correlation is a standard methodology used in statistics for measuring the association between two observed quantities; in our case, the loudness or timing of each note for a pair of performances) when they do not follow a normal distribution. This statistic is used to quantify the degree of association between pairs of expressive profiles. Unlike Pearson correlation coefficients, Kendall's tau rank correlation measures concordance on the direction of the change between two points. This provides, using a single metric, a scale-independent quantification of the relationship between rankings of a given variable, regardless of its absolute value (Stamatatos & Widmer, 2005). Having a pair of observations (x_j, x_i) and (y_j, y_i) belonging to two random variables X and Y, they will be concordant when the sort order of both of them (direction in our case of study) agree. And they will be discordant when one of them disagrees. By correlating pairs of performances we can verify how a particular feature (such as loudness or timing) changes for both performances; for instance, the degree of concordance in which two performances being compared make an *accelerando* at the same points in the score. As such, Kendall's tau rank correlations together with the mean and standard deviations can be efficient measurements for

analysis when expressiveness is defined as the deviation from the norm given by the score.

The formal definition for the Kendall τ coefficient is as follows:

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{n(n-1)/2} \quad (1)$$

In the current study, Kendall Tau rank correlations are used to measure different expressive features in order to find out what the consistency between each of the recordings is and to analyze which features may be more representative of the differences between performances. Consequently, Kendall's correlations are calculated for each pair of performances on a note level (event) and on each of the different expressive features selected (timing and loudness).

The Tau coefficient is expressed within the interval [-1,1], with positive values representing agreement between the two rankings measured (e.g. 1 with identical expressive timing between two different performances), and negative values representing disagreement between two rankings. A value of -1 represents perfect disagreement between performances (e.g. a crescendo in one performance and decrescendo in another). A value of 0 represents independence between performances regarding the use of a variable (differences in use are equally balanced between two performances).

2.2.3 *Local Maxima Phrase Detection (LMPD)*

Cheng and Chew (2008b) proposed a methodology to analyze phrasing strategies in expressive performances. They suggest relating local maxima in the loudness or tempo curves with the occurrence of performance phrases (or sub-phrases). That is, the number of peaks (local maximum: M) and valleys (minimum: m) to be found on the tempo and loudness curves of different sections of an analyzed performance. In addition to counting the occurrence of possible local maxima to compare expressive phrasing strategies within the analyzed portion of the piece they suggest three expressive descriptors:

- Phrase strength (S): a measure of the clarity of a given phrase, which estimates the difference between a local maximum (M_j) and the two adjoining local minima (m_j and m_{j+1}) for each note in the score (j).

$$S = 1/2[(M_j - m_j) + (M_j - m_{j+1})] \quad (2)$$

- Phrase volatility: the standard deviation of all phrase strength values within a given music fragment. This is done by measuring "the degree of quantity of variance from the average phrase strength" (Cheng & Chew, 2008a), i.e., the magnitude of variability in phrase strength, from the average phrase strength.

- Phrase typicality: which quantifies the 'popularity' of a phrase by quantifying the proportion of performances that coincide on placing a local maximum at the same point of the fragment analyzed. The more performers coinciding on the placement of local maxima, the more common will be that specific 'expressive gesture'.

2.3 DATASET SELECTION

The music excerpt analyzed here consists of the first twenty-one bars plus the following seven notes (343 notes in total) from the Prelude of Bach's Suite in G Major (BWV1007), that is from the first note to the fermata found towards the middle of the movement. The reader can find a creative commons version of this movement in Appendix a. There are two reasons for choosing this piece. The first reason is that, since Pablo Casals recorded it for the first time in 1936, this Prelude has become one of the most recorded pieces of the baroque solo repertoire and is, therefore, an ideal case study to demonstrate the applicability of analytical methods in music performance. The fact that so many recordings are available makes it possible to study how performers may want to vary their interpretations across time. Within the field of Early Music many studies can be based on analyzing recordings. For instance, we may be able to trace the aesthetical developments and the effect that musicological findings might have had on these performances. The second reason is that the score's regular isochronous rhythmic structure facilitates analysis of different approaches to phrasing without confounding effects from heterogeneity in rhythmic structures annotated in the score. Furthermore, neither the selected score excerpt nor the performances analysed are complicated by any ornamentation.

For the purpose of this study, a dataset of recordings of BWV 1007 played on period instruments was collected. The recordings analyzed were performed by Anner Bylsma, Jaap ter Linden and Pieter Wispelwey, with at least two different recordings of each performer (listed in Table 2.1). In the case of Wispelwey, who has published three recordings of the Bach Cello suites, I chose the two showing a greater difference (his first and third recording), as well as a broadcast (unpublished) live performance. Since assessing the use of period instruments based on the information supplied with the recordings might be misleading (Tidhar, Dixon, Benetos & Weyde, 2014), I estimated the frequency of the first note by visually analyzing the spectrogram frequencies distribution. Within this dataset, the expressive features analyzed are timing and loudness. In the following lines, I will explain how these features were measured.

2.4 DETECTION AND ANNOTATION OF NOTE ONSETS AND LOUDNESS

2.4.1 *Timing annotations*

In order to differentiate (and isolate) notes from an audio source it is necessary to identify the beginning (onset) of each recorded note. While many automatic and semi-

automatic methods to do so have been proposed, detection of onsets of low frequency string instruments remains an unsolved challenge, since the attack of each note is not always clearly discernible (Collins, 2005). After trying several automatic approaches, given the short length of the excerpt chosen, I decided to use a manual annotation approach in order to maximize reliable timing measurements. For the manual annotation I used a procedure similar to the one presented by Robert Ashley, using the graphical audio analyzer Sonic Visualiser to annotate aural and visual cues by hand (i.e., by looking at the beginning of peaks in the spectral representation). After verifying that the number of onsets corresponded to the number of notes being analyzed, data was exported as a time series of onsets for further computation and analysis. The beat-per-minute representation used in the rest of this chapter is measured per note and smoothed over the beat level, in this case a crotchet. The effects of smoothing and possible implications for the representation of timing have been previously addressed by Chew (2012).

Performer	Recording Label	Catalogue Number	Recording date	Duration	Abbreviation
A. Bylsma	RCA	RD 70950	1979	2′12″	B1979
A. Bylsma	Sony Vivarte	S2K 48047	1992	2′49″	B1992
J. ter Linden	Harmonia Mundi	HMU 907216.17	1996	3′19″	L1996
J. ter Linden	Brilliant Classics	93132	2006	3′14″	L2006
P. Wispelwey	Channel Classics	1090	1990	2′27″	W1990
P. Wispelwey	unpublished, AVRO	-	2001	2′30″	W2001
P. Wispelwey	Evil Penguin	EPRC 012	2012	2′06″	W2012

Table (2.1) Dataset analysed on this chapter

2.4.2 *Loudness extraction*

To extract a loudness representation I used the Short Time-Varying Loudness model proposed by Glasberg and Moore (1990), which quantifies loudness (in sones) for each note played. The motivation behind choosing this model was that it accounts for psychoacoustic phenomena such as frequency-dependent hearing thresholds, level-dependent compression and masking, and, therefore, better reflects the perception of the listener. This is obtained using the implementation included in the Genesis acoustics toolbox. [1]

1 http://www.genesis-acoustics.com

2.5 DATA ANALYSIS AND RESULTS

In addition to illustrating the impact of definitions on analysis of expressiveness, with the current study I wanted to assess any consistent differences between the first and second recording of each performer and find out whether this could be related to their characterization or musical individuality. Note that with this set of recordings I do not attempt to represent a holistic 'idea' of each of the performer's musical personality (or individualism) but rather their curated 'idea' about how the piece should be performed at the time of recording. Given the well-established artistry of the performers, the recordings are expected to be the result of a thoughtful compendium of choices made a priori and afterwards curated (in most cases, the recordings have been edited through a post-production process with the final approval of the performer). Therefore, these recordings reflect, if not the individuality of a performer, the expressive choices together with their up-to-recording-date technical ability to represent those choices. These recordings can thus be interpreted to reflect the state of artistry of performers at different moments of their careers.

2.5.1 *Timescapes analysis*

In order to compare pairs of performances, different bell shape correlation timescapes were generated with the online tool available at the Mazurka website. Rather than illustrating this approach with all possible combinations of timescapes (49 plots in this case), Figure 2.3 shows the timing correlations per performer between two recordings.

The differences between the two recordings of Anner Bylsma seem to be more homogeneous (more uniformly darker), reflecting lower correlations and hence a clearly different approach in general timing of the whole performance. On the timescape generated with the recordings of Jaap ter Linden we can observe a clear difference after the middle section and greater differences in the first part (half of the plot) of the excerpt being analysed. The timescape plot of Pieter Wispelwey's recordings shows very pronounced contrasts (black-white) showing clearly different approaches in the phrasing of specific locations in the score.

2.5.2 *Means and Standard Deviations analysis*

Figures 2.4 and 2.5 and Table 2.2 depict means and standard deviations in tempo and loudness for each performance. The differences in the use of tempo between both recordings are more pronounced in the cases of Bylsma and Wispelwey, with Bylsma's recordings exhibiting the greatest mean difference between recordings. No large differences in the use of dynamic means or standard deviations were observed with the exception of the recording of Wispelwey from 2012, but any such differences can be affected by the recording techniques and volume approaches specific to each recording (e.g., microphone placement or signal compression may alter the loudness on the

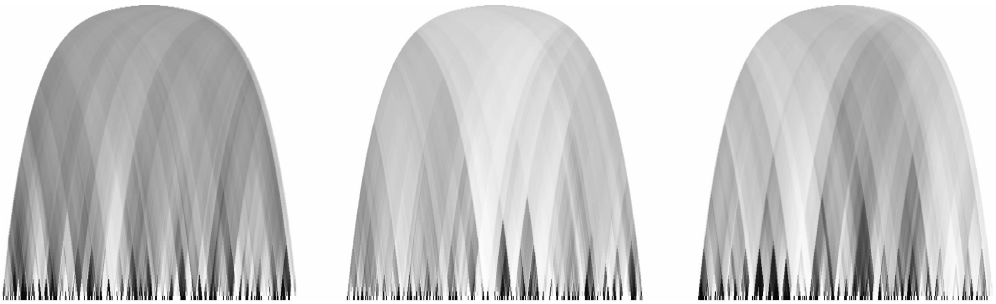


Figure (2.3) Timescapes visualizations between subsequent recordings pairs of (from left to right) Anner Bylsma (1979 - 1992), Jaap Ter Linden (1997 - 2006) and Pieter Wispelwey (1990 - 2012).

recording final product).

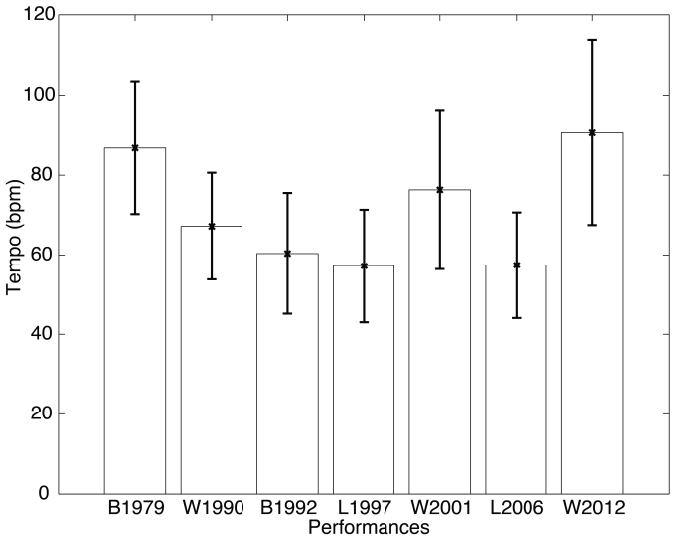


Figure (2.4) Tempo means and standard deviation for each performance.

Despite the small number of performances analyzed we can see a certain trend in the tempo mean through the last decades, interrupted by the recording of Pieter Wispelwey in 2001, who opts for a similar tempo mean as that of Bylsma in 1979. But no intuitive interpretation can be done on the use of loudness through the last decades.

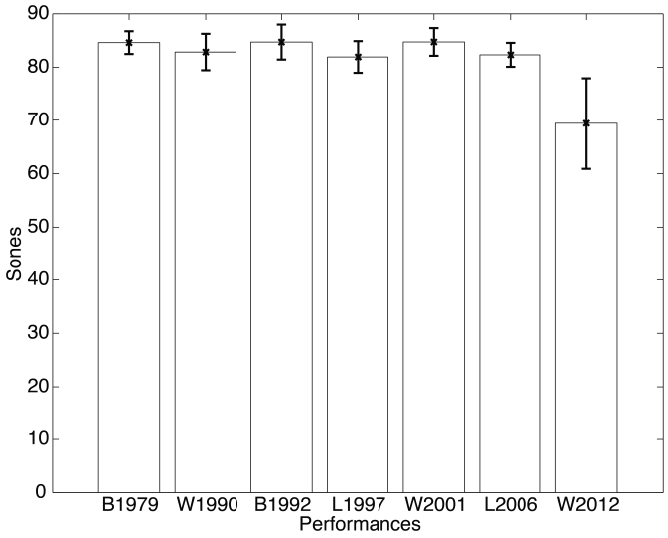


Figure (2.5) Loudness (extracted per note) means and standard deviations for each performance.

	Timing (BPM)		Loudness (sones)	
	Mean	Standard Deviation	Mean	Standard Deviation
B1979	86	12.74	84	2.12
B1992	59	11.82	84	3.26
L1996	56	10.45	81	2.96
L2006	65	10.40	82	2.23
W2001	74	19.99	84	2.59
W2012	88	23.12	69	8.51

Table (2.2) Mean and standard deviations for the performances analysed

2.5.3 *Rank Correlations*

Kendall Tau correlation coefficients of timing and loudness were obtained for each pair of performances, and are listed in Tables 2.3 and 2.4. As we can see in Table 2.3, all the timing correlations are positive and significantly different from 0. It is remarkable to observe that, in most correlations, the coefficient is higher between recordings of the same performer than between recordings of different performers. This might raise questions as to whether the second recordings of the performers might be constrained by different elements involved in the creation of their first recording (for instance, motoric memory), cognitive approaches to their flexibility in the use of expressive features,

or other possible causes.

	B1979	B1992	L1996	L2006	W1990	W201	W2012
B1979	1						
B1992	0.43	1					
L1996	0.25	0.27	1				
L2006	0.34	0.31	0.43	1			
W1990	0.39	0.43	0.39	0.40	1		
W2001	0.31	0.34	0.37	0.39	0.49	1	
W2012	0.37	0.42	0.41	0.35	0.47	0.56	1

Table (2.3) Kendall Tau correlation matrix comparing the Tempo curves of all pairs of performances

In the analysis of loudness, we must be aware that the loudness ranks could also be affected by the use of compression during the post-processing of the recording. Yet we can observe that the rank correlations coefficients are much lower in the loudness measurements than those of timing, with 8 out of 14 pairs exhibiting a negative correlation Tau coefficient. This could mean that the trajectories in loudness are quite different for most of the recordings, probably because the performers are, in this particular piece, not constrained by dynamics (unnotated on the score). Rather, they have diverse phrasing approaches based on their interpretation of the harmonic rhythm implicit in the score, and they have very different approaches in their different performances.

	B1979	B1992	L1996	L2006	W1990	W201	W2012
B1979	1						
B1992	-0.05	1					
L1996	0.11	0.01	1				
L2006	0.04	0.05	0.36	1			
W1990	0.29	0.07	0.15	0.08	1		
W2001	0.004	-0.10	0.12	0.07	-0.05	1	
W2012	-0.06	0.20	-0.008	-0.04	-0.16	-0.03	1

Table (2.4) Kendall Tau correlation matrix comparing the Loudness curves of all pairs of performances

The differences between the rank correlations in loudness and timing could also be due to the fact that the rhythmic structure (isochronous notation through the whole analyzed excerpt) together with the implicit melodic structure of this particular piece allows less variability in the phrasing strategies expressed with timing than the ones expressed with loudness. These data show that performers have a greater diversity in

the expressive directions of loudness than in timing. In fact, the correlations in loudness between first and second recordings of both Wispelwey and Bylsma are negative, indicating different approaches to the use of loudness in the phrasing trajectories for each of the recordings.

2.5.4 *Local Maxima Phrase Detection (LMPD)*

The LMPD methodology presented above was applied. Figures 2.6 and 2.7 depict phrase typicality in counts per bar for both timing and loudness. Figure 2.6 shows much more coincidence in typicality than Figure 2.7, perhaps indicating that timing, as opposed to loudness, is a more relevant feature to differentiate motifs and structure within this particular piece.

The fact that timing is correlated at specific points of the score may be indicative of the performers emphasizing timing in a similar way as they themselves experience the aural transmission of the harmonic context as well as the underlying structure in the piece. In contrast, Figure 2.7 shows very little typicality, consistent with the rank correlation analysis that loudness is a more flexible (or less stable) expressive feature than timing.

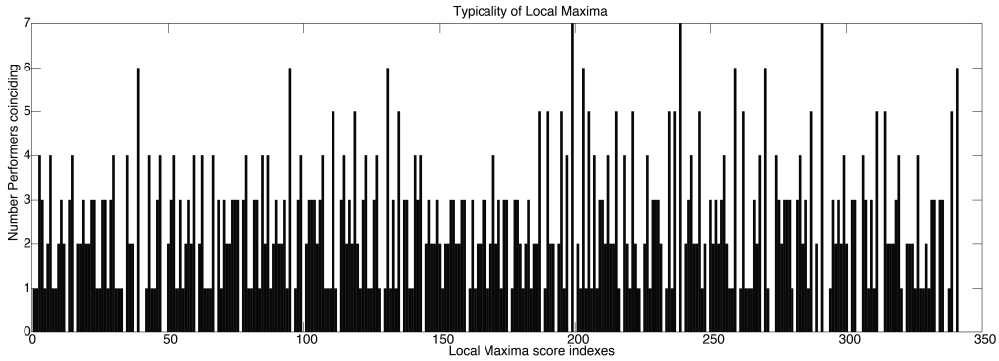


Figure (2.6) Timing Phrase typicality (based on local maxima) within the corpus.

Table 2.5 shows the analysis of phrase strengths and phrase volatility for each of the performances. The correlation between loudness and timing on the phrase volatility measurements is -0.27 while the correlation on the phrase strengths is 0.45. However, the canonical correlation between the Timing and Loudness columns is 0.57 and 0.36. Also, the phrase strengths clearly differ more in the loudness analysis (with a standard deviation of 6.55) than in timing (with a standard deviation of 3.56). Anner Bylsma's recordings show the greatest differences in the number of phrase strengths regarding loudness between the two recordings. The phrase volatility (deviation from the overall mean among all recordings) is much bigger in terms of timing than loudness. This is probably due to the performers using different phrasing strategies, expressed

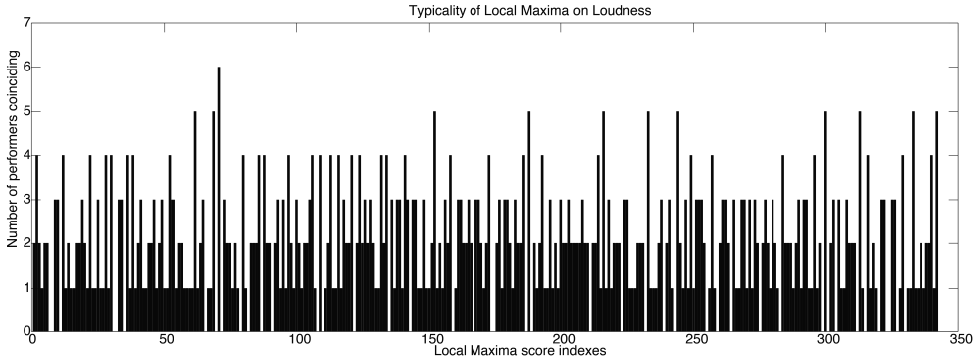


Figure (2.7) Loudness Phrase typicality (based on local maxima) within the corpus.

as lengths in timing, combined with gradual *accelerandi* and *decelerandi*. That is, the arcs length in expressive timing vary more in size than the ones of loudness, yet the loudness is less correlated. This is probably closely related to the harmonic rhythm influencing the expressive features. Depending on the musical context, a performer may choose to emphasize shorter or longer phrasing through timing. The low volatility in loudness can also be explained from a harmonic rhythm perspective. For example, the scalar motive of the pedal progression (the repeated low G during the first 11 bars, see a), together with the metric structure, may be consistently emphasized with loudness, in this way allowing for less volatility in loudness in comparison with timing. However, further experimentation and analysis would be necessary to verify whether the loudness volatility in the dataset presented could have been affected by, other, eventual manipulations of loudness (such as compression) during the recording process.

	Timing (BPM)		Loudness (sones)	
	Phrase volatility	Number of phrase strengths	Phrase volatility	Number of phrase strengths
B1979	12.79	113	1.33	104
B1992	11.17	109	2.32	91
L1996	10.91	102	1.45	97
L2006	10.39	106	1.49	101
W2001	14.88	108	1.73	95
W1990	17.80	107	1.84	95
W2012	10.45	104	4.10	84

Table (2.5) Phrase Volatility and Phrase Strengths for Timing and Loudness

2.6 CONCLUSION AND DISCUSSION

Within the musicological literature, very few studies have dealt with the analysis and comparison of performances of a music piece using more than one recording per performer. The advances in music information retrieval, performance science, and music cognition are facilitating a yet growing development in the field of digital humanities. These developments allow addressing new research questions as well as objectively quantifying some of the expressive features that may play a role in the development of a performance school. Methodologies such as the ones presented in this chapter broaden research possibilities, while perhaps narrowing the possible semantic gap between the subjective anecdotes collected in previous, interview-based performance research and the acoustic facts demonstrated by quantitative analysis. In addition, many cognitive modeling and educational tools can benefit from these approaches in order to improve their development. While most of the differences in performance here discussed may be heard and shared among individuals, having the right computational tools to analyze, quantify, and model different performances may ultimately aid in understanding the cognitive processes involved in the act of listening and performing.

This study, by presenting and applying three alternative methodologies for the analysis of expressiveness of loudness and timing in several recordings of an excerpt of an iconic Baroque composition, has shown that, by combining different methodologies, we can begin to explain the interrelations between performers and their subsequent recordings from a musical perspective, as well as to visualize and measure individual approaches to the expressive strategies used in relation to the score structure. Within this set of performances, it has also been observed that the timing correlation between pairs of performances is, in fact, higher when comparing recordings of the same performer, even when there is more than a decade between the recordings (as in the cases of both Bylsma and Wispelwey). Furthermore, greater differences were found in the correlation per pair of performances for loudness than in the correlations for timing. This implies that, at least for this particular excerpt, performers differ more in their use of loudness as an expressive feature. This finding is in line with previous research done in a similar musical context (Bach's Violin Partita) carried out by Cheng and Chew (2008b).

We must note that, in addition to loudness and timing, many other features (aural and non-aural) play a role in the definition of expressiveness. While the two recordings of Jaap ter Linden may seem to be more consistent than the ones of Pieter Wispelwey or Anner Bylsma in the use of timing or loudness, they might differ much more in other expressive features such as the development in time of articulation and its relation to timbre, an aspect not measured here. Neither has this study reckoned the perceptual validity of our analytical interpretation, except for, indirectly, the psychoacoustic model of loudness used. Within the timing domain, it has been shown that timing is intrinsically linked to tempo, and that observed changes in timing may at least partly reflect

differences in tempo. The representation of timing at a certain tempo may therefore not be generalizable to performances across a broader range of tempi. This claim may compromise the comparison of performances when aiming for an explanation of the perceptual reality of it. This study has rather focused on analyzing the phrase volatility and phrase strengths of the whole excerpt as a quantitative tool, not aiming to explain yet what the perception of these differences is. It is expected however that the 'characterization' of a listener's perception of expressiveness might be related to the perception of features such as phrase strength and volatility of both timing and loudness. Also, probably, these expressive choices are strongly linked to listener expectations biased by the cultural context surrounding these performances, as the greater number of performances coinciding in the same phrase strength and volatility events of a piece, the stronger the expectations of the listeners will be towards these specific events of the same piece. More empirical research is thus needed to perceptually validate the relevance of the proposed measures of expressiveness and to better understand possible hierarchies in performers' choices of expressive features, in the relation between expressiveness and performance characterization, as well as in the possible communication process implied.

While methodologies and techniques are still being developed, there is as yet little uniformity in methodological approach beyond the use of mean and standard deviations. In future work, new methodologies could serve to validate the notion of expressiveness defined intrinsically. For example, deviations from a "norm" could be defined in relation to a performance itself rather than in relation to the score which, arguably, might not be readily accessible to the average listener while experiencing a performance (Grachten & Widmer, 2009). While there are already statistical tools that help to cluster expressive trends within a performance, the fields of computational and cognitive musicology could contribute to obtain an intuitive perceptual and musical explanation. Future work should be done to assess the extent to which performing non-isochronous rhythmic structures may constrain the flexibility of performance with regard to timing or loudness. In addition, more insight is needed regarding possible effects of aesthetics on instrumental practice. For instance, is there a greater variability in timing in the performance of the cello suites played on modern cello than on baroque cello? Is the use of phrasing generally different nowadays than 30 years ago? Could musicians also benefit from these methodologies to develop novel expressive gestures and approaches to their performances of this Prelude? While this is also feasible, a more immediately relevant aim in assessing interrelations between performances is to reveal and understand the fingerprints left by the legacy of music recordings.

3

AN INTRODUCTION TO COMPUTATIONAL MODELING OF PERFORMANCE EXPRESSIVENESS

3.1 INTRODUCTION

As it has been explained in Chapter 2, in order to extract the characteristic patterns of a performance, we need an approach in which such patterns can be found across features and datasets. However, one of the possible shortcomings of the methods presented in the previous chapter is that many of them do not aim to model or reproduce such regularities, but rather serve as tools for visual inspection or post-hoc analysis. When dealing with large datasets and complex systems in which the behavior might only be understood through the interaction of several possible variables, we may instead aim to formulate such behavior by a computational model. A computational model is a systematic description or theory by which some of the factors causing the behavior of a system can be both simulated and computed in order to study this kind of behavior.

Goebl, Dixon and Poli (2005) differentiate between alternative computational modeling strategies in expressive performance, assessing the most common ones as being the analysis by measurement and analysis by synthesis. The **analysis by measurement** strategy includes those models that focus on the analysis of regularities in the performances recorded by humans and converted into a model. Within the modeling approach, the analysis by measurement approach requires a definition of the hypothesis on the behavior of the chosen variables to be verified by any statistical method.

The **Analysis by synthesis** strategy is best represented by the KTH rule system from Friberg et al. (2009), in which several intuitive score based rules are embedded in the model. Among the rules established by this model are the deterministic interactions between loudness and tempo, for which loudness increases or decreases in a linear fashion to faster or slower tempo changes respectively.

In addition, Goebl et al. (2005) also categorizes as modeling strategies the **machine learning** strategy, in which an algorithm aims to extract predefined rules included in a performance, the **case-based reasoning** strategy, which aims to learn some performance based meta rules to be applied to other unknown pieces, and the **mathematical theory approach** proposed by Mazzola and Zahorka (1991) in which the structure belonging to the score and performance features are isolated and decomposed within a mathematical formalization.

The main difference between the machine learning and the analysis by measurement approaches lies in that the machine learning approach has, in principle, no strong assumptions on how the model should behave and, instead, leaves the learning function up to the extraction of relevant patterns. Using analysis by measurement, however, the questions are often very specific, and the behavior of the model is constrained to the hypothesis and pre-selection of their representative examples. Therefore, while one mo-

deling approach focuses on "automatic" discovery, the other has its base on hypothesis testing.

The computational models used in this dissertation are based on a combination of the analysis by measurement strategy and the machine learning (see 3.2) strategy, and they are chosen to achieve an understanding of the relationships between expressiveness and the idiosyncrasy of performers. The models are trained to learn the prevalent 'norm' defined by a performer or group of performers to afterwards study which from the hypothesis drawn regarding the production or perception of expressiveness may get validated by the computational models used.

Once the machine learning model extracts the pattern regularities within a number of performances, we can test and analyse whether an alternative performance (and unknown to the model) may deviate greatly from the predictions obtained through such a predictive model, and how those deviations may relate to the expectations of a hypothetical listener. Thus, the machine learning methods used in the following chapters are chosen to model the idiosyncrasy of a performer but are also intended to be a basis to elucidate how this idiosyncrasy relates to a listening process *"in generating the subjective experience from the perceptual input"* (Pearce, 2011).

The aim of this chapter is to motivate the use of machine learning as well as provide a basic understanding of the methods used in Chapters 4 and 5 of the thesis. For a more detailed discussion on the machine learning methods, see Bishop (2013); Lipton, Berkowitz and Elkan (2015); Mackay (2003); Mitchell (1997) and Goodfellow, Bengio and Courville (2016). In section 3.2 I will introduce general notions of machine learning, in section 3.3 I will explain the basics of neural networks, and in 3.4 I will elaborate on the strategies using machine learning for the modeling of idiosyncratic patterns of expressive performance.

3.2 MACHINE LEARNING BASICS

Machine learning is a field of Computer Science in which the goal is to learn patterns or behaviors from data by using statistical techniques and algorithms. Given a performance measure and a task, the performance of the algorithm in solving the task will improve with experience (by learning) (Mitchell, 1997). The learning in such algorithms thus results from the adjustment of their internal parameters to the data they are trained with.

In this dissertation, the machine learning algorithms used will learn to predict loudness or tempo after having been trained to recognize regularities in the way of, *e.g.* a performer or a group of performers using them. Conceptually, based on the exposure of the model to the recordings contained within the training dataset they will learn the most common ("heard") version of expressive gestures within it.

Depending on the complexity of the model and how the parameters are tuned, the model may be able to predict better or worse the characteristics and behavior of unseen data. This is what is commonly referred to as generalization. Having, accordingly, a model in which the parameters are very well fit or too poorly fit may lead to overfit-

ting or underfitting when being exposed to new data. Overfitting refers to the case in which the model is very accurate in predicting the data with which it has been trained but it might be very inaccurate when being exposed to unseen data. That might be because the model is memorizing the data instead of recovering generalizable patterns. In such cases, it may perform very badly when trying to predict sequences with slight variations in the input.

Figure 3.2 shows an example of how the generalization power of a polynomial regression model may be affected by overfitting or underfitting the degrees of the polynomial features (and, thus, its complexity as a model). On the leftmost plot from Figure 3.2 we can see how the data is underfitted (and biased) as a consequence of using a polynomial of degree 1. On the contrary, as we can see on the rightmost plot from Figure 3.2, the higher the degree of the polynomial, the greater the variance of the model, which translates in the model being very sensitive to the noise of the data rather than the general properties of it; seen in the True function (orange line). The plot in the center from the same figure, shows presumably the "best" fit possible as it coincides almost exactly with the underlying True function (in this case, a cosine function) of the data.

Thus, having a model in which its parameters are too biased or have too much variance tuned and fitting very well the data of the training set may lead to poor generalization. Other reasons for having a poor predictive model may be related to having too many (in the case of overfit) or too few (in the case of underfit) training iterations given to the model to learn "good enough" parameters values, rather than those that fit perfectly the training data but may deal very poorly with unseen data.

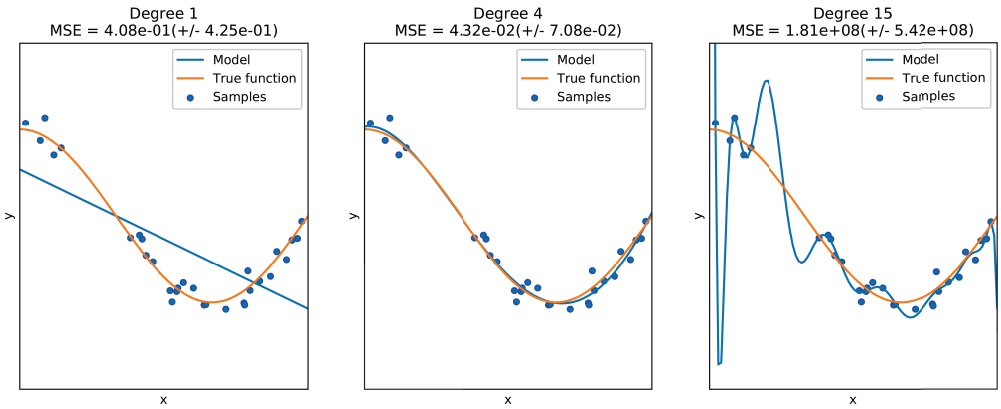


Figure (3.1) Plots in the left and right show, respectively, underfitting and overfitting in the linear regression model due to an increase of the degree of polynomial features.

1

The machine learning algorithms used in this dissertation are chosen to learn which underlying function (f) fits better the relation between the input features (X) and the output (Y) being able to generalize to unseen data. These kinds of algorithms belong to the category of **supervised learning** and they are called supervised as the target or

label from which they have to learn is always predefined. This means that examples $(x, y) \in X \times Y$ are provided. The underlying function $f : X \rightarrow Y$ can be learned by a deterministic or a probabilistic model.

An alternative approach to supervised methods in machine learning is that defined by unsupervised algorithms. These algorithms capture the characteristics, features or structure of a dataset by learning the probability distribution or associations that may have generated the data itself and, as such, reveal patterns in the data. In **unsupervised learning**, there is no corresponding Y from which to learn such a relationship. In contrast, they must be able to "learn" what belongs to the structure of the data (X), from what is unstructured noise contained within it. A common example of unsupervised learning is that of clustering, by which the data given is grouped as similar or dissimilar based on a proximity (distance) measure.

As an example of a supervised learning method and in order to introduce some of the general concepts in machine learning, in the following section, I will explain linear regression.

3.2.1 *Linear regression*

In linear regression, the goal is to learn the function between the input variables (X_n) and the target variable (Y) by fitting a regression line that best approximates the input data points. For doing so, Linear regression needs a parameter, (w_1), that indicates the slope of the line, and an intercept, w_0, that indicates the point in which the regressor line crosses the ordinate (Y axis) of the graph.

Simple linear regression is expressed as:

$$Y = w_0 + w_1 * x_1 \tag{3}$$

In this case, the intercept term, $x_0 = 1$ (therefore omitted in the equation). The regression line passes through the Y axis based on the value of the bias term w_0, which is often represented as well as b or β. Not having an intercept (also called, bias term) would imply that the regressor line would pass through the origin $(0, 0)$, which compromises the fit of the function and predictive power. Each of the w represents a weight, or parameters, over which the linear relation is defined. w_1 represents the change in Y divided by the change in X. In the case of linear regression, the parameter (or parameters in the case when there would be more than one feature) w is linked to an input feature of the experiment, to analyzing its behavior in respect to the predicted variable (Y) .

When we want to include more input variables (features), this is expressed as:[2]

$$Y(X) = w_0 + w_1 * x_1 + w_2 * x_2 = \sum_{i=0}^{n} w_i x_i \tag{4}$$

In linear regression (and many other machine learning algorithms) when including categorical features in the input (those that take a fixed number of values), a transformation is necessary in order to map the categories to input vectors. This transformation is often done by using one-hot encoding, in which each of the columns' features will be represented by a binary vector in which each of the combinations represents their absence or presence. For example, if we only had two markings in the whole corpus: *f* and *p*, we would have a vector for these features in which *f* would be represented as [1, 0] and *p* as [0, 1].

I exemplify this in table 3.1, with a hypothetical performed melody of four notes for which we want to predict Tempo based on Loudness when played by two different performers (A and B). In this toy-example, we model the relation between the use of expressive loudness in a piano performance and the use of tempo. In particular, we study the effect of expressive loudness on tempo when a piece is played by either Performer A or Performer B. Both belong to the categorical variable Performer. In this case, the (dummy) variable Performer B indicates whether this performer is playing or not, and Performer A (being the reference performer) is represented by 0. The toy sequence illustrated consists of 4 notes concatenated with values for Tempo and Loudness as shown in the table.

Tempo	Loudness	Performer B
1.4	1.4	1
1.14	1	1
0.97	2.4	1
1.52	1.6	1
1.35	2.3	0
0.97	2.4	0
1.52	1.6	0
1.35	2.3	0

Table (3.1) Toy example of regression of Loudness on Tempo differentiating between two performers (categories) playing the same musical excerpt (4 notes). The variable Performer is used to indicate to which of the two performers correspond the values of Loudness or Tempo. When the values in the Performer columns are 0, the weight w_2 is canceled.

2 The expression $\sum_{i=0}^{n} w_i x_i$ can also be often found as $w^T x$, in which w^T refers to transposition.

$$\text{Tempo} = w_0 + w_1 * \text{Loudness} + w_2 * \text{Performer} \tag{5}$$

In regression tasks, the measure most commonly used to evaluate the predictions of the algorithm, is the mean squared error (MSE) between the true values, Y, and the values predicted, $\hat{Y}$. In this study, Y is represented by those belonging to each of the performances of the dataset and $\hat{Y}$ by those predicted (generated) by the model. Due to the square, MSE emphasizes the extremes by making the large errors larger and the small ones smaller.

The mean squared error is defined as:

$$\text{MSE}_{\text{test}} = \frac{1}{m} \sum_{t} (\hat{y}_{\text{test}} - y_{\text{test}})^2 \tag{6}$$

A complementary measure to MSE found at times in literature is Mean Absolute Error:

$$\text{MAE}_{\text{test}} = \frac{1}{m} \sum_{t} |\hat{y}_{\text{test}} - y_{\text{test}}| \tag{7}$$

3.2.2 *Gradient descent*

The ultimate goal of a linear regression algorithm is to find the appropriate values for the parameter weights so that, in a hypothetical dataset containing only a dependent (y) and an independent variable (x), the regression line may be as close as possible to all the points plotted in the $X - Y$ graph.

While a simple equation, such as the one presented in (3), can be solved by using the "analytical" method, the amount of computational power will suffer from employing such a method when increasing the number of features (in X). An alternative to the analytical method is numerical optimization. Within the "numerical" methods, a common optimization technique used to minimise a differentiable Loss function (such as MSE) by updating the model parameters, is gradient descent.

During the iterative optimization in gradient descent, the Loss function ($J(w)$) is used within the training set of the data to minimise the error between the output values ($\hat{Y}$) predicted by the model and the true values of the data (Y). As such, the parameters' values will be updated to minimise the Loss (or Error) until a local minimum is found.

Generally, the gradient of a function indicates in which direction the function grows more. Therefore, the iterative optimization of the gradient descent algorithm works by updating the weights *W* in a direction opposite to the gradient ($\nabla_w J(w)$) of the Loss function with a step size (or learning rate) η. Conceptually, the algorithm descends (by

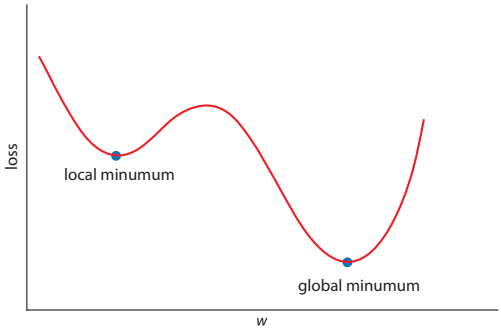


Figure (3.2) Local and global minimum for a Loss function

iterative steps) until it finds parameters that minimize the Loss function. This minimum could, however, be local in the general case. In the example of linear regression, we would start by randomly initializing parameters and perform iterations of the gradient descent algorithm until convergence.

For the error measure in regression tasks, the mean squared error (MSE) is usually preferred as a Loss function over the mean absolute error (MAE), since it can be more easily differentiated.

The update performed during gradient descent is defined as:

$$w = w - \eta \nabla_w J(w) \tag{8}$$

Figure 3.3 shows the plots of the different weight values chosen for a Gradient Descent. The equivalent values are shown in Table 3.2. As we can observe in this example, the update of the weight values has a decreasing effect on the Loss function.

Under sufficiently general conditions, the gradient descent algorithm will converge to a global minimum if the Loss function is convex, or to a local minimum if it is non-convex. The convergence or divergence will also depend on the value of the learning rate, since it will make the steps towards the minima bigger or smaller.

A commonly used alternative algorithm to the gradient descent (GD) is stochastic gradient descent (SGD). SGD follows the same principle as GD, with the difference that it samples data points one at a time. Thus, having $(x_i; y_i)$, where $_i$ indicates a single sample, the weights w are updated after each training sample based on the gradient of the error on that (single) training sample. One of the main reasons why SGD is preferred to GD is that it converges much faster and it has often been found to converge to better local minima (Ng, 2012).

The update performed during stochastic gradient descent is defined as:

$$w = w - \eta \nabla_w J(w; x_i, y_i) \tag{9}$$

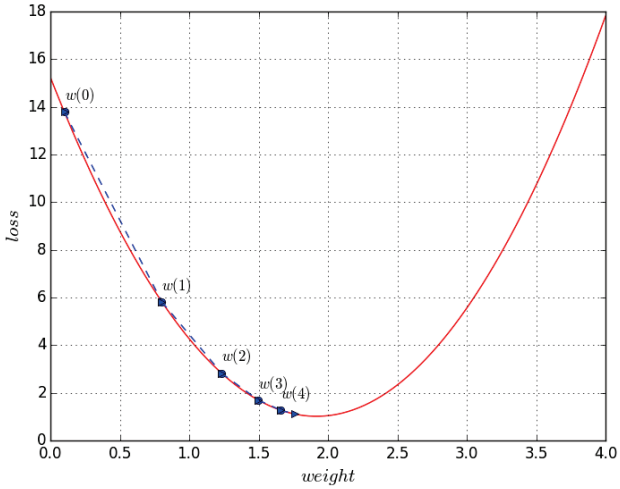


Figure (3.3) Gradient descent updates on Loss function (J) plotted against one weight (w). Learning rate of $\eta = 0.5$. The dashed line indicates the update of the weight per iteration. Table 3.2 shows the correspondent weight updates values.

weight	loss / error
w(0): 0.1000	13.6197
w(1): 0.8139	5.3998
w(2): 1.2515	2.3107
w(3): 1.5197	1.1499
w(4): 1.6842	0.7136
w(5): 1.7850	0.5497
w(6): 1.8468	0.4880

Table (3.2) Weight updates per iteration (0-6) and resulting error values from Figure 3.3. As we can see, the weight updates has a decreasing effect on the Loss function.

With the gradient descent algorithm, we can find local minima in the Loss function curve. In the case of linear regression, this is not a problem, as having more observations than predictors will lead to a convex solution in which the global minimum is guaranteed. In non-linear methods, such as multi-layer perceptrons, there is a risk of having a gradient descent getting stuck in local minima, as the Loss function of such methods is often neither convex nor concave. While an argument on the need to find the global minima is to be able to find the true minima of the Loss function, a counter-argument is that finding such global minima may lead to overfitting the learning function and, consequently, to poor generalization. The randomness introduced

when using SGD aims to avoid getting stuck and reach convergence at 'better' local minima.

Several modifications to the Gradient Descent and Stochastic Gradient Descent algorithm have been proposed with the goal of finding better local minima or speeding up the optimization process (very relevant when training large sets of data). The modification most commonly found in recent literature on neural networks is ADAM (Kingma & Ba, 2015), which is used within the predictive models in Chapter 5.

3.2.3 *Parameters and hyperparameters*

When using machine learning algorithms, we must make a distinction between 'parameters' and 'hyperparameters'.

The parameters of the models are those that need to be learned (optimized) from the data by (re-)training the models to existing (or new) data. The behavior of the algorithm in relation to the parameters often leads to categorizing them as either parametric or non-parametric. In general, parametric algorithms have a fixed number of parameters that determine the capabilities of the algorithm to learn an underlying function. Non-parametric models assume that the data distribution cannot be defined in terms of a finite set of parameters. Therefore, the number of parameters are potentially infinite and will grow as the amount of data grows. Yet, the classification of algorithms as parametric or non-parametric responds to a historical terminological convention, nowadays often blurred as the models get more complex and entangled and their categorization, as such, less clearly demarcated.

The hyperparameters, instead, are those variables fixed and predefined before the model training and parameters optimization. They are related to higher-level properties of the model (such as the value of the bias). Thus, the hyperparameters are not determined by the learning algorithm itself, and are often chosen based on performance on a validation (e.g., using Grid Search Cross Validation; see Section 3.2.4).

3.2.4 *Model selection*

In order to prevent overfitting the models, and with the goal of making them generalizable, it is customary to split the dataset into train, validation, and test sets. The train and validation splits of the dataset can be used to tune and select the best possible models (hypothesis functions). The test set is left for the very final stage of the experiments and evaluation of the model, and it should never be used during the model selection process.

Having a range of values available as hyperparameters, the goal of dividing the train set into train and validation is to run the models through different combinations of hyperparameters in order to choose those values that lead to the best performing model on the validation set.

The size of the dataset splits may condition the model performance and results obtained. The splits on the dataset are sensitive to the following trade-off: reducing the

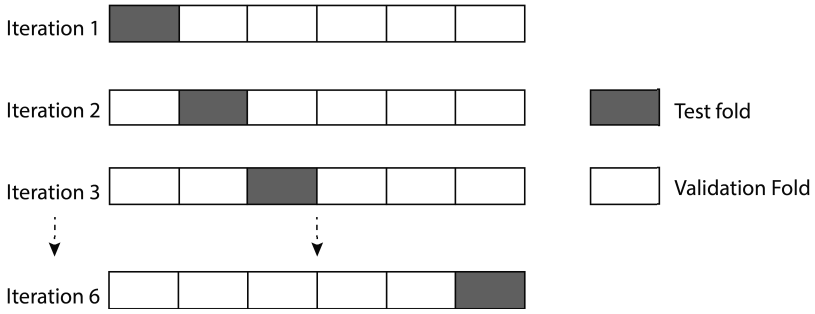


Figure (3.4) Cross Validation through 6 equally divided folds of a dataset

size of the train/validation set often leads to greater variance in the parameter estimation, while reducing the size of the test set may lead to greater variance in the predictive performance. Unfortunately, in many of the machine learning models available, there is no perfect solution to deal with the splits of the dataset.

The adjustment of hyperparameters can be done by different methods. The most common one is Grid Search, which works by doing an exhaustive search over specified hyperparameter values for an estimator. Having a grid of values for the hyperparameters, by doing Grid Search we can evaluate the Loss function on the validation set after trying each possible combination of hyperparameter values.

A common alternative method to Grid Search, is Random Search. In Random Search, a fixed number of parameter settings is randomly sampled from a given distribution. If the weight of the hyperparameters in the Loss function is known, a Grid Search will be more accurate and effective than a Random Search. Yet, this is rarely the case, and recent literature has shown that when the weights values are unknown, using random search might lead to better fitting models (Bergstra & Bengio, 2012). Furthermore, exploring the sample space delimited by the Grid Search will exponentially increase with the number of values specified within the Grid. Thus, using Grid Search might be only convenient when the size of the Grid is rather limited and an intuition or experience on the response of the model to the values given is available.

When having a small dataset, we can use Cross Validation in the train and validation set to better exploit the data and aim for a more robust and generalizable model. The Cross Validation technique consists in partitioning the train/validation dataset into k splits; having a validation set corresponding to split X_{k_n} and a train set of $X_{k_n} - k_n$. The evaluation process is iteratively repeated by each of the K corresponding folds having a record of what the performance of the model is in each of the iterations done. The number of k-folds will be determined by the size of the dataset, being inversely proportional to its size. That is, the smaller the data, the more folds should be present. The number of k-folds, however, can also be determined by the computational power demanded by the algorithm and time constraints of the tasks to be carried.

Cross Validation is often combined with Random or Grid Search to speed up the process of model selection. If the data to be modeled is independent and identically distributed (i.i.d.), the k-fold cross validation can be randomly combined and split. This is not the case when modeling sequential data in which we must respect the order of events in the dataset (in our case, for instance, the sequence of notes in the melody). In such cases, the validation split must always follow the training split.

In our case of study, the mean squared error (MSE) is used as both a Loss function (as part of the gradient descent) over which we can evaluate the performance of the network with a set of hyperparameters values during the train/validation step, but also a measure over the test set during the final evaluation of the model.

After Cross Validation, once the best fitting parameters are chosen, the models' generalization must be evaluated in the held-out set (or test set). In order to be rigorous with the evaluation of the final model, the test set should always be left out of the model selection process. This is because the hyperparameters will be tuned according to the train and validation sets. Therefore, the performance of the model would be biased and give a less trustable scientific output.

To test the hypothesis based on the results of the models on the 'test' set (for instance, to evaluate whether a combination of features may lead to better predictions than another combination of features), we can use, for instance, a t-test or Wilcoxon test (Wilcoxon, 1946), which tests the null hypothesis that two paired samples come from the same distribution. The samples to be compared are the MSE errors obtained for each of the experiments predictions. Being non-parametric, the Wilcoxon test makes little assumptions about the probability distributions and it does not assume normality in the distribution of the population.

The resulting curve obtained from subtracting each of the data points in Y from $\hat{Y}$ is known as the residual error. What is contained within the obtained residual (either idiosyncratic properties of a signal or just random noise) will depend very much on the choices of the model and, fundamentally, on the properties of the data being modeled.

3.3 NEURAL NETWORKS

Neural networks (also referred to as connectionist models) are a type of computational models originally inspired by some of the biological principles of the neural behavior in the brain. Conceptually, a neural network is represented by a set of artificial neurons or nodes (or units) connected by a set of directed edges, which represent the biological synapses between the neurons (see Figure 3.5).

One of the first approaches to model a Neural Network was proposed by McCulloch and Pitts (1943). In their model, a neuron may receive several inputs from an external source X, or from other neurons. Each input x_i is multiplied by a given weight w_{ki} on the correspondent edge and connected to a sum junction which adds all the weighted values. The sum junction may also receive an input of a bias component. Finally, the resulting value from the summing junction is "filtered" by an activation function ϕ that serves the purpose of limiting the amplitude of the output of the neuron. In the

original model from McCulloch and Pitts (1943), the activation function consisted of a step function, but in later developments of Neural Networks architectures, other activation functions are used. In fact, having a Neural Network with a single node and a linear activation function is equivalent to the linear regression explained in Section 3.2.1.

The single layer perceptron neural network can be represented therefore as:

$$\begin{aligned} z &= \sum_{i=0}^{n} w_i x_i \\ Y(x) &= \phi(z) \end{aligned} \tag{10}$$

The most common outer activation functions (applied in the output node(s)) are:

- Softmax exponential function, which is often used for multiclass classification (with k nodes in the output layer)

$$\hat{y}_k = \frac{e^{a_k}}{\sum_{k'=1}^{K} e^{a'_k}} \tag{11}$$

- Sigmoid function, which is used for multilabel classification

$$\hat{y}_k = \frac{1}{1 + e^{-k}} \tag{12}$$

- Identity (or linear) function, which is used for regression

$$\hat{y}_k = k \tag{13}$$

In such a single layer perceptron model, depending on the task, there may be several neurons at the input layer connected to one or more neurons in the output layer. The way in which the perceptrons are connected together is commonly referred to as the neural architecture. Choosing for the right architecture within the neural network will be decisive in its behavior and, consequently, in the results obtained for the given task.

3.3.1 *Feed-forward neural network*

Feed-forward networks are a type of neural network architecture in which the perceptrons are arranged in layers. A feed-forward network including one or more hidden layers is known as a multi-layer perceptron (MLP). In an MLP architecture, all neurons

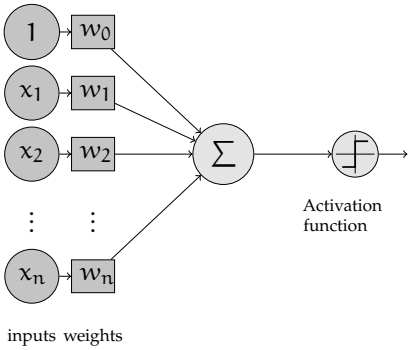


Figure (3.5) Model of a neuron according to McCulloch and Pitts (1943)

on each layer are connected to all neurons of other layers, but not to any others of the same layer. Therefore, the information from the incoming input is fed forward across the defined layers towards the output layer.

In an MLP, each neuron in the hidden layer has as well an activation function. The "inner" activation function is often different from the one in the output layer in both the role within the topology and the behavior. The most common inner activation functions found in the literature are the sigmoid function, the hyperbolic tangent (*tanh*) function $\phi(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$ and the rectified linear unit (Relu) function $f(x) = max(0, x)$.

In feed-forward neural networks, the weights update is done with the backpropagation algorithm (Werbos, 1990). This algorithm, calculates the partial derivative of the output error with respect to all weights in a backwards fashion. That is, starting from the error obtained at the output, it updates (differentiates) iteratively the weights going backwards through each of the layers towards the first layer or input.

In the backpropagation algorithm, the gradient descent calculation is done with the chain rule, which allows for decomposing a derivative as a product of its individual functional parts and, like that, calculate the derivative of the Loss function with respect to each weight (parameter) within the network. Thus, keeping a record of the differences within every connection during the forward pass, the algorithm calculates the gradient by means of the chain rule based on the error propagated backwards.

In order to infer the best weight values within the neural network, the process of feed-forward computation and backward propagation has to be repeated several times. Each cycle of completing both steps is referred to as an epoch. The number of epochs needed to find convergence in the neural network depends on the data characteristics and the neural networks architecture.

Some of the hyperparameters that are often tuned before or during the cross-validation to avoid overfitting are:

- Number of hidden units

  The number of units (or neurons) included in the hidden layer can also have an impact on the generalization capabilities. Therefore, trying different sizes in the

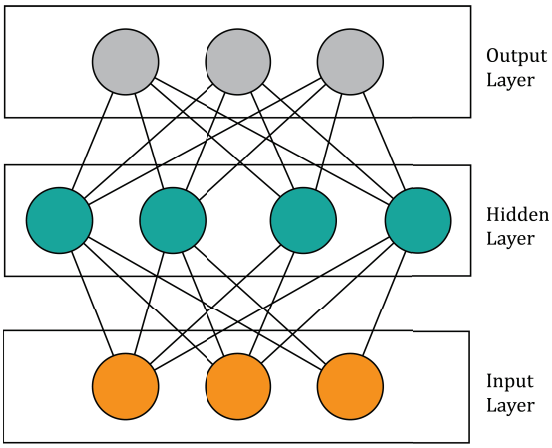


Figure (3.6) Multi-layer perceptron

hidden-layer during the hyperparameter tuning is a relevant step. Normally, the number of hidden units should be in between the number of units at the input and output layers, or the mean value between both.

- Regularization

  By regularization, we aim to prevent overfitting by reducing the complexity of the neural network. As shown in Figure 3.2, if the magnitude of the weights is too large, it tends to cause overfitting. Common choices for regularization are L2 regularization and Dropout.

  – L2

    L2 penalizes large weight values by summing the square values of all weights and multiplying them by a (hyperparameter) scaling factor, that controls for their magnitude. This is effectively equivalent to parameters weight decay during learning/optimization. It is defined as:

$$R(f) = \frac{1}{2}\lambda \sum w^2 \tag{14}$$

    L2 then it is simply added to the Loss function:

$$MSE_{test} = \frac{1}{m}\sum_{t}(\hat{y}_{test} - y_{test})^2 + R(f) \tag{15}$$

  - Dropout

    Dropout consists on specifying a percentage of units per layer to be randomly deactivated during training. These units and their connections are, therefore, dropped out randomly on each of the epochs during training, which prevents the network of depending too much on specific networks and in order to have a more balanced representation within the layers. In the neural networks herewith presented, all units are used at test time when applying Dropout.

- Learning rate

  The learning rate (η) hyperparameter, used in the gradient descent (equations 8 and 9) is used to define the amount of learning on each epoch during the training phase. The value used for the learning rate usually is very small.

- Momentum

  Momentum is a method used to accelerate SGD in the right direction. The momentum hyperparameter specifies how different the values of the weights should be on each epoch in order to find convergence. The role of the momentum is to damp the oscillations of the gradient descent progressing slowly to the minimum.

$$\begin{aligned} v_t &= \gamma v_{t-1} + \eta \nabla_w J(w; x_i, y_i) \\ w &= w - v_t \end{aligned} \tag{16}$$

  The momentum term (γ) (Qian, 1999) updates the vector v_t in the gradient descent by a fraction of the previous time step in the gradient descent (Ruder, 2016).

  Momentum is applied with the aim of preventing the gradient descent of getting stuck at local minima, and letting the gradient being more effectively adapted through each of the epoch iterations. That is, with the momentum, the gradient descent is extended to account for the time steps. It complements the zig-zagging descend (towards the minima) approach of the SGD by affecting its speed and consequently also the positioning of the descending particle during the zig-zagging.

  Other variants of the momentum are the Nesterov momentum, which pre-calculates and evaluates the momentum by approximating the position of the next step in the descent, and the first and second vector moments of the ADAM gradient descent (Kingma & Ba, 2015).

3.4 MACHINE LEARNING AS A PERFORMANCE MODELING STRATEGY

The performance modeling strategies defined by Goebl et al. (2005) (see Section 3.1) differentiate between the analysis by measurement strategy and the machine learning

strategy. That is, on whether the research approach departs from a hypothesis in the performance rules, or whether the performance rules are discovered by the machine learning algorithm. The methodology and machine learning algorithm used to discover such rules may, however, challenge such demarcation.

Goebl (2001) cites the study of Widmer (2002) in order to illustrate the "machine learning" approach. In his paper, Widmer (2002) proposes a supervised classification algorithm to categorize a number of performance rules predefined by a combination of both a music theory model of melody (Narmour, 1992) and some categorized abstractions in the use of timing, dynamics and articulation. Such abstractions, however, contain several arbitrary assumptions in the categorizations. For instance, according to this categorization, articulation will be considered *staccatto* if the ratio between notated duration and performed duration is smaller than 0.8 and *portato* if the same ratio is longer than 1.0. This rule based discretization can also lead to the following assumptions: 'When performers play with a ratio longer than 1.0 it is because they intend to play *portato*, or, 'When performers play with a ratio longer than 1.0, listeners perceive it as *portato*'.

We should be aware that, in this view, the rules discovered could also be an artifact of their categorization rules, rather than of the performer's playing, or how this is perceived by the listener. If that would be the case, the rule based discretization of such continuous variable may therefore not correspond with the reality (and intention) of the performance (or its appreciation), and rather be contextual. For example, what is considered *staccato* in a performers style by such rule, could as well be a technical artifact resulting from the performers adapting their playing to aspects such as the tessiture (in relation to loudness), the characteristics of the instrument used for its performance or the acoustics of the room in which it is recorded (when related to articulation). Hence, while Widmer's study is a very relevant contribution in the field of performance modeling, his rule based features approach is not completely agnostic, and the interpretation might be limited (or even biased) by the categorizations boundaries. That is, the categorization process already includes several pre-defined assumptions on what can be "discovered" by the algorithm. On the other hand, this raises the question of whether we may expect an algorithm to find interpretable "norms" when the encoding input and output of the algorithm may not be predefined to us. In many unsupervised algorithms, this is indeed one of the main challenges we often have to deal with.

3.4.1 *Previous uses of machine learning to model performers idiosyncrasy*

A number of studies in the topic of expressive performance idiosyncrasy have used a combination of the analysis by measurement strategy with machine learning methods. Repp (1992) used principal component analysis to differentiate common phrasing structural dependencies from "eccentric" ones in a set of 28 performances played by 24 pianists of Schumann's *Träumerei*. Madsen and Widmer (2006) proposed using a self organizing map, which is a type of unsupervised neural network algorithm with an

alternative method to backpropagation (namely, competitive training of neighbors) to calculate "performance archetypes" based on string matching. Stamatatos and Widmer (2005) introduced an ensemble learning approach to automatically classify performers based on their use of timing and loudness. Grindlay and Helmbold (2006) used a (sequential) Hierarchical Hidden Markov Model to predict performers individuality and synthesize expressive performances after training on a professional pianist.

Within the automatic classification of performers, Ramirez and colleagues also applied different machine learning algorithms to classify saxophone performers playing jazz standards (Ramirez, Maestre & Serra, 2010) and violin performers playing classical music (Ramirez, Perez, Kersten & Rizo, 2010). Finally, Gingras, Asselin and McAdams (2013) used Linear Mixed Models to disentangle piece and performer dependencies in the use of timing by twelve performers playing three different harpsichord pieces.

3.4.2 *Neural networks in expressive performance modeling*

The use of neural networks for expressive performance modeling or analysis is not abundant. Bresin (1998) trained a multi-layer perceptron on human performances of Schumann's *Träumerei*. In his study, it was shown how this model could generate expressive performances by learning some of the rules included within the KTH system (Friberg et al., 2009). Jacobs and Bullock (1998), used a neural-network model to generate legato articulation when playing scales and arpeggios. Serrà et al. (2013) used, among several other methods, a neural network to test the hypothesis of onset timing deviations as a feature to automatically classify music pieces.

Concerning rhythm categorization and timing discretization of expressive performances, other studies that used connectionist models include: Large (1996) and Eck and Schmidhuber (2002), using neural oscillators to infer a metrical representation, and Desain and Honing (1992), in which recurrent attractor networks are implemented to quantize expressive timing to a score representation and as a perceptual model of rhythm categorization.

With the advances of deep learning in the last decade, the use of neural networks has been extended to performance modeling. In addition to the study included in Chapter 5 of this dissertation, other recent applications of deep learning to expressive performance modeling have been developed at the same time I carried this research. Malik and Ek (2017) propose using Long Short Term Memory (LSTM) networks (introduced in Chapter 5 of this dissertation) to generate performances focusing on the prediction of dynamics and tempo on different music genres. Oore, Simon, Dieleman and Eck (2017) use LSTMs to generate improvised (the score is also composed by the model) expressive performances based on timing and dynamics. To the best of my knowledge, no current research has been done including LSTMs and idiosyncratic expressive performance modeling, which is the topic of research in this dissertation.

The research in performance modeling suggests that the use of machine learning methods combined with analysis by measuring is a practical approach to study the elements that define idiosyncrasy in a performers style. Using the right computational

models to find the constraints and regularities in the production and perception of such expressive characteristic gestures, might help to further understand expressiveness in music as an essential element in the communication process. In the following chapters, I will present research done using machine learning models to extract performance patterns in the collective and individual use of tempo and loudness.

4

MODELING TEMPO AND LOUDNESS EXPRESSIVENESS AT SCORE MARKINGS

4.1 INTRODUCTION

In common music notation, in addition to the score-notated pitches and rhythms, composers may use score markings to emphasize expressive changes or to reinforce the expression of a music passage or section (Grachten & Widmer, 2012). Thus, they are an extra resource to indicate to performers a composers' idea on how the score should be performed and how such markings may relate to the piece structure.

Within the western music repertoire from the Baroque period until the first half of the XXth century, composers have been progressively including more expressive markings and instructions for performers. The development of expressive markings notation has been parallel to that of the expressive possibilities of the music instruments as well as to the popularization of the musical score as a written medium (Chanan, 1994). In this regard, the inclusion of expressive markings in the score suggests an aim to narrow the gap in the communication process between composer and performer as described by Kendall and Carterette (1990). Elucidating how the expressive constraints defined by a score may affect the performance of that particular piece is necessary to understand both the idiosyncratic expressiveness of performers, as well as the collective stylistic approaches to the interpretation of music.

Most expressive markings (in western music) are intended to guide the performer during their practice and execution. These markings suggest how to make use of tempo, timing, dynamics or, even pitch and timbre. In the performance of western music, it is common to find styles in which characteristic expressive gestures result from the combination of different expressive dimensions. An example of these expressive gestures can be found in the use of *ritardando* on structural cadences, in which timing and loudness gestures are often varied in a correlated fashion. For instance, the notes belonging to the Vth and Ist chords are slightly slowed down while the notes belonging to the Vth chord are stressed in loudness. This is often done in the performance of Classical and Romantic western art music to resolve the tension between the dominant (Vth) and the tonic (Ist). The romantic repertoire is commonly characterized by the performers' use of *tempo rubato* and large dynamic contrasts at structural points of the compositions (Benetti Jr., 2013).

In anticipation of the dataset to be described in Section 4.2, in Figures 4.1 and 4.2 I illustrate the possible relation between score markings and expressive performances. In particular, these figures (4.1 and 4.2) show the average obtained in tempo and loudness between two different Chopin Mazurkas played by 11 pianists. In those Figures, one can see that the relation between both expressive curves suggests being correlated to the performance structure as defined by either tempo or dynamic markings. The

figures also suggest that, in some pieces, tempo and dynamics interact at structural points of the compositions. These observations suggest that modeling the combination of high level expressive features (such as loudness and tempo at score markings) may lead to more complete performance models, as well as bring insights into the idiosyncratic approaches of performers. Furthermore, studying how the structure of a score is related to the interaction (and possible dependencies) between expressive dimensions, may add to explaining the mechanisms behind the interpretation, perception and categorization of music (Palmer, 1996).

In order to investigate the interactions between tempo and loudness, Pampalk, Goebl and Widmer (2003) developed a visualization tool that allows combining loudness and tempo performance trajectories into a two-dimensional representation. The same visualization tool was later used aiming to identify individual performance signatures (Widmer, 2003) by modeling performer-specific patterns (Madsen & Widmer, 2006) in combination with string matching representations. Serving the same purpose, other methods have been proposed with the aim of discovering idiosyncratic rules (Widmer, 2003). Yet, due to the complexity of the problem, the datasets available and the sensitivity of the algorithms (Widmer & Tobudic, 2003), it remains a challenge to elucidate what aspects may constrain the interaction between the different expressive features.

Few studies can be found concerning the possible dependencies between score markings and performance expressiveness. In an illustration of the difficulty of the task, Todd argues that *"often there is no direct relationship between dynamic markings in the score and actual performance [...] the expression marks in a score are used only as a rough guide by performers"*(Todd, 1992). In line with these observations, the use performers make of expressiveness at score markings has generally been considered as arbitrary or even within an improvisatory context (Cobussen, Frisk & Weijland, 2010). However, following the definition of expressiveness of Chapter 2, by which expression is understood as the deviation from the norm defined by the performer and/or perceived by the listener, we may expect that certain score annotations and markings may contribute to establishing such norms. That is, by allowing performers to emphasize their intentional expressive gestures at structural points (Palmer, 1989) in which the markings are already notated, and communicate their idiosyncratic approach of those structural points to listeners. Score markings, therefore, serve as a constraint over which the idiosyncratic qualities of performers may manifest more clearly. Accordingly, it is expected that score markings will have an effect on the performers' choices along the interpretation, and therefore, patterns of expressive gestures are to be found within them.

The complexity of modeling these expressive gestures, together with a lack of publicly available (and large enough) datasets, may explain the scarcity of research of performance analysis and modeling towards score markings. Grachten and Widmer (2012) have shown how the combination of dynamic markings with expressive features using linear basis models (*e.g.*, multiple linear regression) may improve the modeling of expressive loudness rendering on a note-to-note level. Kosta, Bandtlow and Chew (2014) found that the use of expressive dynamics across a group of performers is not linear or rule-based. As shown by Kosta et al. (2014), performers have different approaches

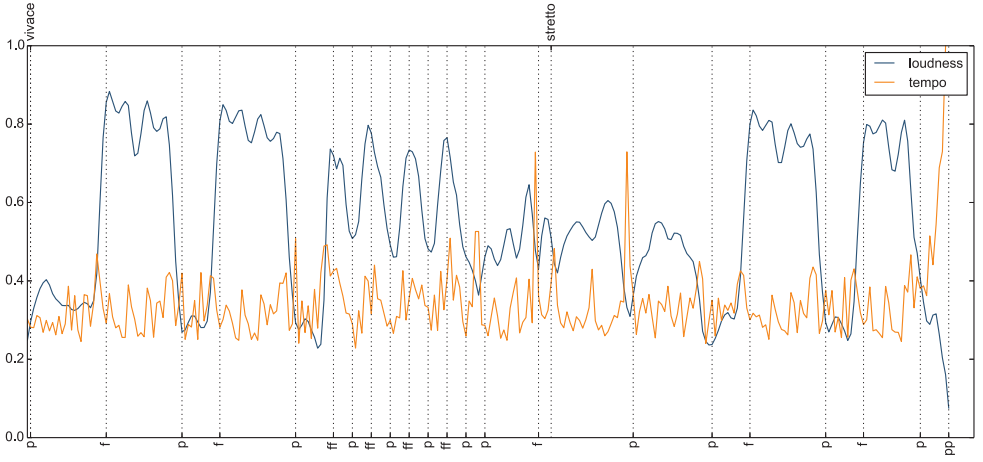


Figure (4.1) Normalized means of tempo and loudness curves for 11 pianists playing Mazurka op. 06-3.

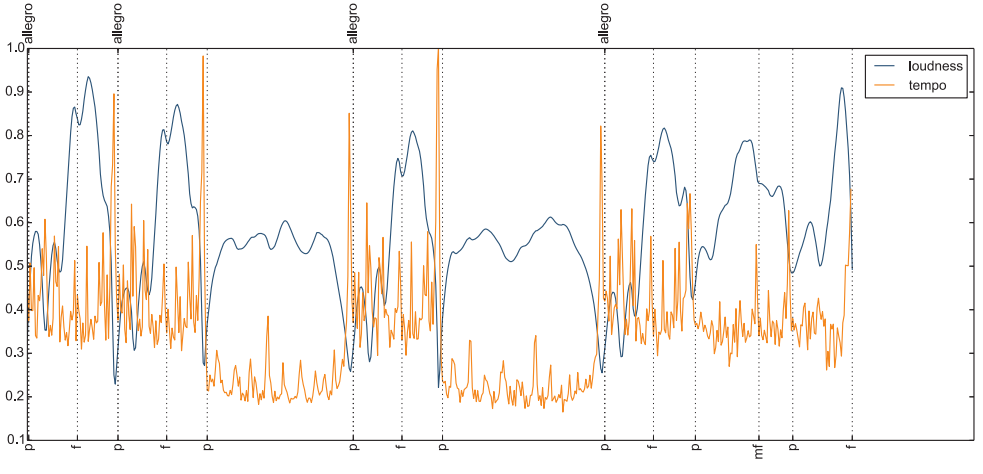


Figure (4.2) Normalized means of tempo and loudness curves for 11 pianists playing Mazurka op. 56-1

to the represented dynamic markings and there is not an ordinal relation between the representation of expressive dynamic markings and the use they make of loudness; a *piano* might be louder than a *mezzo-forte* depending on the approach the performer has to a particular piece. Thus, performers make use of the local and global context of the marking, taking into account the structure of the piece and how it relates to the thematic context. While these studies have focused on finding and learning relations between loudness and dynamic markings they do not consider possible interactions between loudness and tempo at either dynamic or tempo markings.

In the novel approach herewith presented, I study the interactions between performance variables and score-based features while distinguishing among those constraints (possibly) derived from the score and shared among a group of performers, and those constraints idiosyncratic to a performer. By using a machine learning approach I analyse whether there are interactions between timing and loudness at score markings and whether these are better modeled by the idiosyncratic style of each performer playing several pieces or by the shared stylistic (expressive) constraints embedded in a piece when played by different performers.

4.1.1 *Hypotheses and Experiments*

Based on the arguments and observations presented, this study departs from two hypotheses:

- first, I hypothesize that the expressive choices of performers in loudness and tempo may be constrained by the markings and their structural relation of these to the score. It is expected that the performers' expressiveness on loudness and tempo will be constrained by the score. Therefore, the expressiveness of performers in tempo and loudness could be better predicted by our models when knowing which tempo or dynamic markings are written in a particular score.

- second, I hypothesize that the performers' use of expressive tempo and loudness at both tempo and dynamic markings will be better predicted when combining them (as complementary features) than when considering them as isolated features. For instance, tempo is expected to be better predicted at tempo markings when our models contain information about both tempo and loudness from the two preceding bars to the one in which the marking is placed. I therefore examine how tempo may contribute to the prediction of loudness at specific dynamic score markings (*e.g. piano*, *forte*, etc), and how loudness may contribute to the prediction of tempo at specific tempo score markings, (*e.g.*, *lento*, *moderato*, etc).

In the interest of investigating and testing such hypotheses by means of computational analysis, two sets of experiments are presented:

- In the first set of experiments, the goal is to model the prediction of tempo or loudness per Mazurka for each pianist, based on how all other pianists (10) have performed that particular Mazurka.

- In the second set of experiments, the models focus on whether tempo or loudness can be better predicted per Mazurka for each pianist, based on how the same pianist played all other Mazurkas (25) of the dataset.

With these experiments, the aim is to elucidate whether the interactions between score markings, tempo and loudness are better predicted by learning shared stylistic approaches or by the performers' idiosyncratic approaches to the pieces in this dataset. In the next section, I will present the material and methods chosen to test the hypotheses presented.

4.2 MATERIAL AND METHODS

4.2.1 *Dataset*

For the purpose of this study, I examined piano recordings annotations of twenty six Mazurka pieces by Frédéric Chopin (1810-1849) played by eleven pianists. The *Mazurka* is a general term used to refer to a series of Polish folk dances in triple meter. Maria Szymanowska (1789-1831) was probably one of the first composers including Polish popular dances into the art music tradition (Golos, 1960). Chopin, however, is probably the best known western music composer of Mazurkas, due to his extensive repertoire. In particular, Chopin adapted to the piano several of the popular Polish songs and rhythms, composing at least, 59 Mazurkas between 1825 and 1849.

Originally, the rhythm in a Mazurka contains a very regular pattern such as the one illustrated in Figure 4.3 over which different choreographies and melodies can be semi-improvised and danced. Similarly to a *waltz*, a Mazurka (especially in the accompaniment) is typically accentuated as a strong-weak-weak stress pattern.

Figure (4.3) Mazurka rhythm

The prototypical Chopin Mazurka is structured as follows: Introduction, three main parts with corresponding subsidiary sections, and coda (Rink, Spiro & Gold, 2011). The three main parts may be found in different ways depending on the Opus number. In a structural analysis shown by Rink et al. (2011) from the Op. 24-2 we can find:

- *intro*
- first part, with sections: A - B - A'
- *codetta*
- second part, with sections: C - D
- third part, with section: A"
- *coda*

Having a common structure, the Chopin Mazurkas repertoire is a convenient frame to study whether the structure effectively constraints the performers' idiosyncratic expressiveness.

The dataset used in this study is based on the one presented in Kosta, Bandtlow and Chew (2018), which is an extension of the CHARM Project's Mazurka database [1]. This dataset is, to the best of my knowledge, the only one currently available containing such a variety of performances played by so many performers. With the aim of maximizing the amount of performers playing the same pieces, 11 pianists playing (the same) 26 Mazurkas were selected from this dataset. By having such a large dataset of Mazurka recordings allows to investigate whether interactions between features may exist in the score markings.

The Mazurka pieces included in the dataset are shown in table 4.1. The pianists and date of the recordings are shown in table 4.2.

Mazurka Opus	06-1	06-3	07-1	07-2	17-2	17-3	24-1	24-2	24-4
Dynamic Markings	18	22	13	13	6	9	4	12	33
Tempo Markings	6	2	5	11	4	10	2	4	12
ID	1	2	3	4	5	6	7	8	9

Mazurka Opus	30-1	30-2	33-1	33-2	41-2	41-4	50-2	56-1	56-2
Dynamic Markings	8	14	5	16	5	7	14	14	7
Tempo Markings	3	2	3	2	2	3	3	4	3
ID	10	11	12	13	14	15	16	17	18

Mazurka Opus	56-3	59-2	59-3	63-3	67-1	67-4	68-2	68-3	Total
Dynamic Markings	16	8	11	4	18	11	21	8	317
Tempo Markings	3	2	4	2	2	10	5	2	109
ID	19	20	21	22	23	24	25	26	

Table (4.1) Mazurka Opus used for this study and respective number of dynamic and tempo markings collected from each piece.

1 www.Mazurka.org.uk

Pianist	Barbosa	Czerny-Stefanska	Chiu	Smith	Ashkenazy	Rubinstein
Year	1983	1989	1999	1975	1981	1966

Pianist	Fliere	Cortot	Shebanova	Mohovich	Kushner
Year	1977	1951	2002	1999	1989

Table (4.2) Pianist's name and year of the recording

A possible issue with the experiments presented is the impossibility to control whether the performers of this dataset shared similar score editions of the pieces being recorded. This indeed could have a great effect on the results obtained, as it is known that in the different editions available of the Chopin Mazurkas, score publishers have often deliberately included additional score markings (based on the editors musicological knowledge and their interpretation of the pieces). Nonetheless, the research presented in this chapter departs from the assumption that the score markings to be studied here are linked to relevant structural points of the score intended by the composers, interpreted by performers and shared across editions. Therefore, despite some exceptions, it is reasonable to assume that the majority of the score markings presented in the *urtext* (original) edition were included in the different score editions used during the recordings.

Having the pianists and pieces of the dataset defined, the following score-based and performance-based features were collected:

4.2.2 *Score-based features*

The score markings were extracted from the Mazurkas and are based on the edition by Paderewski, Bronarski and Turczynski (2011) as presented in Kosta et al. (2018). The following features, based on those proposed in Kosta et al. (2014), were obtained from the scores listed in Table 4.1 and used as features for the current study:

1. Marking position at which either loudness or tempo is predicted (e.g. ***f***)
2. Previous tempo or loudness (depending on the study) marking position (e.g. ***mf***)
3. Next tempo or loudness (depending on the study) marking position (e.g. ***pp***)
4. Possible additional marking at which either loudness or tempo is predicted
5. Distance in beats to previous marking of the same expressive feature (either tempo or loudness)
6. Distance in beats to next marking of the same expressive feature (either tempo or loudness)

The score-based features 1-4 are represented by markings and, therefore, categorical, for which they are transformed into a binary representation by using one-hot encoding.

Table 4.1 includes a total of 317 dynamic markings and 109 tempo markings. These are:

- dynamic markings:

  pp which occurs 33 times, ***p*** 140, ***mf*** 18, ***f*** 97, and ***ff*** 29 times

- tempo markings

  allegretto which occurs 11 times, ***moderato*** 16, ***lento*** 10, ***fermata*** 16, ***stretto*** 21, ***vivace*** 15, and ***allegro*** 20 times.

4.2.3 *Performance-based features*

Based on the hypothesis presented in 4.1.1, expressive tempo and loudness annotations were collected for each performance and piece of the dataset. As illustrated in Figure 4.3, the triple meter characteristic for the Mazurka rhythm is often spread over two score bars. In the scope of this study, I am interested in modeling how tempo and loudness change at the bar in which the score marking is placed in the context of the performance. The aim is to capture the context defined by the two previous bars to the marking and the possible effect of the marking within the expressive discourse of the performances. For instance, whether a certain gesture in the use of tempo and loudness for a specific marking can be better predicted when including loudness or tempo from the two previous bars. With such purpose, the corresponding values in tempo and loudness from two bars before the score marking are included as features.

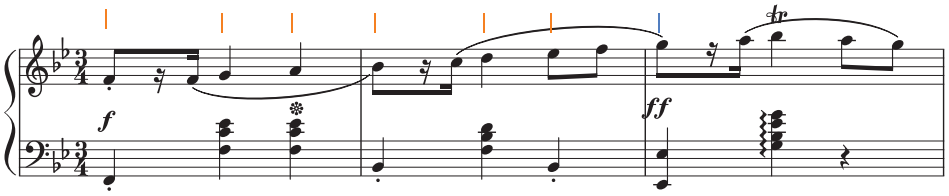


Figure (4.4) Score fragment from the first three bars of Chopin's Mazurka Op. 7 No. 1 in B Major. The beat positions over which loudness or tempo values were extracted are indicated with a vertical line. The orange lines indicate the values (durations or sones) from the previous two bars to the marking. The blue line represents the beat where the duration or sone at expressive marking (in this case *ff*) is found.

The following performance-related features were obtained from the recordings listed in table 4.2:

1. Inter-Beat Interval (IBI)

   The annotations of the Inter-Beat Intervals from the audio recordings were obtained from the dataset presented in Kosta et al. (2018), which describes the onset annotation process by means of the following semi-automatic approach: For each Mazurka piece, indexed by its Opus-Number, one recording is detected as the "reference" recording (in which the onsets are manually inspected and corrected). The beat positions in the rest of the recordings of the same piece were automatically annotated using a pairwise alignment technique in relation to such reference (Kosta et al., 2018).

   The onset detection algorithm is the one presented in Ewert, Müller and Grosche (2009), which uses Dynamic Time Warping (DTW) and incorporates chroma features that facilitate positioning the onsets per chroma (Kosta et al., 2018). The heuristic proposed by Kosta optimizes the choice of the recording used as reference by estimating the minimum match distance between the candidates, computing their Euclidean distance in pairwise manner and, like this, reducing the alignment error. The duration (measured in seconds) between each annotated beat (IBI) is then calculated. The beats of each performance are normalized by dividing by the largest loudness value of each performance sequence.

   The IBI value to be predicted (corresponding to each bar in which markings are placed) is the average of three consecutive beats at the bar in which the marking is placed. The tempo values of the score-beat positions b constitute a sequence $y_n, n \in \mathbb{N}$, where n is the number of score beats in one piece. The tempo mean t at bar y_n is calculated then as:

$$t_{y_n} = \frac{1}{3} \sum_{i=b}^{i=b+2} d_i, \tag{17}$$

   where d is the inter-beat interval in seconds.

   Figures 4.5 and 4.6 show boxplots of the inter beat intervals to be predicted during the experiments for all pieces and performers. Figure 4.5 shows the tempo value ranges for each pianist playing all Mazurkas. As can be observed, Mazurkas 10 (Op. 30-1), 12 (Op. 33-1) and 15 (Op. 41-4) are the ones with largest variance of IBIs, while Mazurkas no. 21 (op.59-3) and 26 (op.68-3) showing limited variance. When inspecting the individual pianists, the variance in the IBIs is slightly narrower in the case of Fliere and Shebanova than in the case of the rest of pianists in the dataset.

2. Beat loudness

   The raw loudness values per beat were also obtained from the dataset presented in Kosta et al. (2018). The loudness corresponding to each beat was extracted

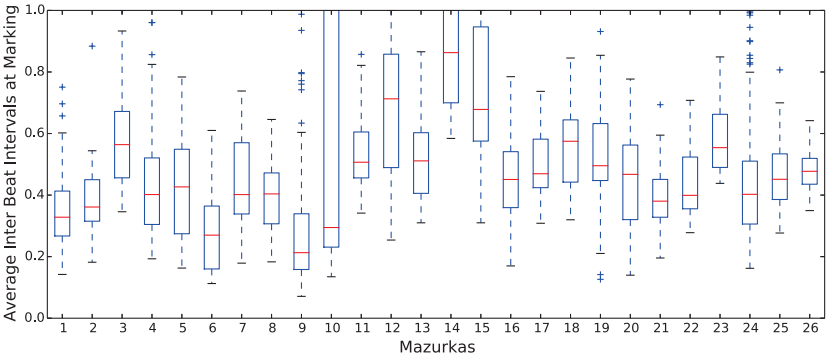


Figure (4.5) Inter-Beat Intervals per piece

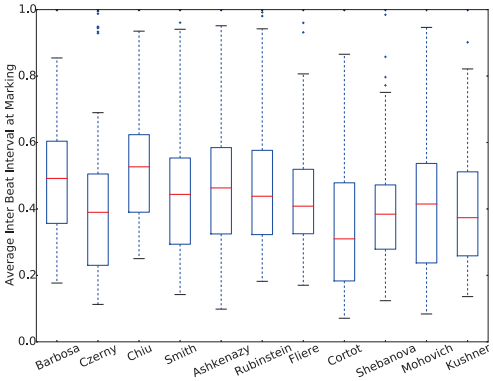


Figure (4.6) Inter-Beat Intervals per performer

using the MA Toolbox (Pampalk, 2004) and is represented in sones (Zwicker & Fastl, 1999). A sone is a linear scale commonly used in psychoacoustics to measure the perceived loudness. In these experiments, the loudness values of each performance are normalized by dividing by the largest loudness value of each sequence.

As with tempo, the loudness values to be predicted and corresponding to each marking are the average of the values found at the beat of the marking and the two consecutive beats. The values of loudness corresponding to the two previous bars represent the loudness at each beat. Considering the dynamic values of the

score-beat positions that constitute a sequence $y_n, n \in \mathbb{N}$ (where n is the number of score beats in one piece), the loudness mean l at bar y_n is calculated as:

$$l_{y_n} = \frac{1}{3} \sum_{i=b}^{i=b+2} s_i, \tag{18}$$

where s is loudness expressed in sones.

The boxplots presented in Figures 4.7 and 4.8 show the ranges in loudness values to be predicted corresponding to the dynamic markings considered. Figure 4.7 shows the loudness ranges for each pianist playing all Mazurkas. Figure 4.8 shows the loudness ranges for all pieces played by all performers. As we can observe, the performances of Mazurkas 1 (Op. 06-1), 4 (Op. 07-2), 6 (Op. 17-3) and 16 (Op. 50-2) contain the largest range of values to be predicted. Instead Mazurkas 11 (Op. 30-2) and 14 (Op. 41-2) are performed with little variability among all performers.

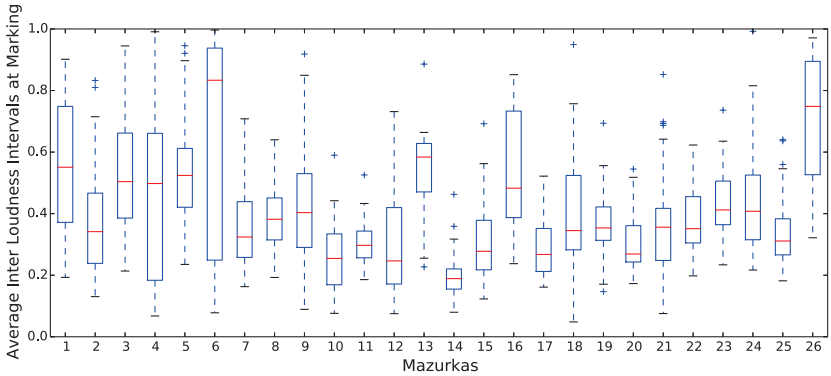


Figure (4.7) Inter-Beat loudness per piece

Table 4.3 lists the different versions of the features set used in all the experiments. Since the prediction task consists on predicting either tempo or loudness at the score markings, the "baseline" set (B) contains only tempo or loudness (depending on the experiment) at markings. The L set contains tempo or loudness values (depending on the experiment) at the markings, as well as the beat loudness values from two bars preceding the markings. The T set contains the tempo or loudness values (depending on the experiment) at the markings, as well as the beat tempo (IBI) values from the two bars preceding the markings.

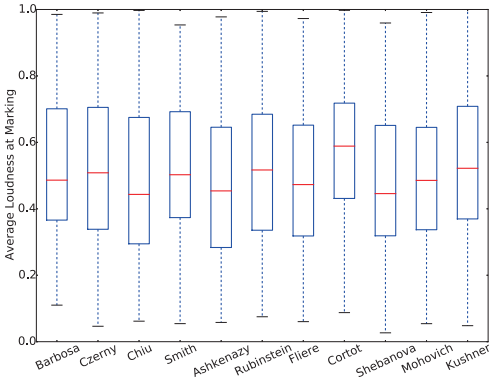


Figure (4.8) Inter-Beat loudness per performer

Abbrev.	Features used
B	Baseline (symbolic score-based features)
L	Baseline + previous two bars Inter-Beat Loudness
T	Baseline + previous two bars Inter-Beat Intervals
A	B + L + T

Table (4.3) Different sets of features used for the experiments

4.2.4 *Models*

The hypothesis aforementioned has been tested by using three alternative machine learning regression models to support and contrast the possible interpretations to be made upon the results obtained. The algorithms chosen were multi-layer perceptron (MLP), random forests (RF) and k-nearest neighbors (k-NN).

4.2.4.1 *k-Nearest Neighbors*

k-Nearest Neighbors is a non-parametric algorithm that can be used for classification or regression. In this dissertation it will be used for regression. Having an input X in the feature space, the label k assigned to a sample X_n is estimated based on the mean of the labels assigned to the nearest neighbors samples. First, it calculates the Euclidean distance between each of the samples labeled. Secondly, it orders the samples in an increasing order. Thirdly, it uses a manually assigned integer number for the k-nearest neighbors or a heuristic to find an optimal amount of k nearest neighbors via cross validation. Finally, it calculates the average of the inverse distance weight among the k nearest neighbors. Having a large number for k is often associated with better capturing the possible outliers in the levels of values to regress over.

4.2.4.2 *Random Forests (RF)*

Random forests are an ensemble learning method obtained from the resulting mean of several individual decision trees used during the training phase. A decision tree is built by selecting at each node the most relevant attribute. The selection of the most relevant attribute is based on the information gain associated with each node of the tree (and corresponding set of instances).

In a random forest, having a great number of different decision trees for the training set, the ensemble learning consists on randomly subsetting at the node level the parameters of each of the decision trees and obtaining the ones that give a better split of the data according to the objective function. This way, the random forest obtains the best possible parameters among all decision trees to represent the best approximation to the training/validation set.

4.2.4.3 *Multi Layer Perceptron (MLP)*

For a detailed description of the multi-layer perceptron the reader is referred to Section 3.3.1 of this dissertation. In Chapter 4, the activation function chosen in the hidden layer is a rectified linear unit (ReLu) function, $f(x) = max(0, x)$. In recent years, ReLu functions have become increasingly popular when working with deep learning architectures containing several stacked layers, as in these architectures, sigmoidal functions have shown to make the training harder. Furthermore, ReLu functions have been shown to speed up the convergence of the gradient descent when compared to sigmoid or tanh functions (Krizhevsky, Sutskever & Hinton, 2012), as well as to improve the results of some neural network acoustic models (Maas, Hannun & Ng, 2013).

4.2.5 *Model selection and evaluation*

To overcome the possible limitations of the dataset size and prevent compromising the generalization power of the models obtained, I used jackknifing (Efron & Gong, 1983), which is applied by splitting each Mazurka in two equal sections. As such, each Mazurka in the dataset is split between training set on one half of the Mazurka, and test and validation sets on the other half. Thus, the test and validation each contain $1/4$ of each Mazurka. The hyperparameters for the methods proposed are selected based on the train and validation sets by using grid search with cross validation. Grid search works by doing an exhaustive search over the specified hyperparameter values for an estimator considering every possible combination of features. Finally, the test set (25 % of the markings for each Mazurka) is reserved for the evaluation of the models. Results are averaged over the two iterations per Mazurka.

MLP		RF	
hidden layer sizes	[50, 100, 200]	n estimators	[5,10,15]
max iterations	[200, 400]	max features	[1,3,10,20]
beta_1	[0.5, 0.9]	min samples leaf	[1,3,10]
		min samples split	[1,3,10]

Table (4.4) Hyperparameters grid used for Multi Layer Perceptron (MLP) and Random Forests (RF)

The models' predictions are evaluated by calculating the mean squared error (MSE) between the true values (those actually played during the performances) and the values predicted by the models. MSE is often preferred in the machine learning literature when evaluating regression, as the mathematical calculations of the gradient are simpler than with, for instance, mean average error (MAE). As explained in Chapter 2, using MSE, large errors have a greater influence in the results than smaller errors.

Finally, in this study, model comparison is done based on a Wilcoxon test (Wilcoxon, 1946), which tests the null hypothesis that two paired samples come from the same distribution. Unlike other (parametric) tests, such as the t-test, the Wilcoxon test (being non-parametric), does not assume normality in the distribution of the populations.

4.3 EXPERIMENTS AND RESULTS

In this Section, I describe the models obtained to study the effect of score markings on tempo and loudness based on a group of performers playing the same piece (Section 4.3.1) or a performer's individual style (Section 4.3.2). It must be noted that the error measure used to estimate the prediction power of each model is based on data normalized between 0 and 1.

4.3.1 *Experiment 1. Score-based models: Predictions of tempo and loudness at score markings*

In this Experiment, I examine the possible interactions between tempo and loudness (at tempo and dynamic markings) shared across performers playing the same piece. For this purpose, the models predict tempo or loudness per pianist and Mazurka having been trained on all other pianists (10) playing the same Mazurka. The aim of this experiment is studying whether there are shared approaches across pianists as a result, possibly, of score-based constraints. In this case the score constraints are defined by the score markings.

4.3.1.1 *Score-based models: tempo prediction at tempo markings*

The results shown in Tables 4.5 and 4.6 indicate that the prediction of tempo at tempo markings (B) are improved in all models when including tempo values from the two

bars prior to the markings (T). In the best performing model (MLP), the Wilcoxon test indicates that the difference between the (B) and (A) models is significant ($p = 0.0153$).

We can also observe that tempo is better predicted (but not significantly) when combining loudness and tempo with baseline features (A) than in the (T) model, which only includes tempo values from the two preceding bars. In both the RF and k-NN models we observe that the predictions of tempo are worse when including loudness features from the previous two bars to the marking.

Models	MLP	RF	K-NN
B : Baseline (score only features)	0.0242	0.0196	0.0230
L : B + previous two bars Inter-Beat Loudness	0.0217	0.0235	0.0223
T : B + previous two bars Inter-Beat Intervals	0.0178	0.0188	0.0179
A : B + L + T	0.0175	0.0198	0.0191
Mean Models	**0.0203**	0.0204	0.0206

Table (4.5) Average MSE for the prediction of tempo per score-based model using different sets of features

	B vs. L	B vs. T	B vs. A	L vs. T	L vs. A	T vs. A
Wilcoxon	0.1513	**0.0153**	0.1068	0.0054	0.0063	0.7702

Table (4.6) Wilcoxon tests between models with different sets of features based on best performing model avg predictions

4.3.1.2 *Score-based models: loudness prediction at dynamic markings*

The results shown in Table 4.7 indicate that loudness at dynamic markings is better predicted in all models when including the loudness values corresponding to the two previous bars to the marking (L), than when including only the baseline features (B). The Wilcoxon test between (L) and (A) for the best performing model (MLP) indicates that this improvement is significant (p=0.0176).

In the MLP model, the combination of tempo and loudness values (A) leads to marginally significant better predictions (p=0.049) than the model including only loudness features (L). However, in the case of k-NN and RF there is no improvement when combining both expressive features. A majority of the models show no improvements when using all features (A) in contrast to either Loudness + baseline features (L). I therefore conclude that, in these experiments, tempo is not contributing to the prediction of loudness at dynamic markings.

Models	MLP	RF	K-NN
B : Baseline (score only features)	0.0179	0.0172	0.0194
L : B + previous two bars Inter-Beat Loudness	0.0126	0.0121	0.0127
T : B + previous two bars Inter-Beat Intervals	0.0179	0.0209	0.0174
A : B + L + T	0.0118	0.0135	0.0143
Mean models	**0.0151**	0.0159	0.0160

Table (4.7) Average MSE for the prediction of loudness per score-based model using different sets of features

	B vs. L	B vs. T	B vs. A	L vs. T	L vs. A	T vs. A
Wilcoxon	**0.0176**	0.0006	0.0619	0.0002	**0.0490**	0.0007

Table (4.8) Wilcoxon tests between models with different sets of features based on best performing model average predictions

4.3.1.3 *Experiment 1 results analysis*

The results obtained in this experiment suggest that both tempo and loudness at score markings are significantly better predicted when including contextual information from the bars preceding the markings. However, I found no contribution of loudness to the prediction of tempo at tempo markings, and no strong evidence of tempo contributing to the prediction of loudness at dynamic markings.

The results shown in Tables 4.7 and 4.5 are averaged over all predictions per pianist and Mazurka. When inspecting the predictions per piece, it was also observed that some pieces show better interactions between tempo and loudness than others. This suggests that the musical material contained within them constrains differently the interactions between these two features. That is, the expressive phrasing across performers (and their idiosyncrasy) when playing the same piece seems to be constrained differently depending on the pieces played due to the differences in the musical discourse. Furthermore, it is also likely that these constrains are due to aesthetically shared cultural approaches.

4.3.2 *Experiment 2. Performer-based models: Predictions of tempo and loudness at score markings*

In this experiment, we study the use of tempo and loudness at tempo and dynamic markings per performer. The models obtained are trained to predict tempo or loudness per piece and performer after having 'learned' from the (same) performer's individual expressiveness when playing all other pieces of the dataset (25). In this way, the model may capture idiosyncratic expressive gestures in the use of tempo and loudness per performer. The aim of this experiment is not only to compare whether the predictions

are better than in Experiment 1, but also to elucidate whether performers are consistent and idiosyncratic in their expressive style across several pieces.

4.3.2.1 *Performer-based models: Tempo prediction at tempo markings*

From the results shown in Table 4.9, we can see that the predictions of tempo with the baseline features (B) are better than when adding expressive tempo features from the two bars preceding the marking (T). In the results of the Random Forests model, we can observe that combining tempo, loudness and baseline features (A) contributes to improve the prediction of tempo (slightly) and to improve the predictions of the baseline. However, in Table 4.10 we can observe that this improvement is not significant.

Models	MLP	RF	K-NN
B : Baseline (score only features)	0.0374	0.0242	0.0218
L : B + previous two bars Inter-Beat Loudness	0.0479	0.0228	0.0240
T : B + previous two bars Inter-Beat Intervals	0.0595	0.0266	0.0375
A : B + L + T	0.0719	0.0213	0.0282
Mean Models	0.0542	**0.0238**	0.0279

Table (4.9) Average MSE for the prediction of tempo per performer-based model using different sets of features

	B vs. L	B vs. T	B vs. A	L vs. T	L vs. A	T vs. A
Wilcoxon	0.5937	0.3739	**0.0076**	0.1823	0.2132	0.1095

Table (4.10) Wilcoxon tests between models with different sets of features based on best performing model average predictions

4.3.2.2 *Performer-based models: loudness prediction at dynamic markings*

Table 4.11 shows that loudness at dynamic markings is better predicted (in all models) when including the loudness values corresponding to the two previous bars to the marking (L), than when including only the baseline features (B). The Wilcoxon test between (L) and (A) for the best performing model (RF) indicates that this improvement is significant (p=0.0044). The results also indicate that, in none of the models, loudness is better predicted when combining tempo and loudness features (A) than when using only loudness features (L).

Models	MLP	RF	K-NN
B : Baseline (score only features)	0.0448	0.0317	0.0530
L : B + previous two bars Inter-Beat Loudness	0.0417	0.0252	0.0362
T : B + previous two bars Inter-Beat Intervals	0.0616	0.0333	0.0495
A : B + L + T	0.0479	0.0270	0.0394
Mean Models	0.0490	**0.0293**	0.0445

Table (4.11) Average MSE for the prediction of loudness per performer-based model using different sets of features

	B vs. L	B vs. T	B vs. A	L vs T	L vs. A	T vs. A
Wilcoxon	**0.0044**	0.2132	0.0076	0.0076	0.0329	0.0099

Table (4.12) Wilcoxon tests between models with different sets of features based on best performing model average predictions

4.3.2.3 *Experiment 2 results analysis*

The results obtained in this experiment show that the predictions of loudness at dynamic markings are significantly better when adding loudness values from the two preceding bars with respect to the baseline. In the case of tempo, the results show that the predictions of tempo at tempo markings do not improve when including tempo values from the two preceding bars in respect to the baseline features.

This might be due to how pianists adapt their use of expressiveness to the structural musical content of the piece being performed, which challenges the recognition and learning of gestural idiosyncratic patterns in the use of loudness and tempo. Furthermore, while tempo shows to interact with loudness in the predictions of tempo, tempo does not show to interact with loudness when predicting loudness. This indicates that pianists emphasize the structural changes indicated by tempo markings with loudness. Yet, performers' changes in loudness at dynamic markings do not seem to be emphasized by changes in tempo.

These results illustrate the complexity of the task as the expressive strategies followed by different performers might be different depending on the musical discourse. This makes it challenging for the models to learn idiosyncratic patterns of expressiveness within the context defined.

4.4 DISCUSSION

In this chapter, I have presented an exploratory study on modeling possible interactions between score markings, tempo and loudness at tempo and dynamic score markings. For this purpose, I designed two experiments in which predictions of tempo and loudness are analyzed based on the effect of combining different sets of features at specific score markings. The first experiment focused on modeling shared uses of tempo and

loudness by training our models on a group of performers playing the same piece of music and predicting it on a performer unknown to the model. The second experiment focused on modeling performer specific uses of the same features shared across performances of different pieces played by the same performer.

In the score-based models (Experiment 1), tempo and loudness showed to be better predicted at markings when including contextual information (expressive performance values from the two bars preceding the marking) for all models. No improvements were found on the prediction of tempo when combining tempo with loudness features preceding the tempo markings. When combining loudness and tempo features preceding the loudness markings, loudness was only better predicted in the MLP model, but no improvements were found in the rest of models (RF and k-NN).

In the performer-based models (Experiment 2), loudness was better predicted in all models when using contextual information from the same feature than when using just score-based features, but none of the models improved the prediction of loudness (L) when combining loudness with tempo (A). In the best performing predictive model, the prediction of tempo was significantly improved when including both loudness and tempo from the two preceding bars (A) in respect to the baseline model (B), but in the other models it was not.

An explanation of why the (T) models lead to better predictions than the (B) models in Experiment 1 (E1), but not in Experiment 2 (E2), is that the tempo markings are 'prepared' differently per piece. That is, tempo markings (in relation to tempo changes) are often expressed in a *subito* (sudden) manner, and thus, the tempo prediction at the bar in which the marking is placed might not be 'prepared' during the performance in the two previous bar to the marking. Since, in Experiment 1, the models are trained on the same piece played by different performers, the models seem to be able to learn the 'preparation' of those markings across performers better. However, in Experiment 2 models are trained across different pieces, and thus, the expressiveness strategies per pianist in the use of tempo might be piece-specific and, probably because of this, they are not well captured within this dataset and models. Furthermore, when inspecting the predictions per performer, each pianist seems to have very different strategies that could be influenced by the piece being performed. This outcome is coherent with the boxplots shown in Section 4.2.3, in which the variance across pianists for tempo (see Figure 4.6) is larger than the variance for loudness (see Figure 4.8). This could explain why, with the dataset herewith analyzed, the models are challenged when learning a gestural pattern on the interactions between expressive loudness and tempo.

The findings herewith presented, support the first hypothesis (see Section 4.1.1); having contextual information preceding the marking improves the prediction of expressive loudness and tempo. In this study, however, no evidence was found supporting the second hypothesis presented, since, in most models, the results showed no interaction between loudness and tempo at tempo or dynamic markings. Accordingly, these results suggest that the interactions between tempo and loudness at score markings depend on the different approaches per performer and piece and, as such, they are not reflected on the shared stylistic approaches across the Mazurkas herewith studied.

The results obtained show as well that the prediction error is smaller on score-based models (trained on the same piece played by different performers) than on the performer-based models (trained on the same performer playing different pieces). Therefore, the score constraints derived from harmony, melody or rhythm around tempo and dynamic markings seem to have an effect on the expressive choices that are shared across performers and, probably, on those made per piece. The results obtained add to previous evidence on expressive constraints based on the score (Repp, 1990) and suggest that, for this dataset and models, modeling score-specific dependencies might be an easier task than modeling performer-specific dependencies. That is, the predictions of the performer-based models are worse, probably because the diversity of the musical content being performed; the musical pieces included within this corpus are quite different from each other besides all of them being Mazurkas. This diversity has a greater influence on the use of dynamics and tempo in relation to these expressive markings as the relation between the expressive variables and markings might vary across pieces more than across performers. Moreover, this suggests that, if existent, the idiosyncratic signature per performer in these experiments is not well captured across these models since they are making use of their artistic freedom, using very different expressive gestures in tempo and loudness depending on the performance context. This, rather than being a shortcoming of the predictive models used and dataset used, exemplifies the complexity of the task and experiments carried out.

Among the limitations of this study, we should take into account both the size of the dataset and the limited amount of markings available. The predictive error obtained when the response variable was loudness is lower than when it was tempo. This suggests that, within this dataset, tempo is a more complex variable to be modeled and predicted. Yet, we must note that, in this dataset, the amount of dynamic markings present in the scores is much larger than those of tempo markings (see Figures 4.1 and 4.2). This is intrinsic to the music material analyzed, as in most scores from the Romantic period, dynamic markings are more abundant than tempo markings. Thus, in this regard, claiming that loudness is easier predicted than tempo, might be a biased interpretation. The results obtained could be confirmed by examining a larger dataset of performances and considering other composers and periods of music.

Finally, we must as well reckon that the performance features proposed for the machine learning models, may not capture the whole temporal length and trajectory of the expressive gestures. While I proposed modeling the relation between the six previous beats and the average of tempo or loudness at the bar in which the masking is placed, it could be expected that the characteristic expressive gestures of certain pianists on specific Mazurkas occurs slightly before or after the bar at which the markings are placed. As explained in Section 4.2, the choice for these features is based on the fact that the Mazurka rhythms in our dataset are best characterized over two measures (see figure 4.3). That is, we may assume that this is an optimal contextual range in relation to score markings. Even when the markings would be shared among different score editions used during the recordings, the interactions may depend as well on the individual artistic approach (and limitations) of the performer linked to the musical

content. In fact, while the score markings are being used as a guideline, it is the duty of performers to bring coherence to the musical discourse, establishing the *"negotiable relation between the score and the performance event"* (Gould & Keaton, 2000). This coherence, however, might be influenced by other musical features than those herewith contemplated, such as, for instance, the shape and size of phrasing in the musical discourse in relation to the rhythmic content or metrical structure.

In furtherance of better capturing the temporal relation between tempo and loudness gestures, as well as the possible structural dependencies implicit in them, in the following chapter, I will present a study on the idiosyncrasy and shared approaches to the same dataset using sequential recurrent neural networks.

5

THE ROLE OF RHYTHM AND METER AS EXPRESSIVE CONSTRAINTS IN SHARED AND INDIVIDUAL USES OF TEMPO AND LOUDNESS

5.1 INTRODUCTION

In many music cultures and styles, music professionals and aficionados are able to differentiate between performers when these are interpreting the same piece or similar styles. All along history, we can find diverse sources in varied contexts exalting the dexterousness and sensitivity of a performer through their interpretations, often being compared to others. The ability to describe and distinguish between performances and performers is possible as a consequence of several representational and control processes being shared between listeners and performers (Sloboda, 2000); so that music can be communicated between both of them. Among the factors that contribute to determine the performer's individual expressiveness and how it may be perceived by a listener we shall include: the performer's mental representation of the piece to be performed, their technical constraints (or control processes) defined by the piece, and their approaches to the cultural context of the piece being played.

With the goal of understanding whether there is a link between the organizational aspects of a piece of music and its mental representation, literature often distinguishes between two main levels of structure (Jackendoff & Lerdahl, 2006). In the context of performance, the macro-level structure refers to the piece form and includes expressive deviations in tempo, rhythm, large scale dynamics, melodic contour, and harmonic relationships. The other main level of structure is the micro-level structure, which includes instead note-level (or short groups of notes) deviations in timing, pitch, loudness, timbre, or articulation. The micro-structure expressive deviations relate to the prosody and (note-level) errors, while the macro-structure refers to the use of phrasing (Sloboda, 2000). The modulation on different expressive features often extends across both macro- and micro-level structures (Clarke, 2002). This is why, both micro and macro structural levels can be considered as entangled and hierarchically dependent.

A performer's mental representation of the macro structural level on the music might be reflected by their stability on the use of expressiveness above the micro-structure beat and note-levels; while the most characteristic idiosyncratic approaches might be reflected on the expressive deviations exercised across (often) smaller units of expression. According to Timmers and Honing (2002), whether the expressive deviations are large or small is related to the different structural levels. Having, for instance, small variations in timing responds to local timing gestures, while note lengthening variations often responds to larger scale trends and phrases. Thus, the expressive features used by performers respond to both long and short time scales and structures, and this should ideally be captured when modeling individual expressive performances.

The study presented in Chapter 4 showed how tempo and loudness at score markings are better predicted when including as features the two preceding bars to the markings. It also showed how, for the features and models presented, score markings had, in most predictive models, no effect on the possible interactions between tempo and loudness expressive gestures. Nonetheless, these findings should not be understood as the general (score) norm; rather, they should be understood as a specific example in which the music style constraints defined by the score markings are not related to loudness and tempo interactions or not present within the scope of the experiments described.

This chapter instead focuses on modeling performers' expressiveness in tempo and loudness using sequential models along the whole pieces as well as studying possible macro and micro structural dependencies (in this case on the beat-level) derived from the pieces performed. In particular, I study whether the interactions between loudness and tempo are defined by the influence of metrical beat position or by the rhythm representation included in each beat. Moreover, I inspect whether such constraints may be performer-based or shared across performers.

The structural organization of a piece by a performer may depend on meter and grouping as the primary syntactic elements in music (Lerdahl & Jackendoff, 1983).

Meter, constitutes the grid by which, in combination with our innate ability to perceive beats (Honing, 2012), rhythms can be grouped and recognized as patterns. Meter thus is defined on basis of the hierarchical distribution of its constituent beats based on the periodic alternation between strong and weak accents (Palmer, 1997). As suggested by Clarke (1985), expressive timing may be affected by the meter framework (Palmer, 1997). Furthermore, the differences of meter across music styles and cultures, suggest that our perception of meter could be culturally biased (Cross, 2009). Recent implementations of probabilistic models based on symbolic (score based) representations of music have been developed in order to model such enculturation processes of meter perception (van der Weij, Pearce & Honing, 2017).

Meter and rhythmic grouping accents (in addition to melody and serial harmony), influence the planning of performance expressiveness in relation to the short and long-term structure of a piece (Drake & Palmer, 1993). Yet, the use of expressiveness by performers can also affect how meter is perceived by the listener (Sloboda, 1985). As shown by Gabrielsson (1973), listeners could effectively group timing patterns depending on how performers use meter, accent patterns and ratio durations in relation to tempo or forward movement. In the same study, Gabrielsson showed how the expertise of performers influences the communication of meter; in this case, defining communication as the listeners' ability to recognize the performed meter effectively.

In another study, Gabrielsson (1974) showed interactions in the use of amplitude and note durations by performers when these would repeat certain rhythmic patterns having a metronomic tempo cue as a guideline. In particular, it was observed that the first note of each bar would be louder and longer. Furthermore, in a similar line of research, Sloboda (1983) showed how performers tend to accentuate meter by the use

of duration, loudness, and legato when being presented with scores in which the same music stimuli would be shifted one note and thus the metrical position would fall in different places depending on the stimuli.

The categorization of rhythm has a relevant role in the expressiveness of timing, as it can constrain the performers' expressiveness to effectively communicate the rhythm representation to the listeners (Desain & Honing, 2003). For instance, an isochronous rhythm pattern will allow performers to be freer in their expressive choices than a non-isochronous rhythm pattern (Honing, 2006a).

The literature presented suggests that meter is a relevant constraint in the idiosyncratic expressiveness of performers and on how it is perceived by listeners. However, to the best of my knowledge, no research has been done on how idiosyncrasy may be constrained by interactions between tempo and loudness in relation to meter or rhythm.

In this chapter, I study shared and individual expressive constraints of tempo and loudness in performers based on different combinations of performance and score-based features. Accounting for the temporal nature of music and the structural (and hierarchical) dependencies between features, in this chapter I will propose using Long Short-Term Memory networks as a specific sequential model.

The rest of this chapter is organized as follows: Section 5.2 exposes the motivation to use sequential models instead of static ones. Sections 5.3 and 5.4 present recurrent neural networks and Long Short-Term Memory networks as the chosen machine learning approach used in this study. Sections 5.5 till 5.6 describe the experiments to be carried out, the dataset and features used, and the model architecture chosen. Sections 5.8 and 5.9 present the experiments' results and their analysis. A discussion of the results obtained and future work to be done is presented in Section 5.10.

5.2 USING SEQUENTIAL MODELS TO ACCOUNT FOR STRUCTURE IN EXPRESSIVENESS

A key aspect when choosing or developing a computational model is reckoning whether the implicit assumptions of such model correspond to the nature of the data and research question addressed. In chapter 4, we analyzed and modeled performances using several machine learning algorithms in which the data points are assumed (by the methods used) to be independent and identically distributed (i.i.d.). Such an assumption implies that each of the data points and the random variables representing them are independent of each other. That is, what happens at event t (in time) is independent of what happened at $t-1$. These models are also commonly referred to as 'static' models.

Depending on the research questions, characteristics of the data modeled, and computational resources, static models may be an appropriate choice. Additionally to the experiments presented in Chapter 4, several other studies using i.i.d. methods can be found in the modeling of performers characterization. For instance, Ramirez, Maestre and Serra (2010) presented a signal processing workflow and a set of audio features to

classify performers based on their expressiveness. In order to evaluate the features proposed in the same study, they tested their hypothesis on 10 alternative static machine learning methods. Molina-Solana, Lluís Arcos and Gomez (2010) followed a similar methodology with the goal of tagging violin performers. Another successful approach based on static methods is the PLCG algorithm developed by Widmer (2003), which is used to discover performance rules. Devaney (2016) investigated the relation between the classification of a few expressive features (such as those derived from timing, loudness, timbre, and pitch) and the perception of inter-performer and intra-performer features by using support vector machines. Moreover, based on a varied set of static machine learning models, Serrà et al. (2013) showed how several guitar pieces can be automatically identified within a group of performers; even when providing very limited information per performance (up to one note deviations). Despite the advantages of the methods mentioned, as a consequence of their static nature, these models are often too limited when aiming to uncover a parallelism between them and the production or perception of music, which is in its intrinsic nature sequential.

As it has been argued in previous chapters, music performance expressiveness often responds to constraints from both the piece structure as well as the idiosyncratic gestures of the performer (it is possibly also biased by a number of cultural stylistic agreements and biological constraints). Performers convey their expression in a sequential and structured manner that possibly responds to hierarchical constructs as well as to score constraints (when such score is available). For example, a performer may have a characteristic expressive gesture in the use of loudness that correspond to the structure defined by the score. For example, a *crescendo* extended over several notes or bars. Furthermore, the expressiveness of performers can also be defined on multilevel hierarchical expressive gestures (Desain & Honing, 1993). For instance, within the crescendo, there could be a phrasing pattern that is extended over several bars. In such example, performers may increase the loudness not only along that phrasing but also by accentuating with dynamics the first beat of each bar or by having a particular deviation to a specific motif of a few notes (repeated) along the piece being played. Taking into account possible structural and sequential dependencies is essential to model music performances. Such expressive patterns can indeed be established in both written and orally transmitted music as well as in (structured) improvised music. This is because the expressive deviations norm can be defined, learned and recognized during the performance itself. Yet, it is expected that such expressive deviations will be more salient and clearly recognizable when the piece is known by both the performer and the listener.

The capability of sequential models to learn and generate different levels of structure in music is a valuable property in our case of study since the performance expressive gestures are embedded in a temporal flux of events which can also affect our expectations over the music listened to (Bailes, Dean & Pearce, 2013). Furthermore, this is a challenging task for predictive models, as performance gestures are often shared over multiple scales obeying to choices in the timing micro-structure and phrasing macro-

structure, and they are often not only limited to a strict hierarchy (Desain & Honing, 1993).

Such structural hierarchies are also often influenced by a shared understanding of the score structure (Rink et al., 2011). For instance, in a school of pianists or any music group cultural setting, performers may play with a more similar approach to expressiveness than those who are not familiar to such a cultural setting. For the same reason, listeners who have been exposed to such cultural setting will have more biased expectations to such expressive style than those who are not acquainted with that particular expressive style.

In addition, it has been shown that listeners have the ability to perceive temporal regularities in a categorical domain despite the deviations exercised in the expressive performances (Desain & Honing, 2002; Large & Palmer, 2002). As such, the exposure of a performer (being as well a listener) to previous renditions of the piece, the cultural setting, and the music aesthetic trends, will influence and bias the listeners' expectations on the different structural levels within a performance (Brattico & Pearce, 2013). This is a key element in the encoding and decoding process of music communication (Kendall & Carterette, 1990) as well as in the ability to recognize characteristic idiosyncratic deviations and gestures in specific performers (Timmers & Honing, 2002).

While some i.i.d. models can learn a hierarchical representation of independent events, these methods can not account for a sequential order of events. Such constraints challenge the modeling of expressive performances as, ideally, they should be able to attend different macro and micro-structural expressiveness while respecting the sequential nature of music.

Within the field of performance modeling, several studies have used different types of sequential methods to model performers' expressive gestures. Widmer, Flossmann and Grachten (2009) and Raphael (2010) propose using a bayesian network to model timing. Vera and Chew (2014) use conditional Gaussians in combination with clustering to address expressive performance stylistic timing. Grindlay and Helmbold (2006) propose a hierarchical hidden Markov model to predict timing and loudness in specific pianists. Cemgil and Kappen (2003) use Kalman filters for tempo tracking and rhythm quantization. Linking predictive modeling of expressive performances and cognition by using sequential models, Desain and Honing (1992) propose recurrent attractor networks to model timing perception and quantize rhythm. In the same line of research, but using instead neural oscillatory networks, Eck and Schmidhuber (2002) models meter perception and Large and Palmer (2002) propose a perceptual model of rhythm entrainment.

In the following sections, I will introduce the properties of recurrent neural networks (RNN), as the kind of sequential model used in this chapter. Focusing on a type of RNN, namely Long Short-Term Memory (LSTM) networks, I will illustrate their application to modeling and analyzing shared and individual constraints in performance. In addition, I will study possible dependencies between different performance and score rhythm related features. Finally, I will discuss the results and future directions of research following on the approach herewith presented.

5.3 RECURRENT NEURAL NETWORKS AND TEMPORAL PATTERNS

Recurrent Neural Networks (RNNs) are a type of connectionist model in which the information is passed across sequence steps and processed at each of those steps. This processing, in theory, allows for modeling possible multiple level dependencies between events and constituents in the data sequence presented. Consequently, they are suitable to model the inherent processing and generation of multi-level structures in music. For example, we can model the expressive approaches by a certain performer when playing a Mazurka and, for instance, learn how this performer consistently lengthens the first beat of each bar within a particular section that is repeated through the piece. That is, RNNs allow for attending different structural levels of expression within the piece.

These types of neural networks are named "recurrent" because the edges that connect the artificial neurons (or units, or nodes) across adjacent time steps recur over themselves through time. In a RNN, the output $\hat{y}_{(t)}$ at time t will be influenced both by its inputs $x(t)$ and the state of its hidden layer $h_{(t-1)}$ from the previous time step. A RNN can be thought of as a layered net (or unfolded network) in which weights are being reused across time steps. Both the input (X) and output (Y) is a sequence of vectors containing real values having a length T. The dimensions at the input X will vary depending on the features used (at the input). For instance, for one feature, it will be $X_{n\times 1}$, in which n is the number of time steps in the sequence modeled. Y, in our case of study, will always be $Y_{n\times 1}$.

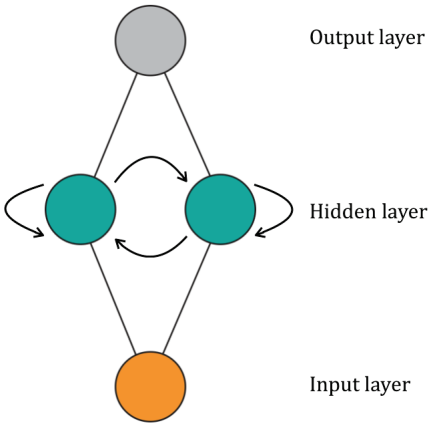


Figure (5.1) Folded RNN, adapted from Lipton et al. (2015). Note that in this folded RNN, the hidden layer arrow does not represent bi-directional connections, but a unidirectional (left to right) connection between the hidden layers. See Figure 5.2 for an unfolded version of the same figure

Formally, (the output of) a basic RNN can be expressed as:

$$h_{(t)} = \phi(W_{xh}x_{(t)} + W_{hh}h_{(t-1)} + b_h) \tag{19}$$

Where $h_{(t)}$ represents the hidden node values at time t, W_{xh} represents the matrix of weights connecting input and hidden nodes, W_{hh} represents the connections between hidden nodes and b represents the bias parameters. ϕ represents the hidden layer (or inner) activation function, which in the experiments to be presented will consist of a hyperbolic tangent *tanh* function, being of the standard choice.

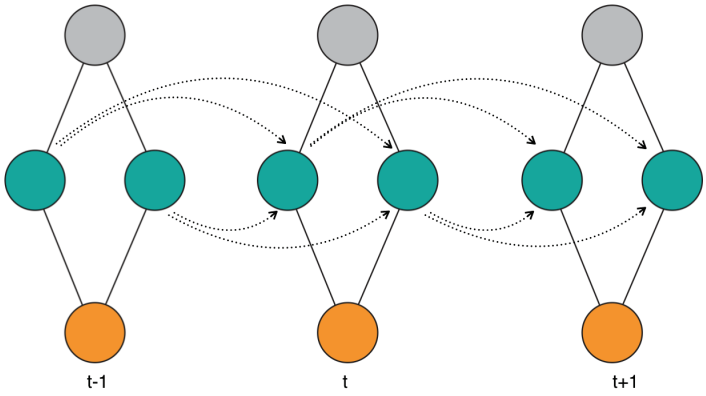


Figure (5.2) Unfolded Recurrent Neural Network across three consecutive time (t) events, adapted from Lipton et al. (2015)

For regression tasks as the ones carried out along this chapter, the activation function used in the output layer is linear, as we do not need any sort of mapping to a discrete representation between the linear transformation given by the hidden nodes and the actual representation of these. In a regression task, therefore, $\hat{y}_{(t)}$ is obtained from:

$$\hat{y}_{(t)} = W_{yh}h_{(t)} + b_y \tag{20}$$

As in feed-forward neural networks, in a RNN, the weights are updated by using Backpropagation (estimating the gradient from the loss function with respect to the weights) but using an extended version of the algorithm called Backpropagation Through Time (BPTT) (Werbos, 1990). In BPTT, the sum of the gradients for all layers (having one layer per time step) is calculated with respect to the error relative to each weight in time. Thus, weights are shared across time steps.

We must note that, in a RNN, each of the time steps represents a "deeper" layer, "since their hidden state is a function of all previously hidden states" (Graves, Mohamed & Hinton, 2013). Thus, a RNN can be considered inherently deep. In the machine

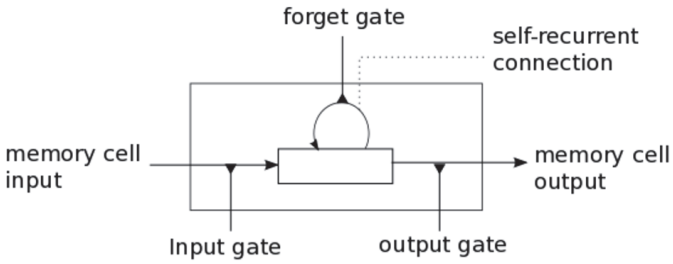


Figure (5.3) LSTM cell As illustrated on `http://deeplearning.net/tutorial/lstm.html` Accessed on 11-11-2017

learning jargon, when referring to deep recurrent neural networks, the depth referred to is most often "in space", rather than "in time". In RNNs, this kind of depth is built by stacking several recurrent hidden layers (in the same manner than feed-forward layers are stacked in MLPs) through which different "representations" of the data learned are passed. Such deep stacks may allow for having different granularities in the time steps being transferred as for capturing a (possible) hierarchical sequential abstraction of the input data in relation to the target output. In the model architectures used within this chapter, the Long Short-Term Memory networks (LSTMs) are not stacked.

Originally, BPTT presented several difficulties for training due to the vanishing and exploding gradient problems (Bengio, Simard & Frasconi, 1994; Hochreiter & Schmidhuber, 1997). This is a phenomenon by which the gradient (or derivative of the error) to be passed across time-steps may decrease or increase exponentially as a function of the number of time-steps (Pascanu, Mikolov & Bengio, 2013). This was a limitation when dealing with long-term dependencies and the representation of structure in long sequences. As shown by Hochreiter and Schmidhuber (1997), given a sufficiently long-range dependency, RNNs would perform as well as chance. In music modeling, problems derived from the vanishing gradient were shown in the generative model for music composition proposed by Mozer (2007). Thus, initially, the RNN combined with BPTT had several limitations when applied to long term structure dependencies.

5.4 LONG SHORT-TERM MEMORY NETWORKS

In order to address the vanishing (or exploding) gradient problem, Hochreiter and Schmidhuber (1997) proposed modifying the RNN architecture by replacing the hidden units with LSTM units. LSTM units consist of a memory cell which stores information along arbitrary time intervals by means of an input gate, an output gate and a forget gate. These gates compute the activation of a weighted sum and allow the LSTM to preserve the error flow through time (Eck & Schmidhuber, 2002). As such, the memory cell permits the possibility to model both short and *longer*-term dependencies than RNNs (Graves et al., 2013).

In the following lines, I include the main components that a LSTM memory cell must have, as explained by Lipton et al. (2015) and Hochreiter and Schmidhuber (1997). An illustration of a LSTM cell is shown in Figure 5.3.

- **Input node (memory cell input),** g, which collects the input from the network at time step t as well as the output from the hidden layer from $t-1$, both are combined by a sum weighted function and passed through a tanh activation function.

$$g_{(t)} = \tanh(W_{gx}x_{(t)} + W_{gh}h_{(t-1)} + b_g)$$

- **Input gate,** i, as in the input node, it collects the input from the network at time step t and from the output from the hidden layer from $t-1$ but, in this case, multiplies it by a sigmoid $[0, 1]$, which allows to control for the amount of information updated at each state.

$$i_{(t)} = \sigma(W_{ix}x_{(t)} + W_{ih}h_{(t-1)} + b_i)$$

- **Self recurrent connection,** s, it is a self-connected recurrent edge with a fixed unit weight given by a linear activation function that carries the internal state of the cell. As it is connected recurrently through all time steps in the sequence it serves as a sort of *tunnel* through which the error can flow across all time steps (with constant weight) and prevent, like this, the vanishing gradient problem.

$$s_{(t)} = g_{(t)} \odot i_{(t)} + s_{(t-1)}$$

- **Forget gate,** f, this function, added by Gers, Schmidhuber and Cummins (2000) and not included in the original paper by Hochreiter and Schmidhuber (1997), is often used when we want the network to forget the information contained in the internal state with a certain decay.

$$f_{(t)} = \sigma(W_{fx}x_{(t)} + W_{fh}h_{(t-1)} + b_f)$$

  The update step of the internal state is done by point-wise multiplication ($\odot$) between the previous internal state and the rest of the components shown.

$$s_{(t)} = g_{(t)} \odot i_{(t)} + f_{(t)} \odot s_{(t-1)}$$

- **Output gate,** o, decides what is going to be in the output from the cell by running it through a sigmoid layer, which decides what parts of the cell are going to the output to afterwards pass it to a hyperbolic tangent activation function (tanh), to scale the output between -1 and 1.

$$o_{(t)} = \sigma(W_{ox}x_{(t)} + W_{oh}h_{(t-1)} + b_o)$$
$$h_{(t)} = \tanh(s_{(t)}) \odot o_{(t)}$$

In a LSTM, the equivalence to a hidden unit input in a simple (vanilla) RNN will be shared over the $g_{(t)}$, $i_{(t)}$, $s_{(t)}$ and $f_{(t)}$ inputs of the LSTM memory cell. Depending

on the network architecture, the memory cell implemented may have slightly different components or activation functions. The ones herewith presented are the most commonly found and included within the LSTM implementation used for the experiments presented in this chapter.

Since their original development, different types of RNNs and LSTMs have been applied to several music research topics such as chord generation (Eck & Schmidhuber, 2002), expectations-based music analysis (Cox, 2010), meter classification (Lambert, Weyde & Armstrong, 2014) or melody generation and prediction (Cherla, Tran, Garcez & Weyde, 2015). LSTMs have been shown to be successful in several predictive tasks in which the data (to be modeled) is sequential (Karpathy, Johnson & Fei-Fei, 2015). In the case of performance idiosyncrasy, LSTM-based models can attend to different structural and, if existing, hierarchical dependencies in the expressiveness exercised along a performance. This is because the LSTM parameters (will) contain information about short and long-range dependencies related to different expressive performance patterns.

Another type of RNNs commonly found in the literature, are Bidirectional RNNs (Schuster & Paliwal, 1997), which, in comparison to unidirectional RNNs, are trained in both positive (left to right) and negative (right to left) directions; having information both from past as from future events.

An impediment for using Bidirectional RNNs in an on-line setup of undefined length, is that they require a fixed endpoint in both the future and in the past (Lipton et al., 2015). As such, they are not appropriate for the study here presented, since the nature of music listening and music making occurs in a unidirectional stream manner. We can't process auditory incoming information from the future, but merely have an estimation on future events based on our exposure and expectations of the auditory stream being listened to. In an offline setup, however, several state-of-the-art models in tempo estimation methods, rhythm (Böck, Krebs & Schedl, 2012) or timing related tasks have made use of Bi-directional RNNs and are often shown to outperform Unidirectional LSTMs. Since the ultimate motivation to carry this study is to better understand the possible underlying mechanisms of expressive music performance or music listening, adopting a bidirectional model was discarded for our experiment purpose. [1]

Finally, another sequential model that has been very much used in speech and music modeling are Hidden Markov Models (HMM) (Rabiner & Juang, 1986). While explaining how HMMs work is out of the scope of this work, the main difference between a vanilla RNN and HMMs is that HMMs depend on discrete states, while vanilla RNNs use real-valued vectors, which allows for a more flexible modeling of the structural dependencies. In this regard, a common argument on behalf of using RNNs instead of Markovian approaches it is that the characteristics of the latter make dealing with long-term structure very impractical (Karpathy et al., 2015). This is because, for an

1 There are however arguments against the non-coherence of a parallelism in the way we process information and how BPTT works even in unidirectional LSTMs (Marblestone, Wayne, Kording & Scholte, 2016).

HMM, "the transition table capturing the probability of moving between any two time-adjacent states is of size $|S|^2$" (Lipton et al., 2015). Furthermore, in order to increase the context of the Markov model we need to create new states by calculating the cross product of the possible states at each time in the context windowed (Graves, Wayne & Danihelka, 2014). Other less compromised solutions than HMM such as Variable Length Markov Models (VMM) or Markov Constraints (Pachet & Roy, 2011) have been applied to music composition recently with convincing results as generation systems, but, to the best of my knowledge, these have not been compared so far with the performance of LSTMs. RNNs are a much more flexible and powerful approach to modeling sequential dependencies than Markov Models since they can represent many more states (become more expressive) without increasing their complexity as much as HMMs.

In the rest of this chapter, we will see how LSTMs can be applied to study the relation between expressive performance tempo and loudness, and meter and rhythm, in the characterization of performers individuality.

5.5 EXPERIMENT DESCRIPTION AND HYPOTHESIS

In this section, I will present two experiments which have as a common basis the study of performer idiosyncrasy, by predicting tempo or loudness per piece and performer:

5.5.1 *Experiment description*

- The first experiment considers the study of expressiveness consistency on individual performers with models inferred per performer based on their use of expressiveness when playing other pieces.
- The second experiment focuses on predicting tempo or loudness per piece and performer based on how all other performers (10) have played the same piece being predicted. The models obtained in this experiment mainly reflect structural constraints derived from the score.

In order to better understand whether there are interactions between tempo and loudness in the use of expressiveness, and if these may be constrained by meter and rhythm, several LSTM models predicting tempo or loudness were inferred based on different combinations of features.

The following combinations of features are considered:

(a) Tempo predictions:

  a) based on tempo (t_b_t)
  b) based on tempo and loudness (t_b_t_l)
  c) based on tempo and rhythm (t_b_t_rh)
  d) based on tempo, loudness, and rhythm (t_b_t_l_rh)

e) based on tempo and meter (t_b_t_m)
f) based on tempo, loudness, and meter (t_b_t_l_m)

(b) Loudness predictions:

a) based on loudness (l_b_l)
b) based on loudness and tempo (l_b_l_t)
c) based on loudness and rhythm (l_b_l_rh)
d) based on loudness, tempo, and rhythm (l_b_t_l_rh)
e) based on loudness and meter (l_b_l_m)
f) based on loudness, tempo, and meter (l_b_t_l_m)

5.5.2 *Hypotheses*

One of the main goals of the study presented herewith is to elucidate whether possible interactions in the use of tempo and loudness may be found as a consequence of performer-based constraints or as a consequence of the constraints defined by the score performed. In doing so, we present an approach to isolate, model and study idiosyncratic expressiveness in the use of tempo and loudness. In order to account for long and short structural dependencies as well as for the sequential character of expressive gestures (and possible hierarchical relations), we propose using Long Short-Term Memory Networks to predicting tempo and loudness.

Another goal of the research presented in this chapter is studying the role of meter and rhythm in the predictive models described. Based on the evidence found in Drake and Palmer (1993), it is expected that performers use accent structures in a systematic way and this may depend on the constraints derived from the rhythm or meter grouping present on the music (scores) performed.

We, therefore, hypothesize that the idiosyncrasy per performer in the use of tempo and loudness may be conditioned by the influence of meter or rhythm on the score structure. Consequently, constraints derived from the meter or melodic rhythm as represented on the score are expected to be reflected on the interactions between tempo and loudness, as used by performers to emphasize structural boundaries.

Based on the findings from Chapter 4, we expect as well that the use performers make of loudness and tempo will be more constrained by the score structure than by their individual style. This would indicate that the expressive approaches of performers may be piece-based. In the same line of argumentation, the possible idiosyncrasies between tempo and loudness will be more evident in the score based models, as the structural dependencies will be reduced to the piece being studied and the models will contain less variance in the performers approaches to interactions at structural points.

5.6 DATASET AND FEATURE PREPARATION

The dataset used for this study is the same one described in Chapter 4, which is obtained from the dataset presented in Kosta et al. (2018). This is a collection of tempo and loudness annotated on recordings of Chopin piano (polyphonic) mazurkas by several professional pianists. To my knowledge, there is currently no other dataset available with such an amount and variety of professional performances of the same pieces. Moreover, the mazurkas dataset is suitable to study possible expressive phrasing dependencies in the macro-structure, which in the mazurkas is often demarcated by the piece form as described in Section 4.2.1, or subphrases extending over eight bars, but also, to study in the micro-structure, within shorter expressive gestures such as the characteristic mazurka rhythm pattern (see Figure 4.3).

In contrast to Chapter 4, in this study, I discarded the mazurka 17 opus 3, due to inconsistencies and formatting issues in the Music-XML score available for the experiment. Thus, the dataset contains 25 mazurkas (instead of 26) played by the same 11 performers as in Chapter 4.

In the following lines, I will describe the different types of features used in the models obtained: audio-derived features and score-based features.

5.6.1 *Performance features*

The performance features extracted and annotated from the audio recordings are the same as the ones used in Chapter 4. These features represent the auditory streaming input of tempo (at the beat-level) and the corresponding loudness values from each performance.

- **Expressive tempo:** for which we use the Inter-Beat Intervals measured in seconds. The onsets for each beat have been previously annotated with a semi-automatic approach with the algorithm proposed by Ewert et al. (2009). This alignment algorithm annotates each onset based on a reference annotation from another recording and using a heuristic to reduce the error on the possible mis-annotated onsets when matching the alignment between the recording being annotated and the reference one.
- **Expressive loudness:** for which the loudness measured in sones relative to the first frames of each beat is annotated.

The performance features of each feature are normalized by dividing each value by the maximum value of that feature in the sequence.

5.6.2 *Score features*

In order to study how the expressiveness of tempo and loudness is constrained by the score as well as to find possible interactions between performance features and specific

score features, beat meter position and beat rhythm features were extracted from each mazurka.

- **Score metrical structure**

  All the mazurkas from the dataset are in a triple 3/4 measure, thus the metrical structure per measure (bar) from each mazurka is encoded as $-1, 0, 1$. In this encoding, each value represents the first, second and third beat of the measure. Using metrical structure as a feature may allow for capturing structural relations within the bar, but also in relation to higher structural orders from the score. For instance, every 1st beat of a bar, or every 1st beat of every four bars, etc...

- **Rhythm**

  For the representation of rhythm in the LSTM input, I included the score representation of the melodic line from each mazurka. This choice is based on the fact that all mazurkas from the (collected) dataset contain (most of the time) a clear salient melodic predominance on the top voice (G-clef), while the "accompaniment" voice (or bass voice), is focused on the low voice (F-clef). Because of these characteristics, most of the rhythm variability is contained within the top voice.

  In order to obtain the top melodic voice of each score performed I implemented an algorithm to extract out of the top voice the highest pitch as the melody line. This algorithm was first introduced by Uitdenbogerd and Zobel (1998) and is commonly referred to as skyline. As discussed by Isikhan and Ozcan (2008), the skyline algorithm is one of the best performing methods for melody extraction when the accompanying voice does not contain pitches higher than the top voice, which is the case for all mazurkas used in this dataset.

  I implemented the skyline algorithm in Python using Music21 (Cuthbert & Ariza, 2010) to remove ornaments like trills or *appogiaturas*. Once the melody is extracted, the results were inspected manually and compared with the original score to discard any artifacts added when converting from MusicXML format to a Music21 object.

  To concatenate the input vectors from the beat-level representation in the performance features (tempo and loudness), the rhythm is encoded per beat measure. The encoding chosen represents each beat measure with a resolution of 12 units. Then, on each of these units, the absence or presence of the "onset" of a note (as indicated by the score) is represented by a 0 or a 1 respectively.

  Figure 5.5 shows the encoding of a (hypothetical) skyline over one measure. Since we are combining all features in one LSTM, we need to adjust the length of meter, loudness, and tempo with that of the rhythm encoding vectors. To do so, each value of meter, loudness, and tempo is repeated 12 times, thereby ensuring that all features are of the same length.

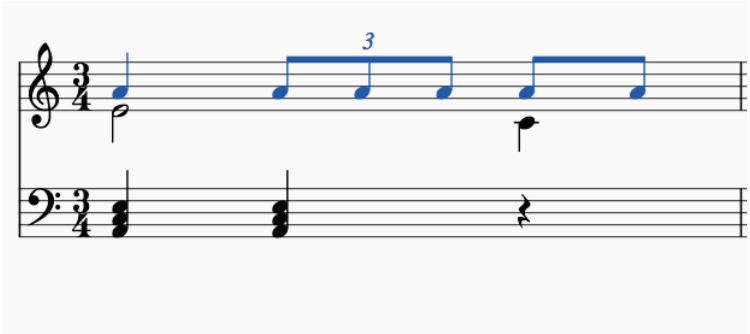


Figure (5.4) Example of how the rhythm represented in the melody line (in blue) is extracted from the polyphonic score. Figure 5.5 shows this rhythm encoded to be used within the LSTMs.

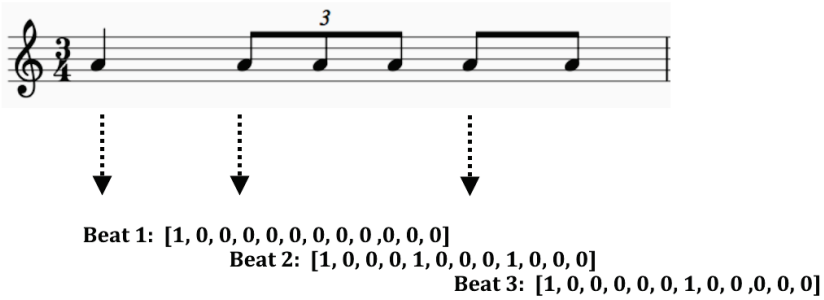


Figure (5.5) Encoding representation of the rhythm from Figure 5.4 with a beat granularity of 12 units.

5.7 MODEL ARCHITECTURE

I implemented the LSTM architecture by using the Python library Keras (version 2.0.2) (Chollet, 2015). The objective function aims to learn the parameters to effectively predict loudness or tempo at event t_n in time based on t_{n-1} and all previous events in the performance.

In this architecture, the inner activation function chosen for the LSTMs is a hyperbolic tangent function (*tanh*). The outer activation function chosen is the linear function. The loss function to estimate the error between the predicted ($\hat{y}$) and the true output (y) during each epoch is the mean squared error. Each model is run on 1000 epochs.

The hyperparameters are chosen by doing grid search cross validation based on the following values:

- Hidden units: 20 or 40

- Regularization:

  - Dropout on the visible layer (between the input and the hidden layer). Values used are 0 (indicating no dropout) or 20%.

  - L2 regularization with values 0 (indicating no L2 regularization) or 0.001.

- Optimizers:

  - Stochastic Gradient Descent, with the following parameter values: learning rate =0.01; parameter updates momentum=0.8; learning rate decay on each update = 0.0; no Nesterov momentum is applied.

  - Adam first order gradient based optimizer, with the default parameters as proposed on the original paper (Kingma & Ba, 2015).

For each of the models and features combinations, the mean squared error between the true values and the predicted values on the validation set is calculated after each permutation. Once the hyperparameters for each model are chosen upon the validation set (those that predict with a smaller minimum squared error), we retrain each of the models with those hyperparameters and test them on the 'unseen' dataset (test set) and calculate the MSE between the truth and the prediction again. Thus, we obtain the best performing models given a combination of hyperparameters in each of the a-1, a-2, etc.

Once we obtain the results from the different experiments, we can compare the error on the different models to estimate which models and combinations of features lead to more accurate predictions; this, being an indicator of how these features constrain the expressiveness of performers.

5.8 EXPERIMENTS

In order to study whether performers are more constrained by their idiosyncratic approaches to the music or by the score, two different types of experiments are carried:

- In the first experiment, predictions are done by mazurka and performer after modeling the same performer's style when playing all other 24 mazurkas. In this experiment the models aim to learn the idiosyncratic stylistic characteristics per performer.

- In the second experiment, predictions are done by mazurka and performer after modeling (learning) how the other 10 performers in the corpus play (only) the same mazurka. This experiment will elucidate on possible constraints based on the score and shared across performers.

In both experiments, we inspect the possible interactions between loudness, tempo, meter and rhythm (as notated on the score melody) and how these constrain the use of tempo and loudness in performance. Therefore, we also analyze whether the prediction error on each of the different models obtained is smaller as a consequence of the models learning such patterns of interactions.

5.8.1 *Persistence algorithm baseline*

The baseline proposed to evaluate the models on each of the experiments is a persistence model (also called naive forecast), which is often used in time series forecasting evaluating tasks. This is a one-shifted output algorithm that works by outputting at each event $t+1$ the values contained in t. The algorithms can be resumed as follows: having an input vector with *e.g.* the true values of a feature $x_n = [t_0, t_1, t_2, t_3]$, create a vector ($\hat{y}$) by shifting y_n by one position.

In order to have the same length than the original sequence, we remove the last value of y_n and include as the first value of $\hat{y}$ the mean of all values contained within sequence x_n. For instance: $\hat{y} = [\mu(t), t_0, t_1, t_2]$. Finally, the mean squared error (MSE) between x_n and $\hat{y}$ is calculated as an indication of the error (smaller MSE is better) estimated by the persistence model. In all of the experiments herewith presented our baseline is defined by this calculation for each mazurka and per pianist.

5.8.2 *Experiment 1. Performer-based models*

In Experiment 1, we obtain models for the prediction of tempo or loudness per piece and performer having trained each of the models on the same performer playing all other pieces of our dataset. With this experiment, we aim to show and study the individual differences in the possible interactions between performance and score features.

Having for each performer a dataset of 25 mazurkas, we partition the dataset in 6 blocks (rounded per iteration), leaving out one of the blocks for testing (`test`) and using the rest for training (`train`) and validation (`val`). The training and validation set are split again in 6, using 5 blocks for the training and 1 for the validation. Having a corpus of 25 mazurkas, for each rounded iteration block (on each experiment and per performer), we use 17 mazurkas for training, 4 for validating the models (finding the best combination of hyperparameters) and 4 mazurkas for testing. During the training and evaluation, we iterate over this process to cover the whole dataset and average results on all mazurkas per performer. As it is customary, the test is left aside and only used once the models from the validation set have been chosen upon their best fit.

Every sequence is padded (with a value not contained within our dataset) according to the length of the longest sequence. We do so by using the Keras function `Prepad`. In the dataset collected the longest mazurka (in score notation) contains 661 beat events.

5.8.2.1 *Experiment 1 results*

The average MSE over all predictions per pianist are shown for both tempo and loudness on Figures 5.6 and 5.7 respectively. Tables 5.1 and 5.2 show the p-values obtained after a one-way ANOVA as measured by Fisher's ratio of all models obtained on the predictions of tempo and loudness.

- Tempo predictions

The results obtained show that tempo is not better predicted when combined with the rhythm representation from the melodic line or when the model combines beat-level tempo, loudness, and melodic rhythm. The predictions of tempo either when adding loudness (t_b_t_l) or when adding meter (t_b_t_m and t_b_t_l_m) are better than in the models using only tempo in the input (t_b_t). Yet, according to the ANOVA test (table 5.1) these improvements are not statistically significant. All combinations of features lead to significant better predictions than the baseline persistence model ($p < 0.001$).

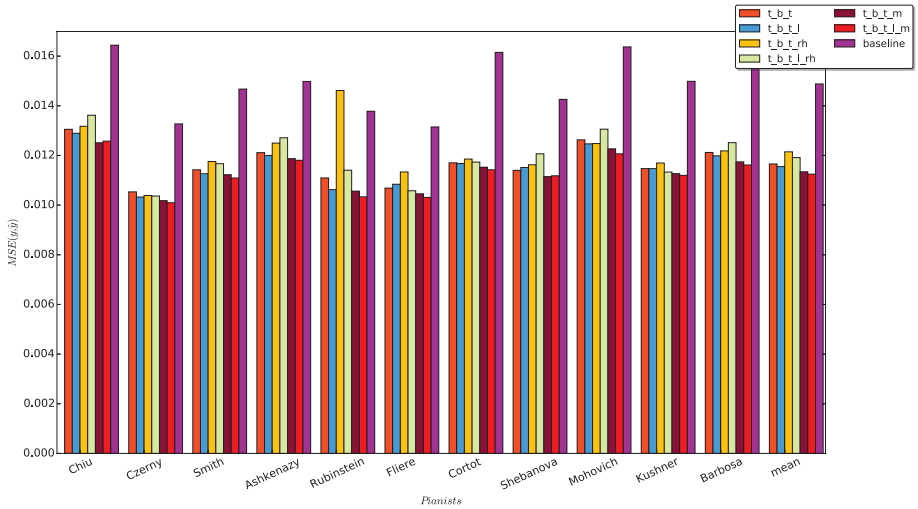


Figure (5.6) Experiment 1: Performer-based models tempo predictions. A smaller error indicates a better prediction

	t_b_t	t_b_t_l	t_b_t_rh	t_b_t_l_rh	t_b_t_m	t_b_t_l_m
t_b_t_l	0.7279					
t_b_t_rh	0.1973	0.1199				
t_b_t_l_rh	0.4606	0.3098	0.5755			
t_b_t_m	0.2979	0.4878	0.0373	0.1106		
t_b_t_l_m	0.1861	0.3227	0.0228	0.0688	0.7463	
baseline	0.0	0.0	0.0	0.0	0.0	0.0

Table (5.1) Experiment 1: One-way Anova Fisher's F ratio p-values over tempo predictions models in Figure 5.6

- Loudness predictions

Figure 5.7 and Table 5.2 show that loudness is significantly worse predicted when combined with (only) tempo (l_b_t_l) or with melodic rhythm features (l_b_l_rh, l_b_t_l_rh). The results also show that meter without tempo (l_b_l_m) leads to significantly worse predictions. It is however remarkable that when including tempo in addition to meter (l_b_t_l_m), the predictions are significantly better ($p<0.001$) than the rest of the models. In respect to the baseline, both the l_b_l and the l_b_t_l_m show significant ($p<0.001$) improvements.

The results obtained in Experiment 1 indicate that the idiosyncratic approaches to expressive loudness within this dataset are significantly improved by modeling the interactions with tempo and meter. This finding suggests that the expressive loudness idiosyncrasy of all individual performers is constrained by the communication of meter, for all the pieces within the corpus.

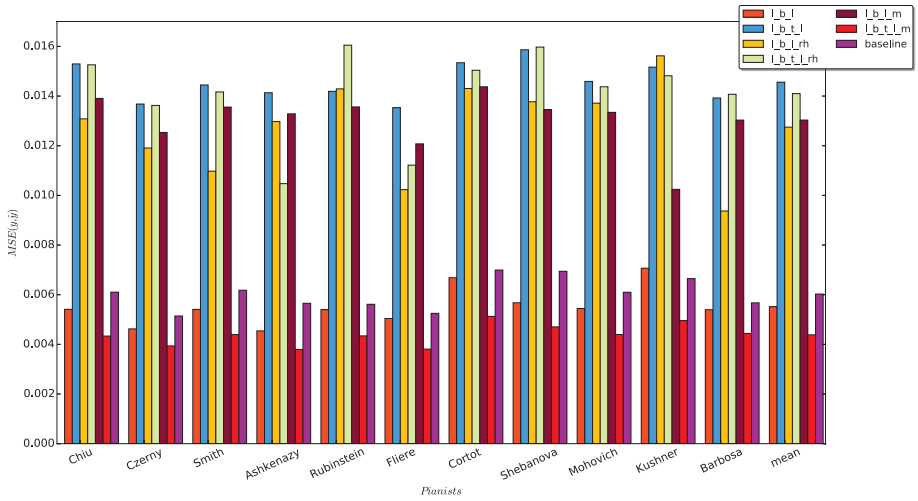


Figure (5.7) Experiment 1: Performer-based models loudness predictions. Smaller error means better prediction

5.8.3 *Experiment 2. Score-based models*

The goal of this experiment is studying how tempo and loudness are predicted per performer when models are trained on other performers playing the same piece. By doing so, the models inferred will learn the score structural expressive constraints shared across performers based on the commonalities of their expressiveness.

In order to avoid (or minimize) overlapping piece information between performances from the train / validation sets and the test dataset, each performance of the piece is

	l_b_l	l_b_t_l	l_b_l_rh	l_b_t_l_rh	l_b_l_m	l_b_t_l_m
l_b_t_l	0.0					
l_b_l_rh	0.0	0.0042				
l_b_t_l_rh	0.0	0.3948	0.0749			
l_b_l_m	0.0	0.0005	0.6455	0.0798		
l_b_t_l_m	0.0001	0.0	0.0	0.0	0.0	
baseline	0.0753	0.0	0.0	0.0	0.0	0.0

Table (5.2) Experiment 1: One way Anova Fisher's F ratio p-values over loudness predictions models in Figure 5.7

split by two. In this way, hyperparameters that are found in one half of the performance and models, are re-trained with the best hyperparameters on the other half of the performance. Unfortunately, this method cannot account for possible repeated motifs written in both half's of the score, but it is the best compromise possible to account for sequentiality while having the longest possible segments on each piece. We should note that the size of the dataset and length of sequences is conditioned by the fact that only beat-level performance annotations (instead of note-level) are available.

In this experiment, the error obtained per piece for all performers is used as an indication of how their expressiveness is constrained by the score and how do they share such expressiveness. In addition to the hypotheses presented in 5.5, it is expected that some pieces will constrain more the performers' expressiveness than others. In the same line of argumentation, the residual error obtained per piece and performer predicted is expected to contain some of the idiosyncratic gestures of the performer predicted. As such, it can be used as an indicator of how performers are constrained by the score structure and differ from the expressiveness of all other performers in this dataset. Yet, the residual error could also contain noise derived from other aspects not related to idiosyncrasy, or be caused by limitations of the modeling.

5.8.3.1 *Experiment 2 results*

The average MSE over all predictions per pianist are shown for both tempo and loudness on Figures 5.8 and 5.9 respectively. Tables 5.3 and 5.4 show the p-values obtained after a one-way ANOVA of all models predictions of tempo and loudness shown in Figures 5.8 and 5.9.

- *Tempo predictions*

Figure 5.9 shows that the predictions of tempo do not improve when combining melodic rhythm with either tempo (t_b_t_rh) or loudness (t_b_t_l_rh) at the input. For some performers, an improvement can be observed when tempo is combined with loudness (t_b_t_l). As in Experiment 1, tempo is, on average, better predicted when combined with meter (t_b_t_m) than when only using tempo as a predictor (t_b_t);

having the best predictions with the model based on tempo, loudness, and meter as input features (t_b_t_l_m). While these improvements are not significant in the ANOVA test, the results suggest that the interactions between tempo and loudness are better captured when including meter in the model.

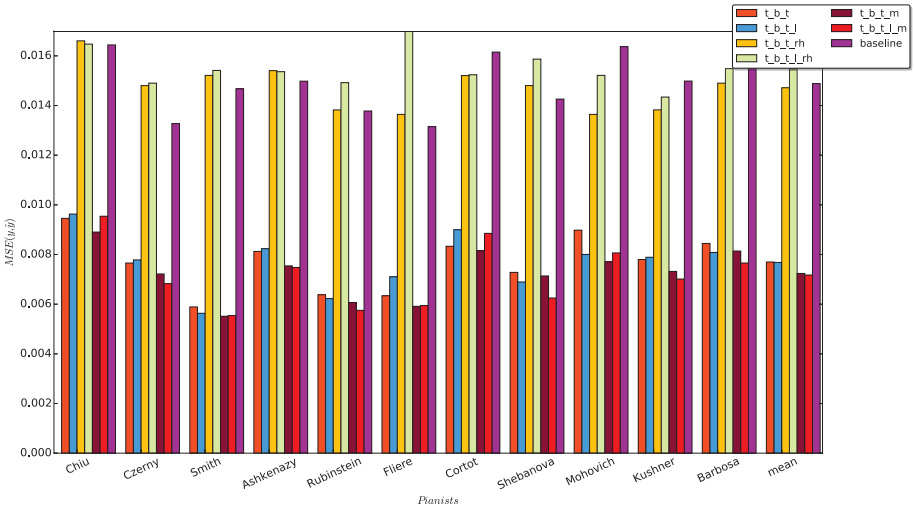


Figure (5.8) Experiment 2: Score-based models tempo predictions

	t_b_t	t_b_t_l	t_b_t_rh	t_b_t_l_rh	t_b_t_m	t_b_t_l_m
t_b_t_l	0.9672					
t_b_t_rh	0.0	0.0				
t_b_t_l_rh	0.0	0.0	0.4474			
t_b_t_m	0.2901	0.3145	0.0	0.0		
t_b_t_l_m	0.2816	0.3025	0.0	0.0	0.8915	
baseline	0.0	0.0	0.7565	0.6591	0.0	0.0

Table (5.3) Experiment 2: One-way ANOVA Fisher's F ratio p-values over tempo predictions models in Figure 5.8

- Loudness predictions

The results on the loudness predictions of this experiment show that loudness is better predicted when combined with tempo (l_b_t_l) and best predicted when combined with meter (l_b_t_l_m). Yet, when combining these features and comparing the predictions to the simplest loudness model (l_b_l), these improvements are not significant. As in the tempo predictions, while the models are not significantly different from the predictions of loudness, the results suggest that the interactions between loudness and

tempo are best captured by meter, indicating that it is a good predictor of score structure constraints in the expressiveness of performers.

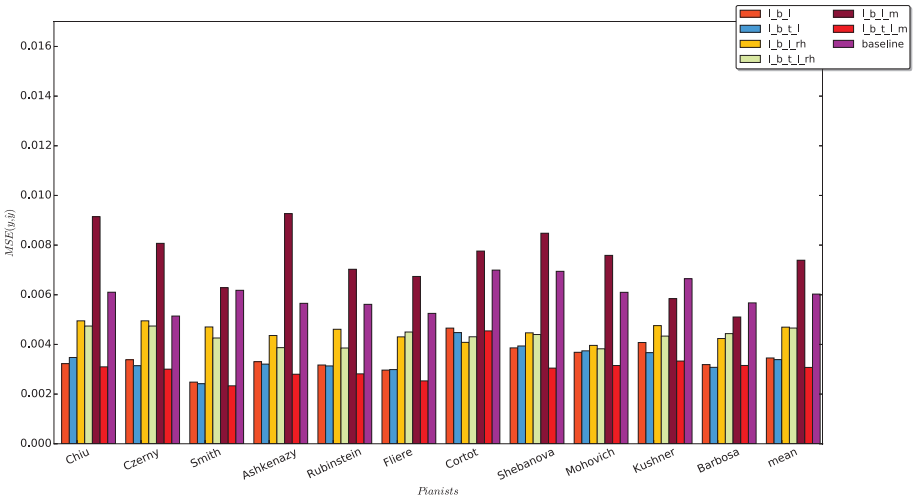


Figure (5.9) Experiment 2: Score-based models loudness predictions

	l_b_l	l_b_t_l	l_b_l_rh	l_b_t_l_rh	l_b_l_m	l_b_t_l_m
l_b_t_l	0.767					
l_b_l_rh	0.0	0.0				
l_b_t_l_rh	0.001	0.0	0.3132			
l_b_l_m	0.0	0.0	0.0	0.0		
l_b_t_l_m	0.1051	0.1641	0.0	0.0	0.0029	
baseline	0.0	0.0	0.0	0.0	0.0	0.0

Table (5.4) Experiment 2: One-way ANOVA Fisher's F ratio p-values over loudness predictions models in Figure 5.9

5.8.4 *Results analysis*

When comparing the results from both experiments it is evident that the models obtained in Experiment 2 (same piece played by several performers) lead to better results than the ones in Experiment 1 (based on the performers individual expressive style when playing other pieces). This finding suggests that the LSTM models are better in learning the performers' expressiveness when conforming to the structural constraints of the piece (as played by other performers) than to their individual idiosyncratic gestures. While for some performers their idiosyncratic expressive gestures might be well

defined and consistent across their performances, the models' prediction error is, on average, larger in Experiment 1 than in Experiment 2 and, thus, such gestures will be adapting to the musical material of each piece. That is, the variance in the musical material contained in Experiment 1, which is trained on 24 different mazurkas (per performer), is much larger than that one in Experiment 2, which is trained on only one piece (played by several other performers). Therefore, for the models obtained in Experiment 2, the temporal structure is more defined and easier to learn (and predict) by the models than in those of Experiment 1.

An unexpected outcome of these experiments is that using the melodic rhythm representation as an input feature has a negative effect (significant ($p<0.001$) in most experiments) in the predictions of either tempo or loudness. While the main goal of including this rhythm representation was to verify possible constraints in the use of tempo or loudness related to rhythm complexity, we shall not conclude that rhythm is not constraining the use performers make of timing and, therefore, tempo. An explanation for such behavior in the networks might be that the approaches each performer takes to rhythm in relation to the beat-level representation are very diverse and as such not captured by the models. This could probably explain why the error in Experiment 2 when adding melodic rhythm is much larger than in Experiment 1.

In all experiments, the best predictions of either tempo and loudness are obtained when combining loudness, tempo and meter. These results suggest that meter may effectively constrain the idiosyncratic expressiveness in the use of loudness and tempo as well as their possible codependency; not only in the short and long term structure of the piece being performed by several performers, but also on the individual approaches of performers to structural phrasing across pieces. That is, the approaches by individual performers to the use of phrasing in both tempo and loudness seems to be constrained by meter structure.

In Experiment 2 (score-based models) we can observe that the predictions of tempo (and somewhat, loudness) vary largely per pianist. This suggests that some pianists may have a more varied approach to phrasing than the rest of pianists. For instance, we can observe that Chiu's tempo is not as well predicted as on other pianists. The idiosyncrasy of Chiu in his use of tempo rubato in mazurkas is described by Cook in *Beyond the Score: Music as Performance*: *"Chiu uses rubato primarily for the shaping of phrases and cadences, as well as for a variety of rhetorical effects, but he does not use it as a basic means of accentuation [...]"* (Cook, 2013).

The results of the models in which meter is included suggest that the hierarchy between different phrasing approaches might be well represented in this structural level. In Sloboda (1985) it was shown that pianists were able to communicate loudness to listeners only with meter. While this finding might be true for a majority of music styles, the way pianists make use of expressive loudness in respect to meter might be much more flexible depending on the relevance given to the score structure by their use of phrasing. As it can be observed through the results obtained, the combination of meter and loudness as features (`l_b_l_m`) does not improve the predictions of loudness in respect to using only loudness (`l_b_l`), which is not the case for neither

of the Experiments presented. This indicates that the relation between meter, tempo and loudness is, at least for the experiment herewith presented, established through the interaction between these three features. Furthermore, these findings are coherent with those presented in Sloboda (1983), in which performers were shown to emphasize structure with changes in loudness and tempo.

5.9 VISUALIZATION OF PERFORMERS EXPRESSIVE IDIOSYNCRATIC DEVIATIONS

An important aspect of the models evaluation, is the visual interpretation of the results obtained per model and performer. In our case of study, we may want to use a visual representation which allows us to asses how accurate the models are in predicting a performance for a particular performer. For example, Figures 5.10 and 5.11 show that the models obtained in Experiment 2 are able to predict tempo and capture accurately structural changes in boundaries of varying lengths; this being one of the main motivations to use LSTMs models in the experiments herewith presented.

Plotting the temporal series' true and predicted values per feature is probably the most direct way to inspect those points in the predictions in which the models have not performed well, or rather, in which the idiosyncrasy of the performer is not well captured by the predictive model. Such an approach, withal, has the inconvenience that it requires inspecting all the sequences predicted and thus, by doing this manual inspection piece by piece, having an overall view of a performers style or characterization becomes a demanding task.

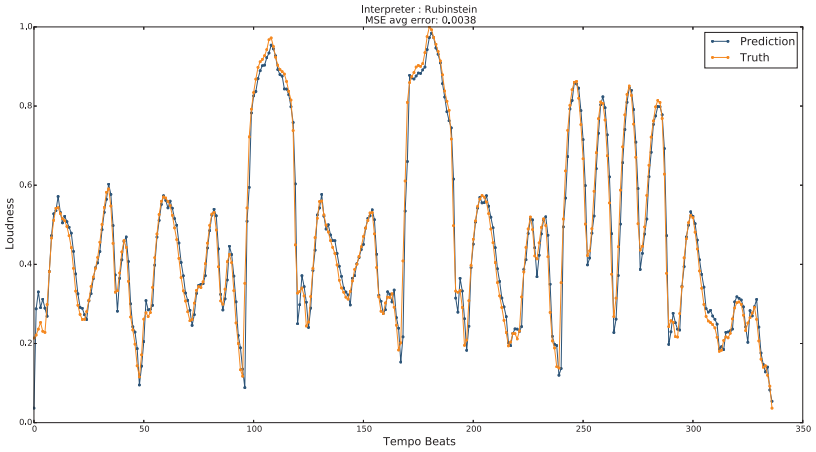


Figure (5.10) An accurate loudness prediction according to the l_b_t_l_m model (Experiment 2)

Other kinds of visualizations and methods can be useful to extract information over how the performers' idiosyncrasy is captured by the models used. An example of

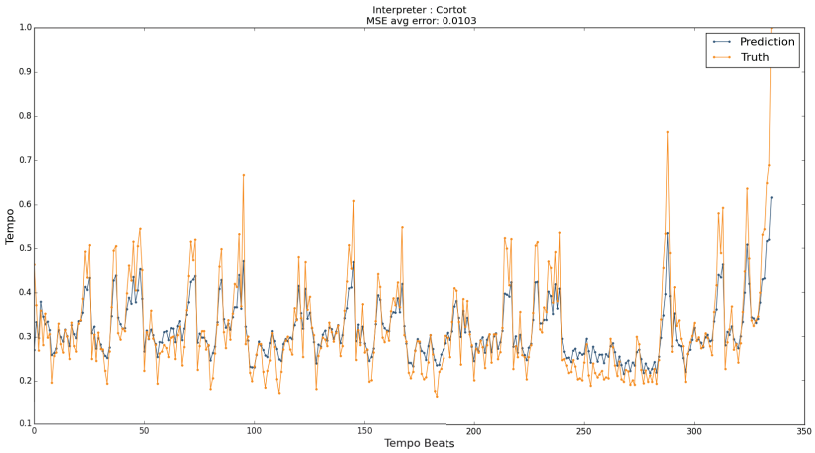


Figure (5.11) An accurate tempo prediction according to the t_b_t_l_m model (Experiment 2)

those is the Timescape and Loudness-scape plots introduced by Sapp (2007) described in Chapter 2, which can be used to analyze different expressive phrasing approaches to different performances of the same piece. With the aim of visualizing individual expressive patterns, Dixon, Goebl and Widmer (2002) presented a method to plot in a 2 dimensional space the tempo and loudness curves trajectory responding to the x and y axis respectively. The goal of such implementation is allowing for visualizing expressive patterns in real-time. Madsen and Widmer (2006) introduced an application of this visualization to explore individual styles. In their application, they used a pattern recognition algorithm based on the worm trajectories mapped to string. Their algorithm shows to be able to identify similar phrasing strategies and rank pianists according to their consistency in the use of matched phrases.

In this section, I present a simple approach to visualize the expressive deviations from each performer with respect to the "sequential" models obtained in the Experiments presented in Section 5.8. The models shown on this visualization are therefore the ones combining meter, tempo and loudness to predict either tempo and loudness as these have shown to be the best predictors for the experiments presented.

The first step for the visualization is collecting per performer the predictions and true values for all beat-level tempo or loudness across all pieces predicted. Having in each figure the x-axis representing the "Normalized Predicted values" and the y-axis representing the "Normalized True Values", if the model would predict perfectly the expressiveness for a performer, it should show a thin diagonal line departing from the (0,0) *origin* till the (1,1) coordinate, which we will refer to here as the hypothetical diagonal. The color of the dots represents the density of coincident values having blue as isolated dots and dark red as the maximum number of coincident dots for all pieces values for each performer. The density is determined by a kernel-density estimate using Gaussian kernels. Kernel density estimation is a non-parametric method to estimate the

probability density function of a random variable. Thus, the color bar scale on the right side of each figure represents the density of coincident dots.

In these visualizations, the further away each dot is from the hypothetical diagonal, the less well predicted this dot is estimated. A performer having a large representation of dots distant from the diagonal will, therefore, be less well predicted than another one in which most dots predicted are around the diagonal. The orange till red areas in the plots will represent the greatest density of performed true predictions. Thus, the warmer colors (orange - red) area often shows where the performers are best predicted according to the models. However, the density of the dots on each figure may also be interpreted as a degree of consistency in a performers' individual style, regardless of how close these estimations are to the diagonal. As it can be observed, the area with a greater density of dots is concentrated differently for each performer. Some performers show a predominance of red dots closer to the ordinate than others. In Experiment 1, this indicates how consistent performers are in their use of large or small deviation in tempo across all pieces according to their own idiosyncratic style. In Experiment 2, however, the area in which the density of scattered dots is larger, indicates in which sort of tempo or loudness values the pianists conform more to the norm as learned by the model (from the expressiveness of other performers). For instance, it suggests that the predicted variable (tempo or loudness) is predominantly used by that performer around the small values (e.g. Figure 5.12, Shebanova, MSE: 0.0114) while others use larger values (e.g. Figure 5.12, Rubinstein MSE: 0.0118) despite having similar MSE.

These visualizations shall complement the MSE error measures obtained in the models by showing the sparsity and density of the predictions. The visualizations in Experiment 1 show much more sparsity in the predictions than in Experiment 2, which aligns with the results obtained in the MSEs averages from both Experiments, being lower in Experiment 2 as a consequence of the less noisy characteristics of the data.

In Experiment 1 (Figure 5.1.2), for the tempo model, the best and worse predictions are Czerny-Stefanska and Chiu respectively. We can as well observe that both Czerny-Stefanska, Fliere and Rubinstein have a consistent use of tempo according to the model, having a rather large density area (in the shape of a diagonal). In the case of loudness, the plots show the differences in how sparse the data dots are for performers such as Mohovich and Cortot, and how much denser (and better predicted) are for Ashkenazy and Czerny-Stefanska.

In Experiment 2 (Figure 5.1.3), the visualizations per performer show how much each performer deviates per piece as played by all other performers. In the predictions of tempo, we can see that Smith, Fliere and Rubinstein are not only the best predicted but contain a rather large area of dots density, which suggests uniformity in the structural approaches to the pieces performed. In the loudness models, we can see however that Chiu and Czerny-Stefanska are predicted quite differently (MSE of 0.0023 for Smith and 0.0031 for Chiu) despite having a rather similar plot. In the case of Cortot, the predictions are the worse, probably suggesting more variety in the expressive approaches. In the plot by Kushner there is however a rather sparse area of dots that is, probably, counterbalanced by an extense thick red diagonal, which indicates that besides having

some outlier dots, the overall style is well predicted by the model. Thus, Kushner uses loudness according to the structure as predicted and learned by the model but often deviates by the norm defined by such model.

The goal of these illustrations is to contrast the results obtained with the MSE of the models. By means of these illustrations, we can point out certain patterns of performance that may not be well captured by the standard use of MSE or Pearson correlations. The visualizations of Experiment 1 thus show which performers are more consistent to their own performing style or which ones are less constrained by their idiosyncratic approaches to tempo and loudness. Therefore, how consistent each performer is to the norm defined by her/ his own playing on other pieces. The visualizations of Experiment 2, instead, may help to understand better how the structure of the pieces predicted as performed by all other performers, constrains each performer.

5.10 CONCLUSIONS AND DISCUSSION

In this chapter, I have used Long Short-Term Memory networks to model and analyze individual performers based on combinations of performance features and score features. I have shown how such combinations of features affect the predictions of expressive loudness and tempo. For such, I have presented two different experiments in which performers were predicted based on individual models or on models based on structural and shared approaches to the same pieces.

The most relevant finding in this study is that the expressiveness of performers in their use of tempo and loudness is constrained by the metrical structure. In particular, in all experiments, the best predictive models are obtained when combining tempo, loudness and meter in the predictions of either tempo or loudness. These results indicate that meter position serves as a relevant cue in the expressive gestures across tempo and loudness in its different hierarchical phrasing levels, and relates to the structure of expressiveness.

The relevance of meter in music has attracted much interest during the last decades across many disciplines. With the aim of elucidating how do we perceive and communicate meter and how its perception it is shaped by our exposure to music, several experiments and models have been realised. The findings presented in this chapter add to previous research literature by showing that the idiosyncrasy of performers in this dataset, when modeling tempo and loudness, is better represented and constrained by the interaction of these two expressive features with meter. In the case of Chopin's Mazurkas, effectively communicating meter by combining both tempo and accentuation while preserving the freedom and expressiveness in the rubato is of much relevance in the definition of individuality (Cook, 2013). The results obtained indicate that, regardless of how the models are trained, the l_b_t_l_m and t_b_t_l_m models seem to effectively learn the different ranges of expressiveness defined by the training set and accounting for both score constraints as well as those idiosyncratic to the performers.

In addition, in this experiment we observed how the predictions of tempo and loudness are much worsened when including the proposed representation of rhythm on the melody section, probably as a result of not being an effective feature representation. Yet, we shall not discard that the hypothesis of rhythm as an expressive performance constrain may not hold for the cases herewith presented, due to the large idiosyncratic expressive rubato of several performers.

Moreover, in both experiments, and for most models, the predictions of loudness are generally better than those of the tempo experiment, this probably as a consequence of loudness curves being more smooth and less variable than the ones of tempo. Moreover, we have also seen how visualizing expressive deviations between the true and the predicted values may be used as a complementary tool to have an overview of how consistent performers are in the deviations from the predicted values obtained on the sequential models used.

The experiment outcomes in this chapter validated one of the initial hypothesis regarding dependencies between performance features and score features. A possible reason why the interactions between tempo use and rhythm are not well captured in the absence of a metrical reference is that these features can operate on different temporal scales (beat-level vs. note-level). That is, for some expressive gestures, the beat-level granularity available within this dataset may have not been enough to capture more detailed interactions. Therefore, the difficulties of the LSTM network to capture such may also reflect the relatively limited size of the dataset. Future research concerned with diagnosing interactions between timing, loudness and rhythm may consider an accurate method to transcribe structure across beat, note and phrase scales.

Finally, the main motivation of using LSTM models, is not only to account for the temporal structure of music and performance expressiveness. Rather, and more importantly, to aim for a better understanding on how the norm expected by a listener relates to the "surprises" that performers elicit on the listener by their use of expressiveness. The relation between expressiveness, perceived musical tension, and score features has been explored already since Meyer (1956), who noted that the violation of expectations could lead to an increase in perceived tension. This concept is thus related to how listeners experience surprise as defined by Huron (2006) and introduced in Section 5.1.

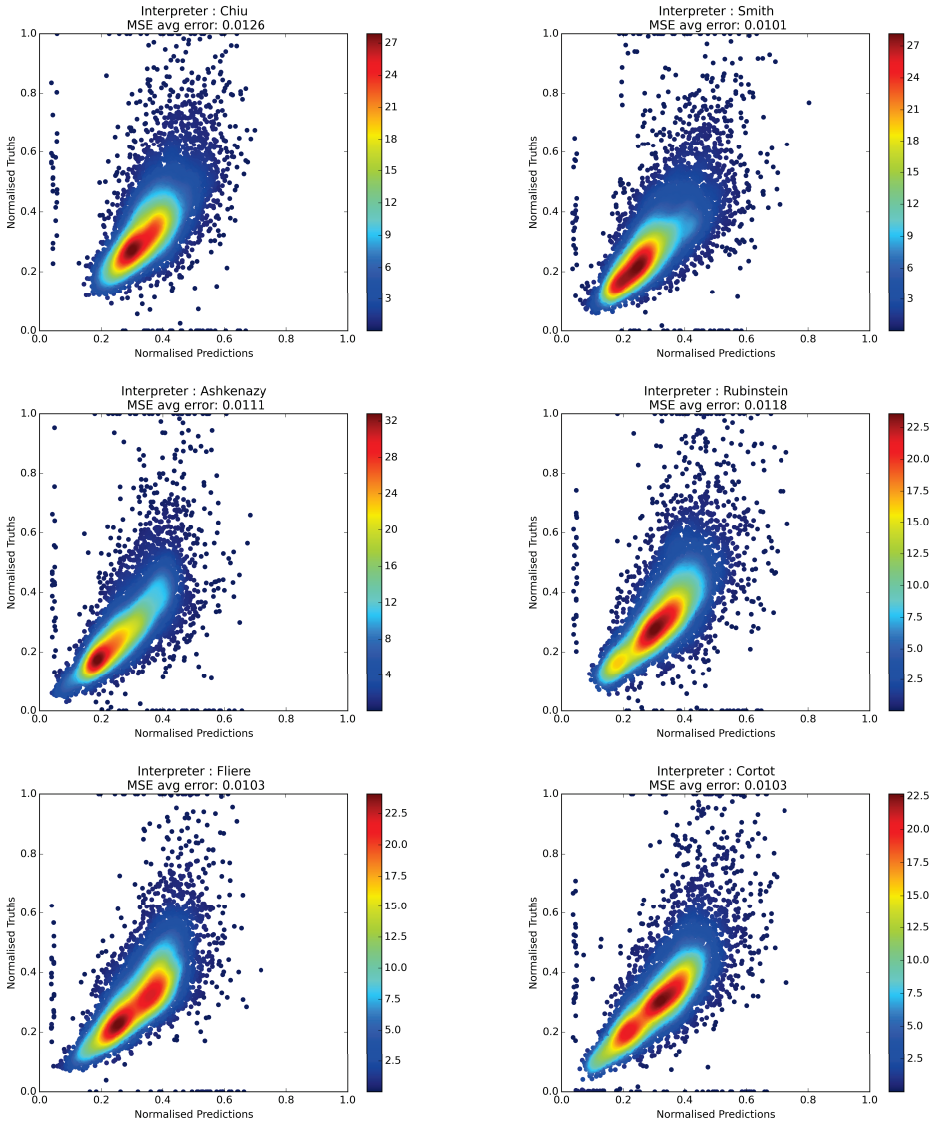
The predictive models herewith presented may relate to the norm as defined by the performers' style through the playing of other similar pieces (in Experiment 1), or to the norm defined by the structure as played by several other performers (in Experiment 2). With the current implementation of the LSTMs, we cannot make use of the expressiveness definition of the norm defined within the course of the performance itself. However, it is compatible with the definition of "expression within a unit as the deviation of its parts with respect to the norm set by the unit itself" (Desain & Honing, 1992). Thus, the LSTMs here presented may be associated with the listening process without an on-line update of the parameters which may represent such norm. A possible future endeavor could be using a Bayesian approach in which the posterior parameters could be updated in the course of the performance and therefore be sensitive to norms defined within the performance itself.

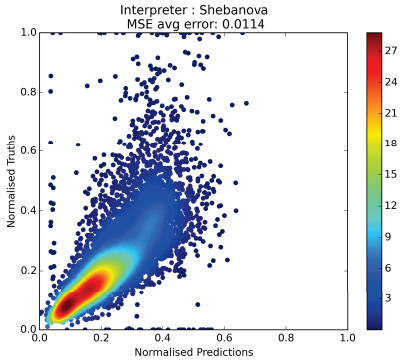
The relation between the perceptual expectations of listeners, the compositional structure of the piece being played and the expressiveness exercised by performers, has recently (re-)emerged as an area of interest in the field of performance modeling and analysis. Gingras et al. (2015) showed how melodic expectation as inferred by IDyOM (Pearce, Conklin & Wiggins, 2005) in the form of information content and entropy are effective predictors of expressive timing (in particular, the first derivative of the inter-onset intervals in a Preludes composition by Couperin. This finding indicates that performers use of melodic score grouping and that of expressive timing is strongly linked.

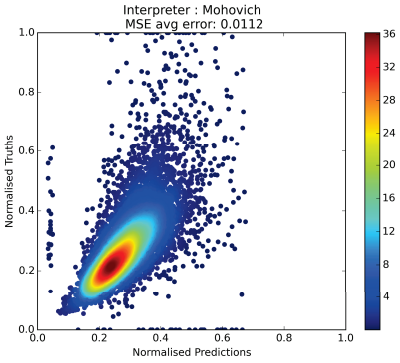
If the LSTM models with the right combination of features are able to capture and learn the expressive patterns defined by a number of performances we may also relate these to some of the expressive norms to be expected by a hypothetical listener. That is, when the listener would be equally exposed to the same corpus with which the models have been trained. While the leap is probably too bold as to assume such parallelism, the application of LSTMs to perceptual processes in music is rapidly emerging within the community. As an example of this, at the time of finishing the study herewith presented, Cancino-Chacón, Grachten, Sears and Widmer (2017) published a study in which they modeled with LSTMs the expressive uses of tempo and beat-level loudness based on performances by a professional pianist playing several pianos sonatas by W.A. Mozart. In their study it is shown how combining IDyOM features with metrical position can lead to better predictions of tempo (but not loudness) than when using only score features. Their results also show the power of LSTM models to account for sequentiality and structure in expressive performances.

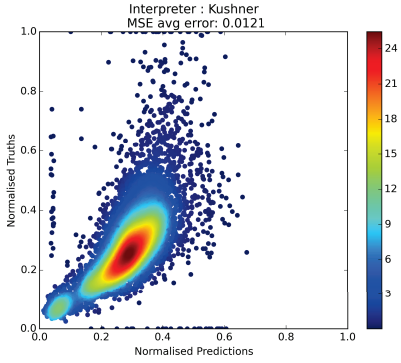
Future research could investigate the use of LSTMs when modeling perceptual processes in the recognition of performers individuality. This seems like a sensible direction to follow to better understand the mechanisms involved in the communication of idiosyncratic expressiveness from performers to listeners.

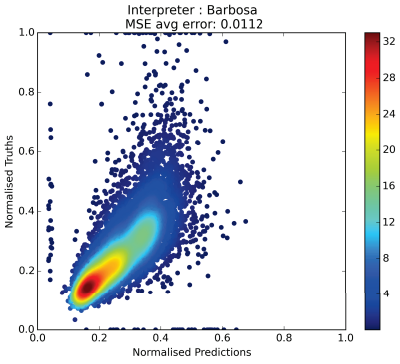
Figure (5.12) Experiment 1 Tempo Expressive Deviations profiles as defined by t_b_t_l_m model




Interpreter : Shebanova
MSE avg error: 0.0114
Normalised Truths
Normalised Predictions


Interpreter : Mohovich
MSE avg error: 0.0112
Normalised Truths
Normalised Predictions


Interpreter : Kushner
MSE avg error: 0.0121
Normalised Truths
Normalised Predictions


Interpreter : Barbosa
MSE avg error: 0.0112
Normalised Truths
Normalised Predictions

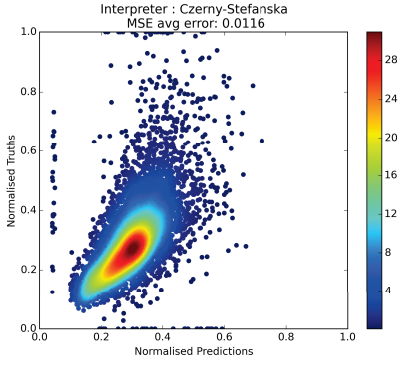
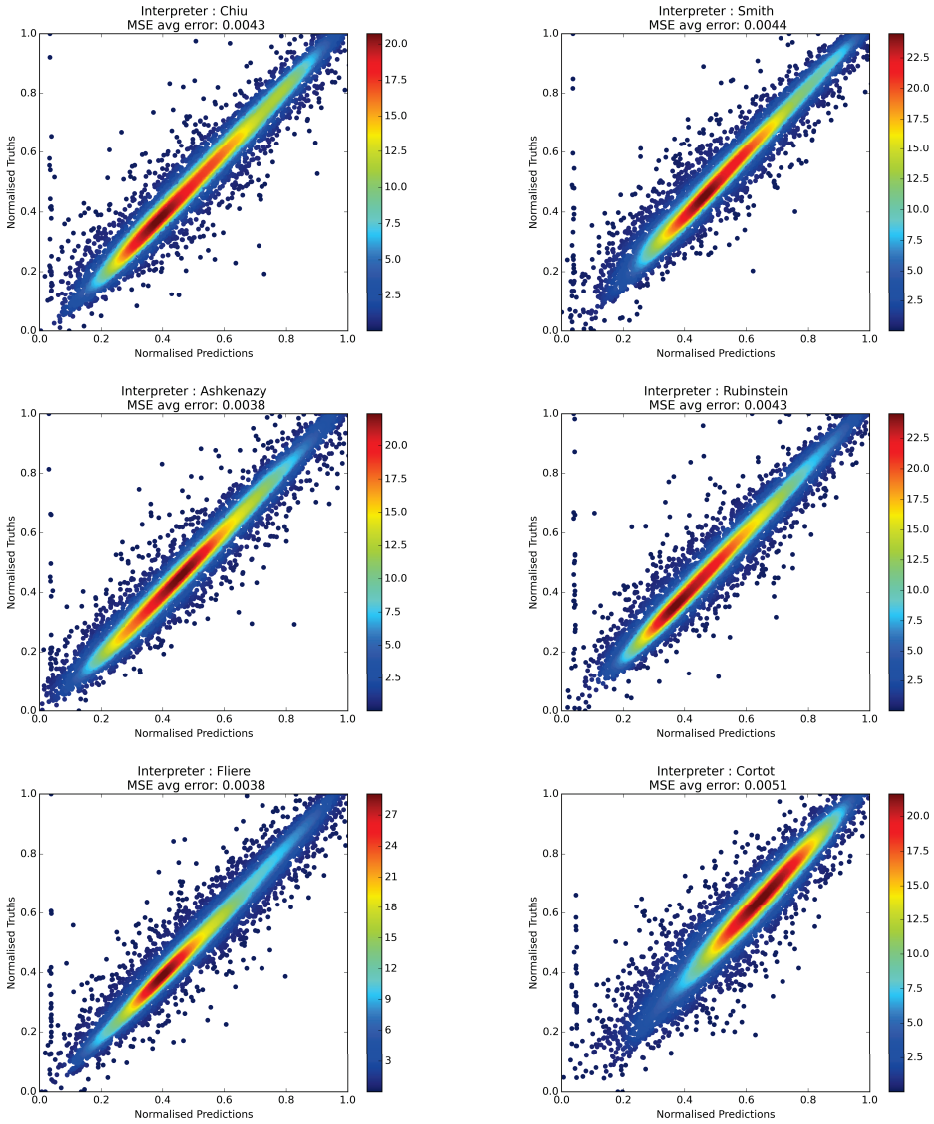
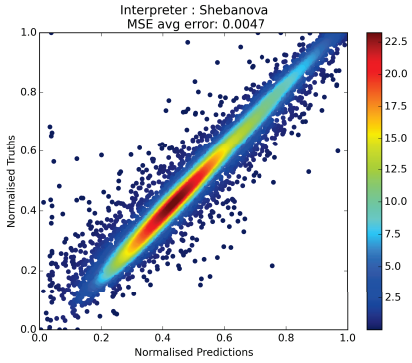
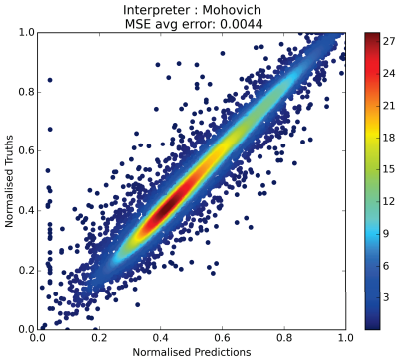
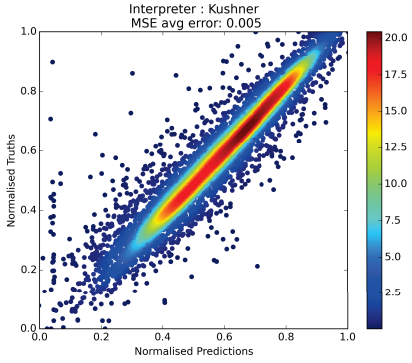

Interpreter : Czerny-Stefanska
MSE avg error: 0.0116
Normalised Truths
Normalised Predictions

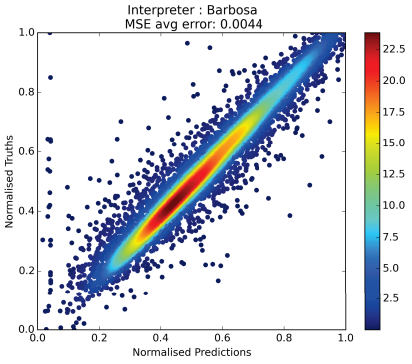
Figure (5.13) Experiment 1 Loudness Expressive Deviations profiles as given by l_b_t_l_m model




Interpreter : Shebanova
MSE avg error: 0.0047
Normalised Truths
Normalised Predictions


Interpreter : Mohovich
MSE avg error: 0.0044
Normalised Truths
Normalised Predictions


Interpreter : Kushner
MSE avg error: 0.005
Normalised Truths
Normalised Predictions


Interpreter : Barbosa
MSE avg error: 0.0044
Normalised Truths
Normalised Predictions

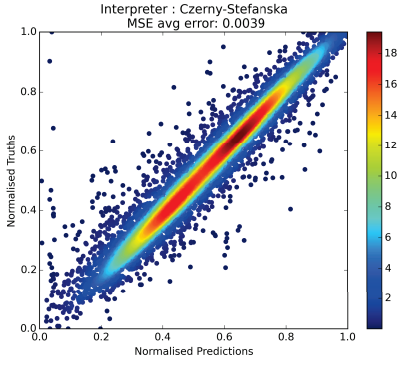
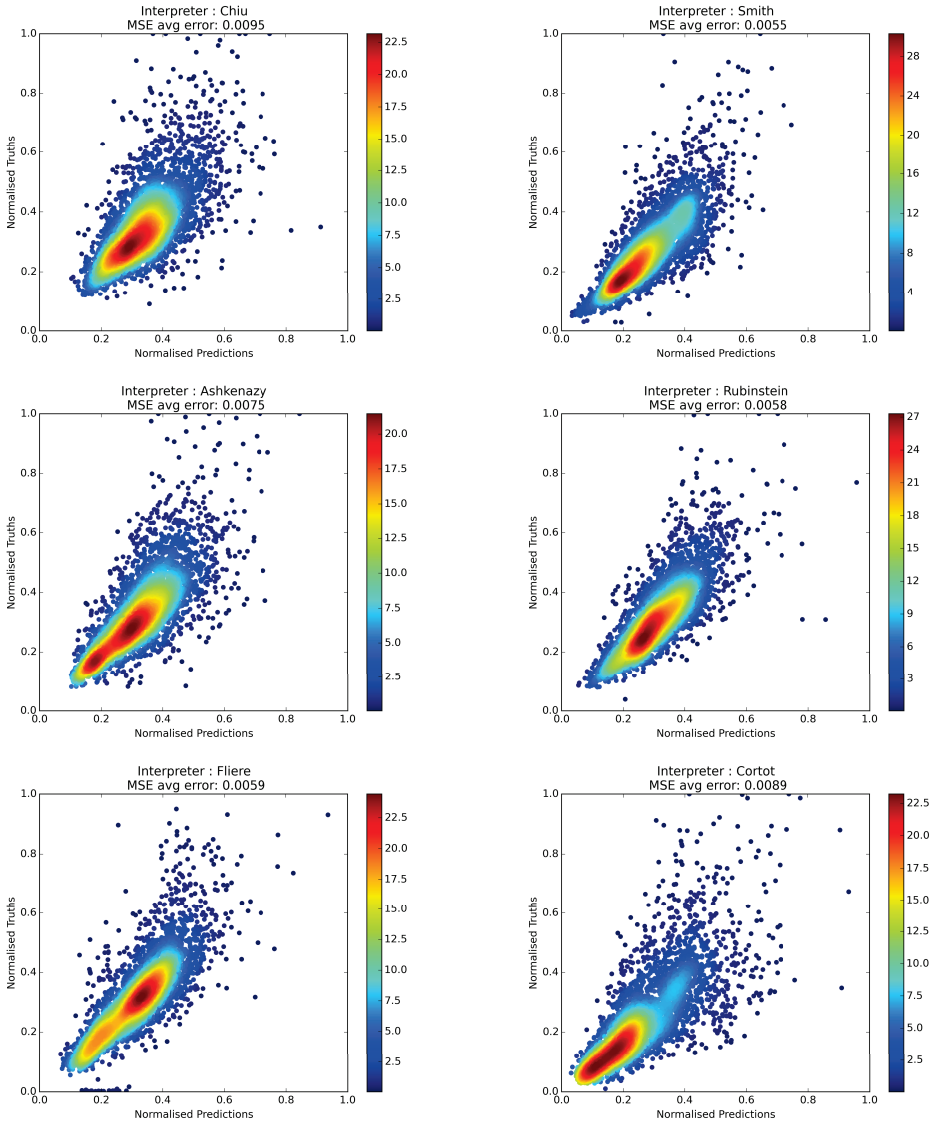
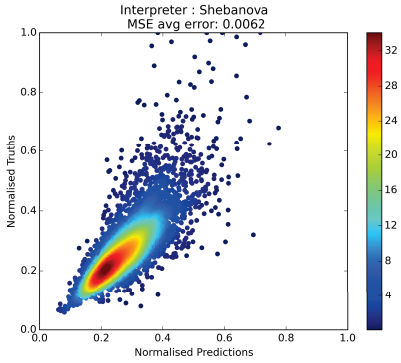
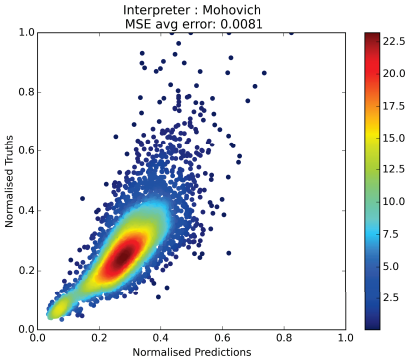
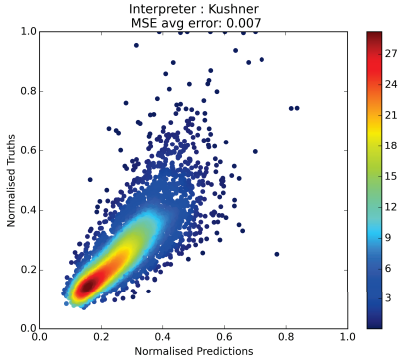

Interpreter : Czerny-Stefanska
MSE avg error: 0.0039
Normalised Truths
Normalised Predictions

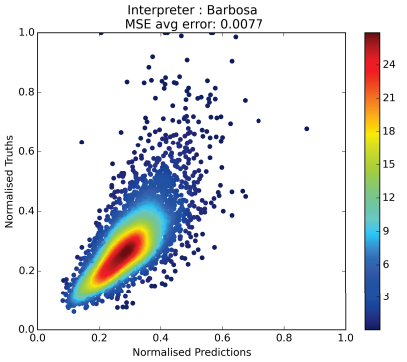
Figure (5.14) Experiment 2: Tempo Expressive Deviations profiles as given by t_b_t_l_m model




Interpreter : Shebanova
MSE avg error: 0.0062
Normalised Truths
Normalised Predictions


Interpreter : Mohovich
MSE avg error: 0.0081
Normalised Truths
Normalised Predictions


Interpreter : Kushner
MSE avg error: 0.007
Normalised Truths
Normalised Predictions


Interpreter : Barbosa
MSE avg error: 0.0077
Normalised Truths
Normalised Predictions

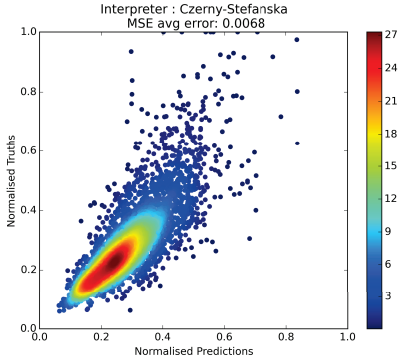
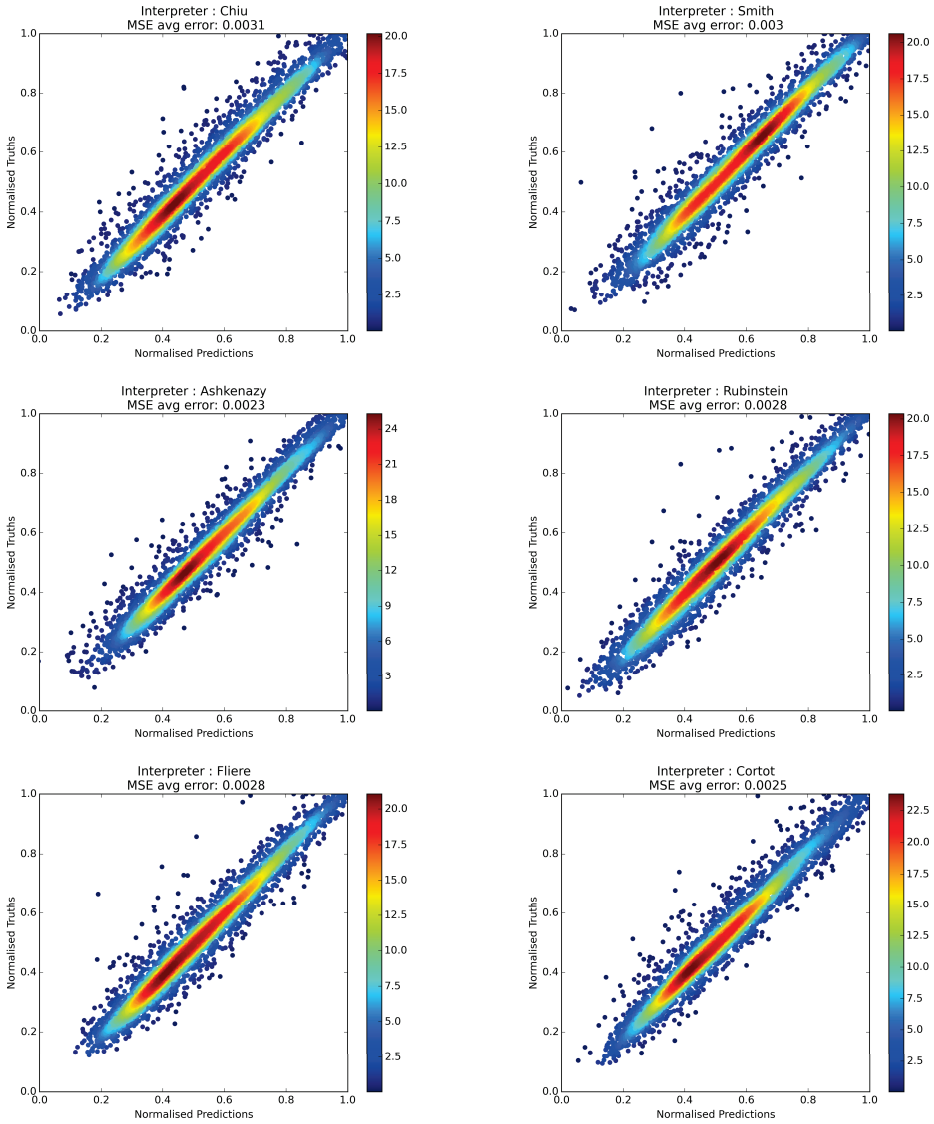
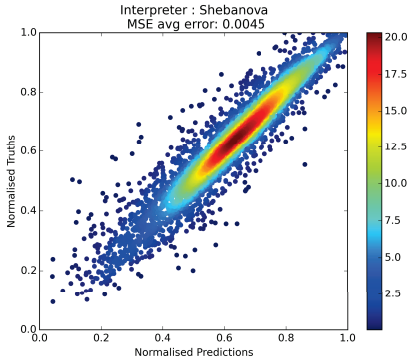
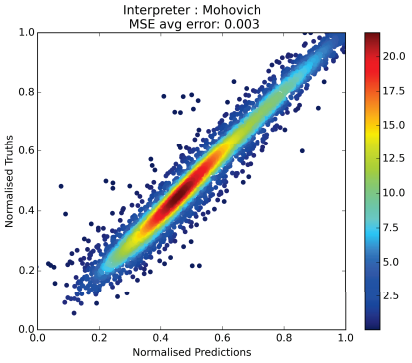
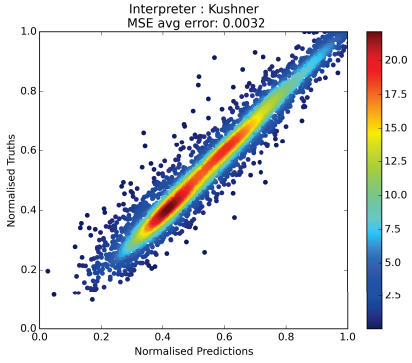

Interpreter : Czerny-Stefanska
MSE avg error: 0.0068
Normalised Truths
Normalised Predictions

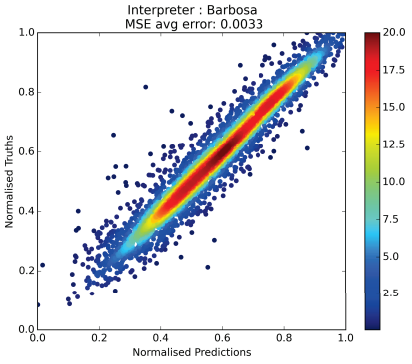
Figure (5.15) Experiment 2 Loudness Expressive Deviations profiles as given by l_b_t_l_m model

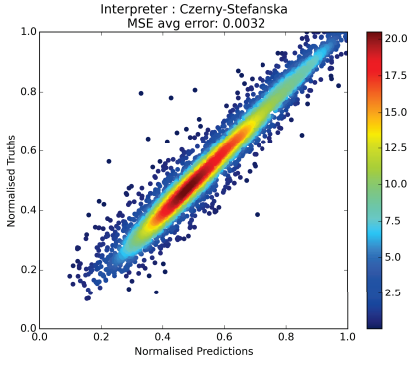



Interpreter : Shebanova
MSE avg error: 0.0045
Normalised Truths
Normalised Predictions


Interpreter : Mohovich
MSE avg error: 0.003
Normalised Truths
Normalised Predictions


Interpreter : Kushner
MSE avg error: 0.0032
Normalised Truths
Normalised Predictions


Interpreter : Barbosa
MSE avg error: 0.0033
Normalised Truths
Normalised Predictions


Interpreter : Czerny-Stefanska
MSE avg error: 0.0032
Normalised Truths
Normalised Predictions

6

A PERCEPTUAL STUDY ON THE ROLE OF EXPRESSIVE LOUDNESS AND TIMING IN A PERFORMERS DISCRIMINATION TASK

6.1 INTRODUCTION

Humans use sound for individual expression and communication to convey meaning and emotion, possibly as an essential survival mechanism (McComb et al., 2014). Before and after we are born, the different expressive acoustic cues we are exposed to may constrain differently how we relate to speech and music (Nakata & Trehub, 2011) and, like this, shape our recognition innate abilities (Trehub, Plantinga, Brcic & Nowicki, 2013). As such, the communication of individuality through expression in music can be understood as a sophisticated development in the production and perception of sound as a survival mechanism.

While the precise nature of any underlying motivations of performers, conscious or otherwise, lies beyond the scope of our current research, the literature and results presented in the previous chapters show that systematic quantifiable differences can be found in the idiosyncratic expressiveness exercised by performers. The aim of the research presented in this chapter is elucidating some of the mechanisms that listeners have in the perception of individual expressiveness in performers. For such purpose, in this chapter, I present a behavioral experiment to study whether listeners can discriminate between performers based on their use of expressive tempo and loudness. Furthermore, I discuss how the discrimination task might be influenced by musical expertise as well as the challenges I encountered in the methodological design of the experiment herewith presented. In the following lines, I will include some of the perceptual experiments I deem to be most relevant in the field of expressive performance.

6.1.1 *A review on expressiveness recognition experiments*

A growing corpus of research has contributed to defining the perceptual features that may play a role in the categorization and discrimination of expressiveness and how listeners respond to it. Repp (1994) studied whether listeners could discriminate between performances that had been tempo-transformed (speed up or slowed down to the same overall duration by a tempo range change of $\pm 15\%$) and the unaltered versions of these performances. The goal of this study was investigating if the perception of individual performance expressiveness would be affected by changes in tempo. The correct responses (55.6 %) of this experiment, were not significantly above chance across subjects ($p < 0.13$) and marginally significant across pairs of items ($p < 0.05$).

These results suggest that relational invariance[1] was preserved in the experiment. Yet, in a follow-up experiment with the same stimuli based on a subjective rating task, Repp (1996) found a significant effect ($p < 0.04$) of tempo affecting the perception of expressiveness when using a range of $+44\% - 23\%$ of tempo range change.

Honing (2006b, 2007) showed how listeners were able to distinguish unaltered recordings and tempo transformed versions of the same pieces. In these experiments, listeners were offered a web based listening environment where they could listen to pairs of performances "as often as needed". Afterwards, they were asked to point out which of the two stimuli was the original one. The results showed that listeners could effectively identify the original recording for classical and, in a less controlled setup, for jazz piano music. In a latter experiment, Honing and Ladinig (2009) showed how listening exposure (how often someone listens to music), rather than musical expertise or formal training, could have a role in the recognition of timing modifications.

The perception across expressive dimensions (e.g. tempo and loudness) has also been assessed in other studies in order to see how each of these may affect the perception of performers expressiveness. Repp and Knoblich (2004) did an experiment to study whether performers were able to recognize theirselves better when listening back to their own playing instead of other performers playing the same pieces of music. This is an indicator of perceptual identification of action through auditory cues. For this experiment, they asked professional pianists to record several excerpts. Once these excerpts had been recorded, they chose fragments of the excerpts free of mistakes and defined two different conditions. One in which the stimuli would preserve articulation, timing and dynamics (ATD stimuli), and another one in which timing and articulation would be preserved but loudness would be "stripped out" (AT stimuli). Seven months after the recording had been made, the same performers were invited to listen back and answer a questionnaire in which they had to answer whether they recognised the performances as theirs and rate it on a 1-5 point scale. During the listening tests, the participants listened only once to the performances, which might have been a relevant aspect of the recognition task. In this study it was found out that pianists were effectively able to recognize equally well with the AT conditions as with the ATD conditions, suggesting that articulation and timing contained sufficient information to recognize their own performances.

Timmers (2005) did a discrimination experiment to study how tempo and dynamics affect our perception of similarity between expressive performances. Based on a pairwise discrimination test, Timmers concluded that listeners' accuracy to asses similarity between expressive performance was best predicted when using global tempo or global tempo times ($\times$) loudness as features. In the same study, it was also shown these features were better predictors than the commonly found features in the literature, such as normalized variations or correlations in tempo or loudness profiles. Timmers' findings also showed that the listeners' musicianship and exposure affected the prediction depending on the piece listened to during the experiment. According to Timmers

1 Relational invariance indicates whether the same structural relationships hold between variables across two or more subpopulations (Mellenbergh, 1989)

such results might be task dependent and influenced by a bottom-up cognitive process. That is, the stimuli perceived influence our perception despite of our background knowledge.

Gingras et al. (2011) showed how listeners could better group performances of the same organ piece by the same performer when the organist retained their idiosyncratic ("natural") expressive gestures as opposed to intentionally play in an "as mechanical as possible" manner. This strengthens the hypothesis of how timing related features may contain sufficient information when listeners aim to group between performances. In order to verify this hypothesis, Gingras et al. (2013) extended this research to the harpsichord, having performers playing several pieces (only in a "natural" expressive manner). Based on this study, Gingras concluded that tempo and articulation were the most salient features of the performers' idiosyncrasy. The results of this study suggest as well that timing is a more stable feature than loudness when characterizing performances done by the same performer. In addition, Koren and Gingras (2014) did a perceptual test to investigate whether listeners were able to capture idiosyncratic properties from performers across different pieces and whether this depended on listeners music expertise. The results of Koren and Gingras (2014) showed that listeners could better group performers playing different pieces based on their use of mean tempo and note onset asynchrony (for instance, when playing a chord in which all notes are written as to be played simultaneously but they are played arpeggiated). The outcome of the same experiment indicated that tempo and note asynchrony were the most significant features to recognize expressive idiosyncrasy in performers, while loudness was not a significant feature in the same task. In this study, however, the role of loudness might have been constrained by the harpsichord itself, since it contains a limited amount of dynamics with a range of, normally, just 2 decibels between *pp* and *ff* (Penttinen, 2006). This is a marginal range when compared to a piano, which has a dynamics range of at least 14 dB between *p* and *f* (Askenfelt & Jansson, 1988). While a limited range in the dynamics is also found in many other the instruments of the baroque period, the harpsichord is commonly known for the limited dynamic expressiveness. As pointed out by both C.P.E. Bach (1752) and Quantz, the "imperfection" of the harpsichord (dynamics) should be contrasted by a suitable (and stylistic) use of timing and articulation (Quantz, 1752). Therefore, Gingras' findings on timing being one of the most relevant cues (followed by articulation) in the recognition of expressiveness, in contrast to others such as loudness, might be contextualized to the stimuli used during these experiments. That is, the characteristics of the repertoire as well as of the instrument chosen for such experiments may have constrained performers and listeners to focus on those expressive features that allow for more variation (timing and articulation) while ignoring those (loudness) that are hardly salient in respect to their expressive unit. It would be interesting to replicate such experiments when performers would play the same stimuli in a piano, rather than a harpsichord, and with a different stylistic approach that would allow for more versatility in dynamics.

We should remark that the perceptual experiments by Gingras and Koren, as well as those by Honing and Ladinig consisted of grouping tasks in which listeners were

allowed to listen as many times as they needed, as opposed to those mentioned by Repp or Timmers, which used a discrimination task instead. That is, in the grouping tasks mentioned, participants were able to listen several times to the pieces before grouping them, while on the discrimination tasks, the learning and discrimination occurs after listening to the stimuli one time.

In another study, Devaney (2016) investigates the relation between the classification of expressive features derived from timing, loudness, timbre and pitch by using support vector machines (SVM) and the perception of inter-performer and intra-performer features from both professional and under-graduate singers. Devaney (2016) concludes that both timing and loudness contain enough intra-performer solidness to be recognized automatically with the SVM. However, none of them are easily identified perceptually even by musically trained listeners. A point of discussion raised by Devaney is the difficulty to disentangle whether listener participants were focusing their attention to specific features.

Benadon (2003) studied the role of dynamics and timbre in the grouping of familiar and unfamiliar performers of jazz music, showing that listeners were able to recognize them after hearing only a few notes. Finally, in a distant but related line of research, the work by Bhatara, Tirovolas, Duan, Levy and Levitin (2011) concluded that timing and loudness are relevant features in the communication of emotional expression.

6.1.2 *Challenges in diagnosing expressiveness perception*

The studies mentioned, suggest that the listening strategies that listeners apply to distinguish between performances, may depend on both their exposure to the music and on the variability of the expressive features inherent to the music. Studying how loudness and timing may affect the identification of performers, is necessary to better understand the musical features involved in such a cognitive process. For these reasons, the main goal of the study presented in this chapter is to verify and extend the findings of studies such as the ones by Repp and Knoblich (2004), Koren and Gingras (2011, 2014), Gingras et al. (2011) and Devaney (2016). That is, by assessing if individuals can discriminate between performers playing the same piece, and if their (potential) ability to discriminate among unknown performers might be better explained by how they retain individual expressiveness in timing, loudness or a combination of these two features.

What role expertise and exposure play in the recognition of expressiveness and individuality in performers, has been shown to be complex to define. Expertise relates to explicit knowledge and exhaustive musical training, while exposure relates to active or passive music listening habits.

Ladinig and Honing (2006) and Honing and Ladinig (2009) showed that when distinguishing between original performances and those in which timing is linearly transformed, expertise is irrelevant, but exposure to music not, this being evidence of a shared ability to recognize expressive timing (which might be enhanced by exposure). In a review by Bigand and Poulin-Charronnat (2006), several experiments are presen-

ted to illustrate how, although expertise may play a role in discriminating subtle differences in musical expressiveness, exposure is enough to acquire several musical capacities such as: perceiving the relationships between a theme and its variations, perceiving musical tensions and relaxations, generating musical expectancies or integrating local structures in large-scale structures. Yet, according to Bigand and Poulin-Charronnat. *"Learning to expressively perform music certainly contributes to boost the processes involved in music cognition and emotion. Notably, good performers are likely to be more sensitive than nonmusicians to the small changes in musical surfaces that have deep emotional impacts on listeners."* (Bigand & Poulin-Charronnat, 2006).

In regard to musical expertise, Kendall and Carterette (1990) showed how musicians and non-musicians were able to distinguish different degrees of expressiveness, and Timmers (2005) showed that both musicians and non-musicians performed equally well when assessing the similarity between expressive performances. In an study by Gingras et al. (2011), performers' expertise was also studied and found to be a possible predictor of listeners' ability to group performers. Yet, in an study by Koren and Gingras (2011) it was shown than performers' expertise was not significant and, instead, the participants' scores in the study seemed to be piece dependent.

In another experiment, Koren and Gingras (2014) studied whether music expertise played a role in identifying harpsichordists when these were playing two different short excerpts from the standard baroque repertoire. In this experiment, it was found that, for one of the pieces pair comparisons, musicians grouping accuracy was significantly higher than that of non-musicians. However, for the other piece or when different pieces pairs were presented, only musicians succeeded better than chance levels in identifying both pieces as being played by the same performer. According to Koren and Gingras (2014), this contrasting finding suggests that the differences between features in the two pieces may have affected the listeners' ability to group them accurately. In addition, they argue that the listener's musical expertise is probably more relevant than the performer's expertise. Yet, it is unclear whether the results obtained in Koren and Gingras (2014) respond to a greater ability by musicians listeners to memorize the (structural) characteristics of a performance. To the best of my knowledge, the role of memory in expressive performance research has not been studied extensively, with the exceptions of Juslin and Laukka (2004) and Juslin and Västfjäll (2008) in episodic memory.

The task of performers discrimination could be related to that of object recognition in one-shot learning, by which humans are able to categorize incoming information based on one or very few examples and, by doing so, they "learn to learn" optimizing the amount of time required to categorize and process new, incoming information. The most common example of one-shot learning is that in which a person is able to find the similarity between faces of people even though their appearance might change slightly. One-shot learning has been widely accepted in the field of vision and counts with several computational models based on different methods (Koch, 2015). In the field of auditory and language perception it has been shown that children learn words at a very fast rate in a context of one-shot spoken word learning (Lake, Lee, Glass & Tenenbaum,

2014). Yet, in music research, it is unclear what are the perceptual mechanisms that allow for the recognition and identification of performers, in the context of one or few samples.

6.1.3 *Experimental design and hypothesis*

Motivated by the diverse findings in previous studies, as well as by the scarce amount of research on the possible interactions between timing and loudness as combined features when identifying performers, this study investigates what role timing and loudness, both as combined or isolated expressive features, may play in the discrimination between performers.

The experiment proposed below consisted on discriminating between performers after listening to pairs of stimuli on the different conditions presented. Having several conditions separating timing related variables from loudness ones, we aimed to find out how each of these features separately or combined may affect the discrimination of performers. By presenting pairs of stimuli as being played by the same performer or a different one in a two alternative forced choice, participants answered to whether each stimuli pair was played by the same or a different performer.

Based on the literature review presented, in this experiment I hypothesize that timing might be a more relevant feature than loudness when discriminating among performers. Depending on the exposure of the listeners to the stimuli, a combination of both timing and loudness would produce better discrimination among performers. Yet, in a discrimination setup, when presented with only one rendition of the stimuli, listeners are expected to better discriminate between performers when they can focus their attention on one of the features, being timing, probably more salient than loudness; this reinforcing previous findings by Timmers (2005) and Repp and Knoblich (2004).In order to test this hypothesis, different conditions were defined based on manipulations of the original performances, to isolate as much as possible the effects and, or, contributions of loudness and timing.

Finally, for this study, I also hypothesize that musicians expertise, based on explicit training (in many cases within classical music), may have an effect on performing the task, probably favoring them in discriminating correctly between the stimuli as being played by the same or a different performer.

6.2 MATERIAL AND METHODS

6.2.1 *Excerpts selection*

The dataset used for the experiments consists of five melody excerpts from piano pieces composed by Frédéric Chopin (1810-1849). The motivation for choosing Chopin is that, as a composer belonging to the Romantic period, his music is often performed with a rather extensive flexibility in timing (e.g. tempo rubato Cook (2013), which allows to better study the effect of this particular feature in respect to the perceived individuality

of performers. Furthermore, thanks to the Mazurka dataset from the CHARM project, there has been thorough research in the analysis and modeling of pieces belonging to this composer, which could inspire further studies connected to the study here presented. The chosen excerpts were extracted from Study Nr 1 in F minor from Three new Studies, Nocturne 1 Op. 37, Etude 7 Op. 25, Nocturne 1 Op. 9, and Nocturne 2 Op. 9. The scores containing the excerpts of all these stimuli are included in Section 6.5 of this chapter.

6.2.2 *Stimuli recording procedure*

Two professional pianists performed the music excerpts of 6.5 which served as a basis to design the stimuli for the experiments. At the time of the recordings of the performances, Pianist 1 was a 31 years old female with 27 years of piano practice experience. Pianist 2 was a 24 years old male with 18 years of piano practice experience. Both had completed Masters of Music in Piano performance in The Netherlands and had won prizes in International Competitions. The performers had extensive experience in performing the piano repertoire from the Romantic period and were able to practice the chosen excerpts during the week before the recording session.

Each performer was asked to record each of the musical excerpts (see 6.2.1) four times (takes). This was done with the aim of preserving ecological validity and accounting for the unavoidable variability among repeated performances of the same piece (even when performer intend to play them identically; Repp (1997)).

A precise tempo indication on each piece excerpt, previously agreed with both performers, was given. The pianists received an aural cue by means of a metronome during two (not written on the original score) bars before the first note of each excerpt. This was done with the aim of controlling for a similar starting tempo in the performances of the pieces, on each of the recording takes. The metronome would be switched off, once the first note of each excerpt was played. If the first note of the piece would start in an upbeat, the cue would extend until the next beat. This procedure was done manually and rehearsed a few times before starting the actual recordings. An illustration of this procedure is given in Figure 6.1.

After each take a short pause of between 30 seconds and 2 minutes was taken. If during the recording takes, the performer would play any different notes or rhythm than those written on the scores, performers would be asked to repeat the take and these takes would not be considered for the experiment stimuli. The performers were told to perform "their interpretation of the excerpt". No other instructions or indications were given.

The recordings were performed on a Yamaha GX-640 Electric Piano using the Grand Piano 1 preset as acoustic feedback and were recorded via the USB MIDI protocol with Ableton Live [2] by using a Macbook i7 with 8 Gb of RAM. Performers used a pair of Sennheiser HD-25 closed headphones for auditory feedback with a constant value

2 https://www.ableton.com/ Accessed on 15th January 2018

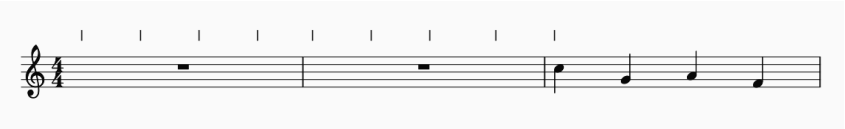


Figure (6.1) Illustration of the method used for initial tempo reference given of each recording take of the experiment stimuli. Each vertical line above the system, represents a sounding cue event from the metronome tempo indication. The last sounding event coincides with the first note of the excerpt to be performed. All the excerpts performed are included at the end of this chapter

for the volume across all performances. The recordings were done in a quiet room of a music school without distracting elements, such as external noises or interruptions during the recording session.

6.2.3 *Stimuli collected*

Mean and standard deviations of different expressive variables from the four performances recorded of each piece excerpt are presented in tables 6.2 and 6.1. Table 6.2 presents data from the length (in seconds) and loudness (in sones) of the performances, while Table 6.2 presents data from the articulation gap between notes (the bigger the value, the more *staccatto* the notes are), and note durations, both in seconds. The loudness was extracted from the audio renditions using the Music Analysis toolbox [3] and the MIRtoolbox [4]. For the sones references, a *piano* is considered to be around 4 sones and a *forte* around 16 sones [5].

	lengths mean (secs)		lengths std (secs)		loudness mean (sones)		loudness std (sones)	
pieces / pianists	L	P	L	P	L	P	L	P
Noct 1(37)	17.783	17.160	0.563	0.635	4.8574	5.2422	0.2980	0.2212
Etude 7	11.317	11.34	0.278	0.429	5.227	5.5655	0.2774	0.1889
Noct 9	22.501	23.072	0.358	0.281	4.9858	5.2989	0.0257	0.0663
3 Studies	15.151	18.414	0.317	0.430	9.0616	9.3064	0.2059	0.2779
Noct 2	20.005	20.065	0.407	0.199	4.3202	4.6001	0.0933	0.1932

Table (6.1) Mean and standard deviations on the performance length duration of the performances (in seconds) and loudness (sones) of the 4 performances per performer and piece as originally recorded

3 http://www.pampalk.at/ma/, accessed on 10-02-2018

4 https://www.jyu.fi/hytk/fi/laitokset/mutku/en/research/materials/mirtoolbox, accessed on 10-02-2018

5 http://www.sengpielaudio.com/calculatorSonephon.htm, accessed on 10-02-2018

	articulations mean (secs)		articulations std (secs)		note durs mean (secs)		note durs std (secs)	
pieces /pianists	L	P	L	P	L	P	L	P
Noct 1(37)	0.712	0.697	0.671	0.822	8.903	8.535	0.292	0.212
Etude 7	0.584	0.568	0.573	0.652	6.974	6.782	0.18	0.137
Noct 1(9)	0.475	0.465	0.488	0.550	10.523	10.612	0.181	0.108
3 Studies	0.366	0.377	0.402	0.440	7.586	8.922	0.149	0.156
Noct 2	0.411	0.419	0.438	0.462	11.404	11.512	0.238	0.139

Table (6.2) Mean and standard deviations in seconds of the articulations and note durations of the four recording takes per performer

6.2.4 *Experiment design and stimuli manipulation*

The performance recordings described in 6.2.2 were manipulated in order to create conditions based on different combinations of features for the discrimination task. Three different types of stimuli were derived from the recordings. The first type of stimuli (unaltered) preserved all expressive elements possibly captured through the midi protocol. These stimuli are thus similar to the ATD condition from Repp and Knoblich (2004). The second type of stimuli (deadpan timing) preserved the expressive loudness contours but used a deadpan version of timing. The tempo used to derive the metronomic score note durations (in mililiseconds) was the BPM initially given during the recording session for each of the excerpts. The third type of stimuli (deadpan loudness) preserved timing and articulation but flattened the loudness (midi velocity) to a constant value for all the pieces. These stimuli are equivalent to the AT condition from Repp and Knoblich (2004). The reference loudness value, was obtained from averaging over all performances of both pianists. Thus, being the same value for all performances of this condition. While preserving the articulation for the deadpan timing condition could be done by averaging different articulations over the duration of each score note represented, this was not done as this would imply creating a third condition to differentiate between loudness, articulation and timing, which would make the experiment duration much longer (around 100 minutes instead of 35). In order to keep the duration of the experiment shorter, timing, articulation and tempo were "grouped" as complementary features altered in the deadpan timing condition. Consequently, in respect to the loudness, the possible slight timbre variations originated due to loudness changes in the re-synthesis were grouped within the deadpan loudness condition.

The manipulation of the MIDI files was programmed on Matlab using the Midi Toolbox (Eerola & Toiviainen, 2004). All stimuli MIDI files (*altered* and *unaltered*) were re-synthesized using the Neopiano plugin [6], which uses samples from a Yamaha C7 Grand Piano as the base for the synthesis. Default presets (without any parameters

6 http://www.supremepiano.com/product/piano1.html Accessed on 17 January, 2018

such as artificial reverb or reflections altered) were used to preserve the characteristics of the original sounds of the electric piano as much as possible.

To verify if the audio renderings obtained from the Neopiano plugin kept the essential aspects of the (original) sounds pianists had listened to during the recording on the Electric piano, the similarity between both audio signals was assessed. For this, the audio rendition on both instruments, based on one of the performances of the "Nocturne 2" was recorded on the computer. The onsets of the audio of both signals were "extracted" with the beat-track algorithm from Dan Ellis [7]. The Euclidean distance between both onset vectors by point locations showed no deviations in their position (8.1759e-05). The possible loudness spectral differences between both stimuli were calculated with the Glasberg and Moore (1990) algorithm also used in Chapter 2 of the current dissertation leaning to a correlation between both STL signals of 0.74. Furthermore, the decay of the sounds from both the original piano as the Neopiano stimuli were checked by measuring the length of a single note played via the VST plugin as through the electric piano. A difference of 0.09 ms was estimated with the Mirtoolbox duration function. We may, therefore, conclude that both the main properties of the sounds used for the experiment preserved those the pianists were exposed to.

From the three different types of stimuli described (unaltered, deadpan timing, deadpan loudness), the following conditions were used (as experiment questions) for each of the five pieces:

- Same performer
  - Both stimuli unaltered
  - Both stimuli with deadpan timing
  - Both stimuli with deadpan loudness
- Different performer
  - Both stimuli unaltered
  - Both stimuli with deadpan timing
  - Both stimuli with deadpan loudness

6.2.5 *Procedure*

The experiment consisted of discriminating between performers after listening to pairs of stimuli in the different conditions presented. A short training session preceded the experiment. In this session, participants listened to three examples of stimuli pairs (containing different excerpts than those used during the experiment) and were informed about whether each pair had been played by the same or a different performer. Once the training session had finished, the experiment would begin.

7 http://labrosa.ee.columbia.edu/projects/beattrack/

Having six different conditions for each of the pieces, the experiment contained a total of 30 questions divided in five blocks (a block per piece). For each question participants listened to two stimuli in consecutive order belonging to each of the conditions presented and were asked to answer the question: "Were the pair of excerpts played by the same performer or different performers? Press 'z' for same or 'm' for different", to which they answered by pressing the keyboard keys assigned.

From the four different performances per pianist and excerpt available, a random performance was chosen in each of the questions to preserve the ecological validity in the idiosyncrasy of each performer. To avoid a possible effect derived from the order in which conditions were subsequently presented, the order of conditions, as well as the order of performers, was randomized for every participant. The core of the experiment was designed on PsychoPy (Peirce, 2007). The randomization of the order of stimuli presented as well as the adaptation to PsychoPy's interface was programmed using Python. The responses of each participant were recorded using PsychoPy. The experiment was performed on a laptop with an i5 processor and 4 Gb of RAM using a pair of closed headphones.

Participants read and signed an informed consent form before starting the experiment. Each experiment had a total duration of 25 minutes. Participants were offered a short break following every block. Once the experiment had been completed, participants were asked to fill out a demographics questionnaire and were compensated according to the regulations of the University of Amsterdam for their participation.

6.2.6 *Participants*

The experiment was performed by a population of 30 individuals who were recruited through flyers or email correspondence. The participants' population was chosen to be balanced by differentiating between musicians and non-musicians. The initial criteria to differentiate between both groups were the following: musicians would be defined as those who had at least five years of continuous performance practice in recent years or had received more than one year and a half of professional music education and were still actively involved in music performance, or a music related profession or studies. Non-musicians were defined as those who did not meet the criteria for being Musicians.

However, 4 participants were excluded in order to improve discrimination between these two groups based on Gingras et al. (2011) and thus, non-musicians were finally identified as those who had not received more than 1.5 years of any instrumental or singing music lessons; like this discarding possible amateurs with extensive expertise but still including people who had followed the compulsory music courses included in most primary school curricula in Europe and USA.

The final population consisted of thirteen musicians and thirteen non-musicians. To balance out for the filtering process, two random musicians were discarded from the dataset. The demographics data of participants is presented in Table 6.3.

Group	No. participants	Men	Women	Mean age (std)
All	26	14	12	27 (4.7)
Musician	13	6	7	29 (5.8)
Non-musician	13	8	5	26 (3.7)

Table (6.3) Participants demographics

6.3 ANALYSIS AND RESULTS

The results collected were analyzed using generalized linear mixed models (GLMM) to search for the effect of the conditions (unaltered, deadpan timing, deadpan loudness) on the discrimination task (same or different performer). GLMMs are an extension of generalized linear models which account for effects that vary per attribute and are relevant to the model. Therefore, in addition to the fixed effects already present in generalized linear models, GLMMs allow for the inclusion of random effects in the model, which may account for subgroups or observational blocks in the data such as, for instance, individual differences among the participants of the experiment herewith concerned. For these analyses, I used an R (R Core Team, 2012) implementation of GLMM by Bates, Maechler and Bolker (2012) included in the lme4 package. [8]

The ability to discriminate correctly between performers was collected under response accuracy of the participants. Timing, loudness, musical expertise and performer were included as fixed effects, while, in order to account for individual variation, participant and piece were included as random effects.

The total of variables considered for the statistical modeling and analysis are:

- Predictors
  - Performer (same or different)
  - Timing (deadpan or unaltered)
  - Loudness (deadpan or unaltered)
  - Participant expertise (musician vs no musician)
  - Participant
  - Piece excerpt
- Outcome variable
  - Response accuracy

In order to assess interactions between variables, several models based on combinations of fixed effects were compared based on the Akaike Information Criterion (as an indicator of the maximum likelihood estimation for each model). Furthermore, different subsets of the data were grouped when certain interactions were found.

8 https://CRAN.R-project.org/package=lme4

Table 6.4 lists the results of the model when including timing, loudness, musical expertise of the participants and performer as fixed effects, in respect to the response accuracy of the performer being same or different. Performer is shown to be significant ($\beta = -0.56, p = 0.0532$) as well as the interaction between Loudness and Performer ($\beta = 0.875, p = 0.015$).

	Estimate	Std Error	z-value	p-value
Performer	-0.5698	0.2947	-1.933	0.0532*
Loudness	-0.3240	0.2551	-1.270	0.2041
Timing	-0.1315	0.2565	-0.513	0.6082
Musical expertise	-0.1301	0.2082	-0.625	0.5322
Loudness*Performer	0.8757	0.3618	2.421	0.0155 **
Timing*Performer	0.1948	0.3592	0.542	0.5876
Musician*Performer	0.2381	0.2942	0.809	0.4184

Table (6.4) Mixed-models analyses of variance on the expressive parameters Performer, Loudness, Timing, Musical expertise. Significance codes: $p < 0.01$ '***'; $p < 0.05$ '**'; $p < 0.1$ '*'

Since the preliminary results showed an interaction between performer and loudness, and since performer showed to have a significant effect on the accuracy, the data was split in two subsets, divided by same and different performer. The models used on the subsets contained the same assignment of effects and random effects as in previous examples. In table 6.5, we can observe that loudness has a significant effect on the response accuracy when performer is the same ($\beta = 0.544, p = 0.033$). Timing or musical expertise do not show to have a significant effect in the response accuracy of these models.

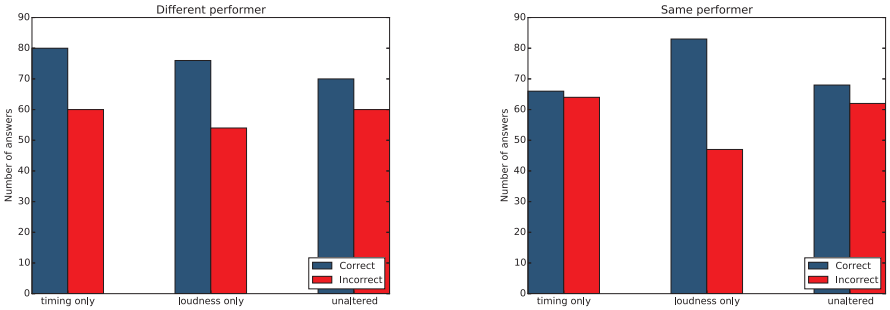
6.3.0.1 *Signal Detection Theory*

Signal Detection Theory (Stanislaw & Todorov, 1999) was used to assess whether there was a possible *response bias* in the participants' answers, as well as the degree of overlap (or *sensitivity*) between the signal and the noise distributions during the experiment. The sensitivity index includes correct responses (hit rates) and incorrect responses (false alarms). d' measures the distance between the signal and the noise means in standard deviation units (Stanislaw & Todorov, 1999). Having $d' = 0$ indicates that participants are not able to distinguish signals from noise. A negative d' indicates that participants are confused and would tend to respond yes when meaning no (Stanislaw & Todorov, 1999). A perfect response of all participants in the experiment would be represented by $+\infty$.

The obtained response accuracy suggests that participants were not able to effectively discriminate between performers ($d' = 0.16$). We must reckon, however, that the variety

	Estimate	Std Error	z-value	p-value
Performer same				
Loudness	0.544	0.255	2.136	0.033**
Timing	0.062	0.250	0.250	0.802
Musical expertise	0.106	0.213	0.501	0.616
Performer different				
Loudness	-0.3266	0.2561	-1.275	0.2023
Timing	-0.1325	0.2574	-0.515	0.6067
Musical expertise	-0.1310	0.2090	-0.627	0.5308

Table (6.5) Mixed-models analyses of variance on the expressive parameters Performer, Loudness, Timing and musical expertise (Musician). Analysis of subsets based on "performer same" and "performer different" conditions. Significance codes: $p < 0.01$ '***'; $p < 0.05$ '**'; $p < 0.1$ '*'



(a) Different performer (b) Same performer

Figure (6.2) Participants answers

of conditions used during the experiment may have had an effect in the response accuracy. Yet, d′ suggests that listeners were, most of the times, guessing.

The response bias in a same/different task such as the one presented in this experiment can be quantified with β. If participants would not be inclined to respond same or different performer we would obtain $\beta = 1$. Values of less than 1 show a bias towards responding same, in our experiment this being same performer. Values of more than 1 indicate a bias of responding no. The *response bias* measure showed that participants seem to be slightly biased to answer that the stimuli presented on each question were played by the same performer rather than by different performer ($\beta = 0.99$). Yet, being so close to 1, the bias shows to be minimum.

6.4 DISCUSSION

The abundant research in computational modeling of expressive performance (Gabrielsson, 2003; Palmer, 1996; Widmer & Goebl, 2004) has shown us what can be measured in expressive performances and recordings, yet we still know very little of what listeners actually hear in them (Clarke, 2002). In as study by Timmers (2005), it is shown that some of the common measurements used to represent performance data do not correspond well to what is perceived by listeners. In line with this argument, Honing (2006a) advocates for the need of new perception based measurements to validate and compare the predictions made by computational models. Reasonably, empirical research may contribute considerably to find out which methods are more suited to validate expressive performance models.

Based on the obtained results, the initial hypothesis of expressive timing being a more relevant feature than loudness, can not be supported by the experimental scenario herewith presented. When analyzing the effect of expressive timing present in the stimuli, we found no evidence of its significance in relation to response accuracy. Loudness instead was found to have a significant effect on the accuracy in the same performer subsets, contrary to what was hypothesized. When looking at Table 6.5 we can observe that, in the same performer condition, loudness is significant ($p < 0.05$), although it is not in the different performer condition. These results suggest that listeners may be sensitive to the contour of loudness and per note dynamic variations, but not significantly for the different performer condition; given that the mean loudness shown in Table 6.1 indicates small differences between the mean levels of the stimuli for both performers. Having such a variety within stimuli, with the goal of keeping ecological validity, is a common experimental approach. This is seen, for instance, in Timmers (2005), in which mean tempo or global tempo is considered as a predictor variable with varying tempi. Nonetheless, these findings could be contrasted with a study in which the velocities from all stimuli would be mean centered, or in which alternative dynamic expressive conditions for the same excerpt would be considered.

An unexpected outcome of this study is the fact that timing (AT, timing only condition) was not found to be a relevant feature to discriminate performers. This outcome does not align with the findings of previous studies, such as those introduced in Section 6.1.1 by Repp (1994), Timmers (2005) and Koren and Gingras (2014). Moreover, it contradicts the hypothesis originally presented in Section 6.1.3 that stated timing being a sufficient and more relevant expressive feature than loudness in a discrimination task. An explanation for this contradictory and unexpected outcome could be that the ecological validity across the four stimuli recorded (per piece excerpt and performer) introduced too much variance; especially, reckoning the size of the participants population in the experiment.

Even though pianists were asked to repeatedly record 'their' version of the excerpts presented, small variations in timing, articulation and loudness occurred in the different performances. These variations, which can be seen on the standard deviations of Table 6.1, may have confounded participants during the experiment; making the task of

retaining expressive characteristics between performers very difficult. In line with this, the fact that participants are able to differentiate among 'same performers' in the loudness only condition, could be affected by the fact that, in those stimuli, time is flattened. Thus, we should consider the possibility that participants were answering 'same performer' while actually recognizing identical timing, instead of similar loudness. If this would be the case, listeners would discriminate upon 'identical inexpressive' stimuli instead of 'similar expressive' gestures.

A confounding factor for the listeners during the experiment might have been the different tempi development within the stimuli used, which is reflected on the durations of each. In regard to the stimuli collection, we must note that the method used during the recordings (by giving a metronomic indication to pianists before playing the first note on each recording take) differs in a fundamental way from the study by Repp and Knoblich (2004). In their study, Repp and Knoblich used a metronome during the entire recording process of the stimuli. This was done with the intention that performers' expressive variations would result from only their use of timing and articulation. Based on the studies by Honing, however, we know that the link between tempo and timing is not easily separable (Honing, 2013). Therefore, deviating in timing while having a constant tempo limits in a fundamental manner the expressive freedom in the phrasing of performers. Especially, in music from the Romantic period, which is characteristic for its rubato (Cook, 2013). Yet, having the freedom to deviate through timing will lead to unavoidable tempo phrasing elongations or shortenings, which might have been essential cues in the experiment presented in this chapter. Thus, there is a trade-off between tempo preservation vs. a more ecological and valid material accounting for tempo variations (as a consequence of timing and phrasing) which should be further explored in the future. For the present study, I chose, for (quasi full) ecological validity. In relation to this, another aspect to be further investigated is the effect on note level deviations for each pair of stimuli presented during the experiment. Future research could reckon studying the use of an appropriate distance measure (or information content index) between both stimuli, in order to attain which deviations are most significant from a perceptual perspective, and how these may affect in the discrimination task.

Not having found timing as a significant feature to discriminate between performers, does not necessarily contradict the previous findings by Gingras et al. (2011), in which listeners performed better at recognizing pieces played by the same performer. We must be aware though, that the nature of the experiment and study here presented is completely different. Thus, the accuracy for each task is probably affected by the experimental design here presented. In Gingras et al. (2011), participants were asked to realize a sorting task and repeatedly listen to the musical excerpts to be sorted. This process probably allowed participants to be more sensitive, through familiarization, to micro and macro-structure level expressive nuances than with the experiment here presented, in which participants were exposed to single renditions and different comparisons in each question.

A challenge commonly found within the studies of expressive performance perception is that, while the research questions are related, the different approaches in the

methodology and nature of the experiments hinder the comparison of their results. Since the initial experiments by Seashore we can mainly distinguish those studies in which the methodology is based on a (one-shot) single stimuli pair discrimination (such as in Repp (1994, 1996), Repp and Knoblich (2004), Timmers (2005)) and those in which the recognition is done following a (unlimited and uncontrolled) number of repetitions of the stimuli listened to (such as Gingras et al. (2013, 2011) or Honing (2006a, 2007)). It is thus unknown what the precise effects of exposure may have been during the experiments of Gingras and Honing, and whether the lack of findings in both Repp's experiments as in this study could be related to the fact that one-shot (single) stimuli learning is not enough to perceive the subtle nuances in expressiveness and discriminate between performers. The sensitivity of listeners to such nuances might have been better captured within a comparison experimental setup than in a discrimination task.

To allow for a comparison between these two methodological approaches, we need to clarify the role of repetition in the listeners' ability to retain expressive nuances across stimuli. Based on the evidence of discrimination tasks in expressive performance, such as the one presented in this study, more research is needed to understanding how the retention and recognition of the features herewith studied may depend on exposure to them during the recognition task.

The most solid findings so far are based on experiments in which the amount of variables are either limited to one or are related (e.g. timing or timing and articulation). Yet, the studies by Timmers (2005) and Devaney (2016), suggest that, with sufficient exposure, the interactions of timing and loudness are well perceived by listeners and help them to discriminate between performances. In the experiment presented in this chapter, the main motivation for choosing a discriminative task was to control for the number of repetitions (none) in pro of covering a larger scenario to investigate the role of loudness and timing. This choice, however, might have been a remarkable constrain. Arguably, the sensitivity test results could probably be due to participants being overwhelmed or saturated by the variety of conditions presented, duration of the stimuli and length of the experiment, which may have surpassed their cognitive limitations. Based on these findings and arguments, we cannot discard the hypothesis of timing being the most important cue in performers' discrimination.

As exposed by Clarke (2002), one of the challenges that listeners have when listening to performances, is to keep track of a series of events that develop in time, and to find sense in them. During the listening process, there is an interaction between exposure (memory) and the 'perceptual present' (Clarke, 2002) which makes complex the elucidation of the perceptual mechanisms involved in the recognition of individuality in performers. Future research could consider a better understanding of the effects and consequences of the different experimental approaches discussed in this chapter, with the aim of finding a more homogeneous methodological strategy. By doing so, we may understand better what retention and attention mechanisms listeners use when several expressive features are present, in order to perceive idiosyncratic characteristics in performers.

6.5 EXCERPTS USED AS STIMULI FOR EXPERIMENTS 1 AND 2

Figure (6.3) Excerpt 1, from Nocturne 2 Op. 9

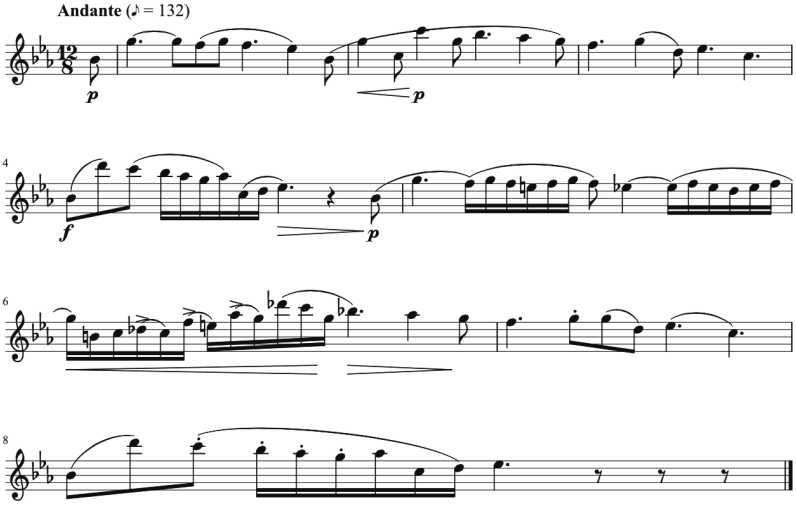


Figure (6.4) Excerpt 5, from Study Nr 1 in F minor 3 New Studies

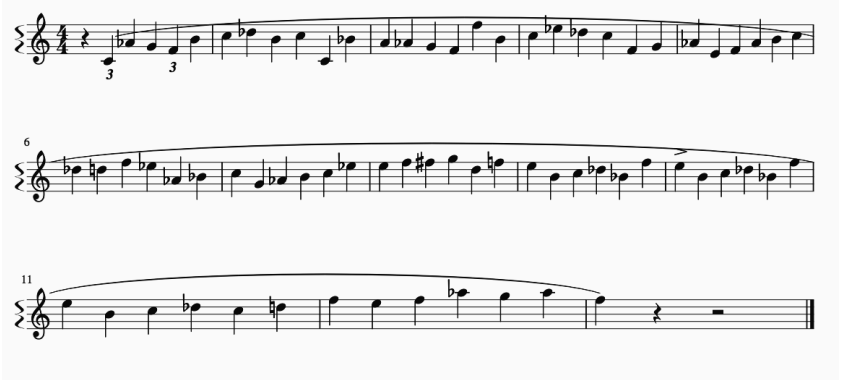


Figure (6.5) Excerpt 3, from Nocturne 1 Op. 9

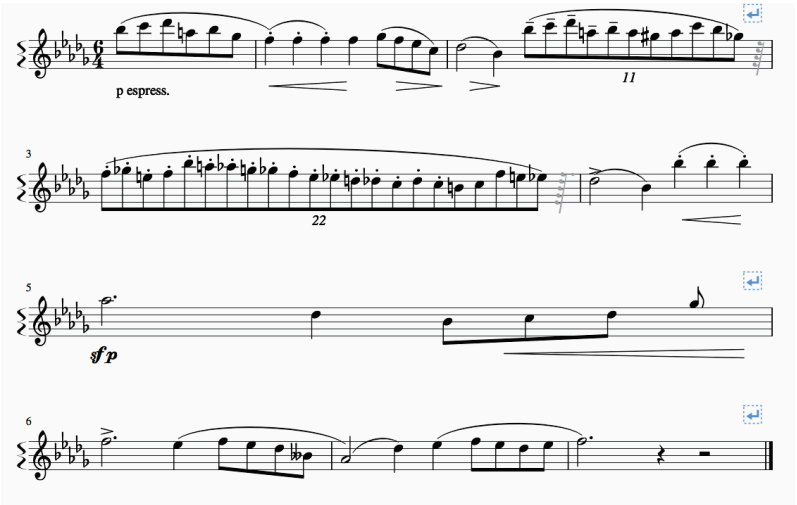


Figure (6.6) Excerpt 4, from Nocturne 1 op. 37

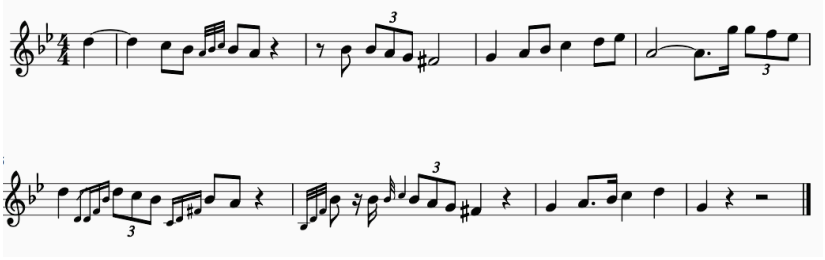
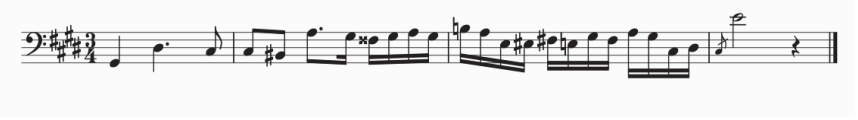


Figure (6.7) Excerpt 2, from Etude 7 op 25



7

RECAPITULATION

The goal of my doctoral research was to better understanding the production and perception of idiosyncratic expressiveness in music performance. With this aim, the research comprehended two main lines of study: The principal line of study was investigating how performers' expressiveness is constrained by their own idiosyncratic style and by the score, and which computational models would be more suitable for capturing such constrains. Furthermore, I used different machine learning methods with the aim of finding possible interactions between expressive tempo, loudness and different score features. A second line of study within the doctoral research focused on better understanding the role of tempo and loudness in the recognition of performers.

Despite the abundant literature on the use of loudness, timing or tempo in expressive performance analysis and modeling, few studies have been published on how the interactions between them and the score relate to the individuality of performers. This dissertation is aimed as an addition to the existent literature in this area of study.

In addition to showing that systematic and quantifiable individual differences can be found in the tempo and loudness expressiveness exercised by performers and how these are partly constrained by the score, this thesis also aims to serve as a guide to systematic musicologists who may want to carry research in performance expressiveness by using machine learning models.

The contributions of this dissertation can be summarized as follows:

- Chapter 1 introduces the subfield of expressive performance modeling under the umbrella of systematic musicology. With this purpose, I outline alternative definitions of expressiveness as well as the study framework which motivates the research presented in the dissertation.

- Chapter 2 presents a proof of concept study showing insights in the expressiveness consistency of timing and loudness of individual performers. I illustrated the approach using different analytical and visualization techniques to show the expressiveness consistency across the different recordings of the same piece and performers across different years. The insights obtained suggested that performers are more consistent across performances in their use of timing than in their use of loudness. Further analysis suggests that timing in phrasing phrasing was probably more strongly conditioned by contextual score features such as melody and harmonic key.

- Chapter 3 offers a review and synthesis on the field of machine learning and computational modeling with applications to music performance analysis and modeling. As such, it introduces several key concepts and motivates the use of machine learning to analyze and model individual expressiveness.

- Chapter 4 contains an study concerning the analysis and modeling of expressive tempo and loudness (as either independent or interacting features) at score markings within the frame of idiosyncratic expressiveness. The results show that both tempo and loudness were better predicted by the models used when these were trained on several (other) performers playing a given piece than when the models were trained per performer playing various other pieces.

  The results also indicate that both tempo and loudness were usually better predicted at score markings when including tempo or loudness values preceeding the marking at which these are predicted. Yet, it was also observed that tempo and loudness did not - in most cases - seem to interact within indications in the score. This invalidates the hypothesis that tempo and loudness interact at score markings, at least within the context of experiments presented.

  These findings suggest that, within the context defined by the dataset, performers plan and direct their expressive gestures on tempo and loudness towards the score marking, possibly as a structural phrasing inflexion. This may explain why the models are better when predicting the same piece, even though the expressive strategies might differ among performers. Alternatively, within a given musical culture, variation in expressiveness may be more closely related to the piece in question than how a performer approaches a broader repertoire. This is an interpretation we can not preclude with the dataset used.

  Future experiments relating score markings and expressive performance based features could address these interactions more robustly using sequential models and appropriate combinations of features.

- Chapter 5 studies whether individual performers' use of tempo and loudness (and the interactions between them), is constrained by meter and melodic rhythm specific to the score. In addition, using the same dataset as in the previous chapter, an assessment as to whether the idiosyncratic use of expressive tempo or loudness is better predicted by performer-based models, as opposed to piece-based models, is offered. Furthermore, this chapter motivates the use of sequential models and, in particular, Long Short-Term Memory networks, to better approximate the nature of music making and listening. To the best of my knowledge, this is the first study of these characteristics.

  The results of these experiments indicate that score-based models lead to better predictions compared with performer-based models. This suggesting that musical structure inherent to the score, constrains expressiveness of the performers more strongly than stylistic idiosyncratic expressiveness of that performer when playing other pieces.

  An alternative interpretation of these results could be that, in addition to the score constrains, other cultural factors such as those shared agreements on the expressiveness exercised could play a role in constraining the expressiveness of other performers. However, testing this interpretation is unfeasible, as we cannot

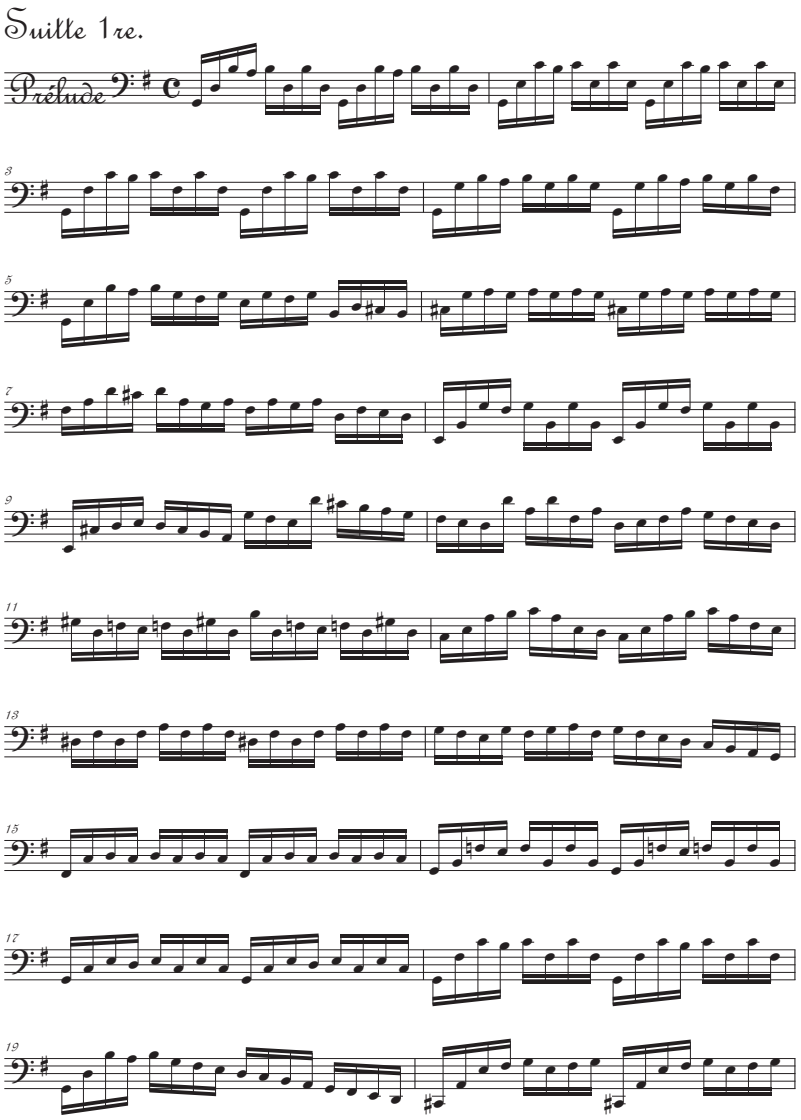
readily obtain data to reflect the possible exposure to performance practice of all individual performers sampled in this study.

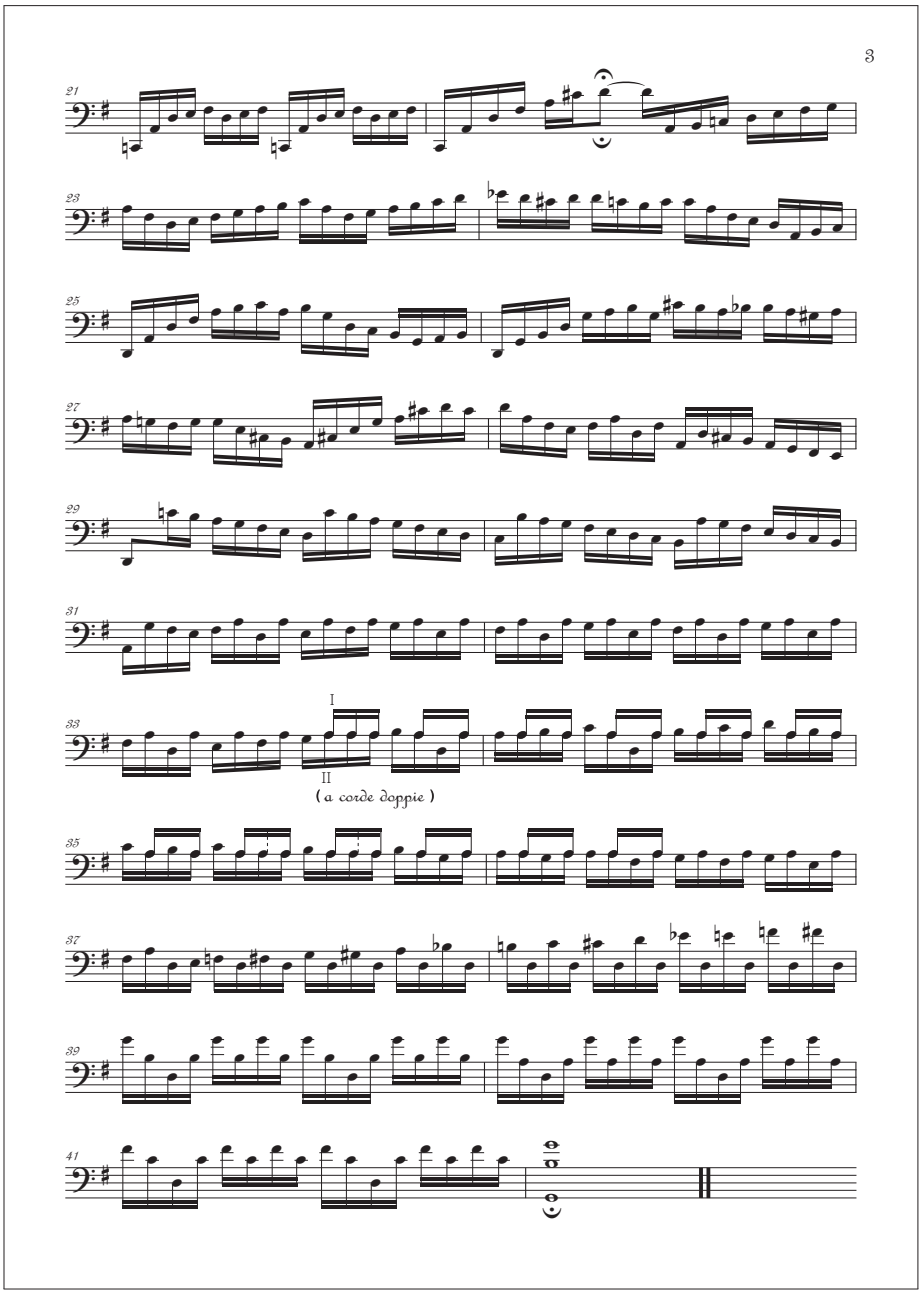
A major insight obtained from this experiment was that the interactions between tempo, loudness and meter allow for predicting both tempo or loudness better than when using tempo and loudness as combined features. However, when meter was not accounted for, interactions between tempo and loudness were not discernible. Such results suggest that metrical structure constrains the expressive gestures of loudness and tempo and how these two aspects of performance interact. Such interactions seem to vary depending upon the piece performed on the long and short-term expressive phrasing exercised by performers. These results may furthermore elucidate how a listeners' expectations of tempo and loudness can also be constrained by the salient structure of the score, beat position and meter perception.

The experiment outcomes in this chapter validated one of the initial hypothesis regarding dependencies between performance features and score features. The reason why the interactions between tempo use and rhythm are not well captured is due to the fact that these features often operate on different temporal scales (beat level vs. note level).

The dataset used in Chapters 4 and 5 was chosen with the motivation of having a balanced comparison between performers, since all of the models are based on the same repertoire for each performer. Future research could contemplate the verification of the results obtained on the performer based models when having a larger dataset. Nonetheless, the current study represents to the best of my knowledge the largest using such a balanced data set.

- Chapter 6 includes a perceptual experiment investigating the role of expressive tempo and loudness (as independent or interacting features) when discriminating between two different performers. While the results obtained invalidated the initial hypothesis, the study provided insights on possible confounding factors. Foremost among these is the challenge to control for expressiveness on timing, tempo and loudness without making the stimuli sound too "unnatural". Furthermore, the experiment offered preliminary insights into the possible cognitive limits and sensitivity of the participants within the one-shot recognition task presented during the experiment.




3
21
23
25
27
29
31
33
I
II
(a corde doppie)
35
37
39
41

b

APPENDIX 2

Application of Hidden Markov Models to music performance style classification via timing and loudness features

C. Vaquero Patricio
Universiteit van Amsterdam,
Music Cognition Group
e-mail: c.vaqueropatricio@uva.nl

E.Chew
Queen Mary University of London,
Centre for Digital Music
e-mail: elaine.chew@qmul.ac.uk

Hidden Markov Models (HMMs) have been widely used in modeling time series data and, especially, in speech recognition [5]. In expressive music performance modeling, HMMs have served varied purposes such as generating expressive performances [2] and score following through machine listening [6]. We explore a novel use of HMMs for stylistic classification of performances of the same music piece. In Western art music it is common to distinguish performers and performance styles by their historically informed approaches, or lack thereof, as well as by the characteristics of the instrument being played. In Baroque music performance, we may differentiate recordings into those played on a Baroque period instrument (tuning around 415 Hz for concert A) and on a modern one (around 440 Hz). Aesthetic performing choices are often based on the performers' artistic and musicological approaches as well their previous exposure to other performances. These choices are commonly manifested in performers' use loudness and timing patterns, which are the elements of musical phrasing. Due to the availability of recordings and quantitative methodologies we may differentiate between the use of these expressive features between modern and Baroque performances. The questions we ask include: How well-defined is the division of Baroque and modern performances? Can quantitative methods capture if a performer plays in one style vs. the other? And, ultimately, can listeners perceive these differences? In this study we focus on quantifying performance style differences using HMMs with ground truth being the aforementioned division between Baroque and modern performances commonly found in Classical music culture.

The method is tested on performances of the first half (343 notes in total) of the Prelude of Bach's Suite in G Major (BWV1007). The excerpt was chosen for its isochronous rhythm and the lack of tempo or dynamic markings in the score. A dataset of twelve recordings performed by Anner Bylsma, Jaap Ter Linden and Peter Wispelwey on Baroque cello and by Mstislav Rostropovich, Jean-Guihen Queyras and Yo-Yo Ma on modern cello were collected. These performers were chosen because they have each recorded the piece at two different periods of their careers. Onsets for every note were manually annotated (twice and averaged) using Sonic Visualiser [7]. The timing ($log_2(tempo)$) is calculated per note duration and smoothed over each beat (four consecutive sixteenth notes) [1]. The loudness values (in sones) for each note played were extracted using a

Short Time-Varying Loudness model proposed by Moore [3]. The set of features used therefore consist of the timing and loudness values separately, the timing and loudness combined, and the 1st and 2nd derivatives of each of these features, in order to study the "acceleration" (phrasing trajectory in our case) and jerk (rate of change of acceleration), respectively.

With the goal of automatically classifying the performances we train our models using the HMM implementation on scikit-learn [4]. We carry different experiments building classifiers models for baroque and modern styles according to the possible combinations of features. Our aim is to model the transition of expressive trends as possible expressive trajectories. We initially represent these trajectories as two different states within the hidden layer. We run experiments gradually incrementing the number of hidden states to 15. Once we have trained our baroque and modern models we test each performance by leave-one-run-out cross-validation and obtain the likelihood of each performance belonging to one or the other style model. We validate these models based on a randomized permutation test with 1000 iterations.

The randomized permutation test results show that the HMM models obtained are significantly different ($p < 0.05$) when the included features are loudness and the first and second derivatives of loudness (with 2 and 3 hidden states) or timing and the first derivative of timing (with 2 hidden states). Models based on loudness and timing combined did not discriminate significantly. A main challenge faced in this study is the limited training data available. Further research should be carried with a larger dataset in order to validate the strength of this methodology. We show how by modeling performance styles using HMMs we may quantify and compare how well a performance fits one aesthetic style vs. another using different combinations of features, and how distant performances are from their respective styles. Further research could use this methodology to analyze the evolution of aesthetic trends by sequentially adding different performances to the models.

References

[1] E. Chew and C. Callender. Conceptual and experiential representations of tempo: Effects on expressive performance comparisons. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 7937 LNAI, pages 76–87. 2013.

[2] G. Grindlay and D. Helmbold. Modeling, analyzing, and synthesizing expressive piano performance with graphical models. *Machine Learning*, 65:361–387, 2006.

[3] B. Moore, Brian R. Glasberg, and T. Baer. A model for the prediction of thresholds, loudness, and partial loudness. *Journal of the Audio Engineering Society*, 45(4):224–240, 1997.

[4] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in {P}ython. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[5] L.R Rabiner and B.H. Juang. An introduction to hidden Markov models. *IEEE ASSP Magazine*, 3(January):4–16, 1986.

[6] C. Raphael. Music Plus One and Machine Learning. In *International Conference on Machine Learning*, pages 21–28, 2010.

[7] C. Cannam, C. Landone, and M. Sandler. Sonic visualiser: An open source application for viewing, analysing, and annotating music audio files. In *Proceedings of the ACM Multimedia 2010 International Conference*, pages 1467–1468, Firenze, Italy, October 2010.

C

APPENDIX 3: USING RHYTHMIC CATEGORIES TO GENERATE EXPRESSIVENESS

C.1 INTRODUCTION

As it has been discussed in this dissertation, research in music performance has shown that even when listeners require no explicit training to perceive expressive timing (Palmer, 1997), exposure (Honing & Ladinig, 2009), memory (Snyder, 2000) and expectation (Huron, 2006) play a fundamental role when recognizing nuances in music timing. Our expectations over the expressiveness exercised when listening to music will be partially determined by our previous exposure to music and, in particular, to the specific style having been exposed to. By better understanding how our cognitive processes relate to the formation of the expectations, we may be able to predict the listeners' response to music, and use this knowledge with a creative purpose in automatic music generation.

One of the challenges faced by computational models of music performance is that the renditions produced by them may often sound either too "simple" or too "unnatural". With the goal of overcoming these limitations, much of the research in this field has been evaluated in the Rencon competition. In which blinded listeners must judge whether a performance is played by an artificial intelligence system or by a human. One of the original goals of this competition was passing a Touring test for musical expression (Hiraga, Bresin, Hirata & Katayose, 2004) and contributing like this to the understanding of some of the mechanisms involved in the perception of music expressiveness. With such goal, many of the proposals till nowadays have focused on solutions constrained by the specific musical expressive style and/ or musical period for which they are developed or over which the algorithms have been trained. This specificity allows accounting for structure and context in the musical discourse of the piece being rendered.

A musical generation system that is entirely style agnostic is often not desired and very complex to be evaluated. Nonetheless, there are several expressive elements of a performance, such as timing, dynamics or vibrato, that are at times desired to elicit a certain amount of tension and awareness during the listening process through a non-linear behavior. As it is observed by Honing (2006a), performers make use of non-linear models to convey expressiveness using different sorts of *ritardandi*. In fact, these non-linear models can be seen then as a communicative resource to refer to a listeners memory and expectations as well as to capture their attention. This non linear behavior, and the challenges it presents when modeling performer specific expressiveness, is commonly found in performers and it has also been observed through the results obtained in Chapters 4 and 5.

Thus, when defining then a model of expressive performance we could reckon incorporating the ability of the model to produce non linear variations within the deviations

defined by the perceptual constraints. This versatility in expressive productions of the model is necessary not only to attend the non linearity in performance, but also to respond to our relation to expectancy and uncertainty as listeners. Having a varied and large corpus combined with approaches such as the one proposed in Chapter 5 may allow to adjust the algorithms stylistic biases and, therefore, generate renditions that may elicit surprise at the most suitable structural points of the piece.

If we aim for an intrinsic definition of expressiveness, we should as well make use of the cognitive constraints in listeners when effectively recognizing a rhythm and without having references to an external score. In the scope of making use of such cognitive constraints in the domain of timing, herewith I suggest making use of perceptual rhythm constraints shared by performer and listener to generate expressive rhythms delimited by perceptual rhythm categories.

As an approach to model expressive performances within different rhythm patterns and mental representations, I propose combining symbolic and graphic representations of rhythm spaces with Lindenmayer systems and logo-turtle abstractions. The model proposed, can be used as an exploratory tool of expressive timing for computational creativity music generation.

C.2 DEFINITION AND VISUALIZATION OF RHYTHMIC CATEGORIES

In the domain of rhythm, expressive timing is defined by the deviations, or nuances, that a performer may introduce in contrast to a metronomic interpretation of a rhythm. The *"ability of listeners to distill a discrete, symbolic rhythmic pattern from a series of continuous intervals"* (Honing, 2013) requires understanding whether the perception of rhythm is categorical or not. Rhythmic perceptual categories then can be understood as mental clumps by which listeners can mentally relate expressive timing to a rhythmic pattern after having effectively detected it (Honing, 2013).

In order to address how a continuous domain such as time is perceived and categorized to be represented symbolically as a rhythm, several behavioral experiments have been done. The two main hypothesis can be resumed in the studies by Clarke (1987) and Desain and Honing (2003).

- Clarke (1987) did two experiments to study the hypothesis that listeners judge deviations as an element out of the categorical domain. In these experiments, the participants were presented with stimuli containing a rhythm in duple or triple meter, having the last two notes of the stimuli being varied with ratios of 1:1 or 1:2. In one of the experiments, the participants then would accordingly have discriminated between pairs of stimuli as being the same or different. In the other experiment, the participants (with a musical training) would have to categorize the rhythm as belonging to each of the ratio categories. In both the discrimination as in the categorization experiments, the results showed to match in the boundary defined. That is, having a clear indication of where a rhythm stops being perceived as binary and starts being perceived as ternary (and vice versa). From these experiments, it was concluded that rhythm is not

perceived on a continuous scale but as rhythmic categories that function as a reference relative to which the deviations in timing can be appreciated.

- Desain and Honing (2003) did an extensive empirical study using a large set of temporal patterns as stimuli to musically trained participants. In the experiment, participants were asked to note down the rhythm heard as in a solfege dictation. By giving an identification task using 66 four-note rhythms with a total duration of 1 second, the authors were able to sample and obtain the perceived rhythmic categories from the whole rhythm space as defined by three inter-onset intervals (IOIs). To visualize the rhythm space, Desain and Honing made use of a chronotopological map. On it, the IOIs are represented in a three dimensional rhythm space in which, each side of the rectangular triangle, represents an IOI. Therefore, all the rhythms presented to the participants consisted on four notes and their IOIs according to a metronomical measure to determine whether the rhythm is represented closer or further away from the centre (or origin) of the triangle. Having a four note isochronous rhythm (all IOIs with the same duration) would then lead, according to a metronomical interpretation, to a representation of itself in the exact origin of the chronotopological map. By having this representation, the location of the notated rhythms would coincide at some positions within the triangle if all participants would notate the rhythm heard the same. Like this, Desain and Honing (2003) were able to extract different perceptual agreements and define topos in the categorization of rhythm.

In order to cover a broader range of tempi, the stimuli chosen for their experiment were presented at 40, 60 and 90 beats per minute. As it is shown in their results, tempo may affect the representation of the rhythmic categories, varying the shape and size, depending on the combination of them (e.g. the 40 BPM and duple rhythm category will be different than the 40 BPM and triple).

Desain and Honing (2003) also showed how, in most categories, the most frequently identified pattern (marked as modal in Figure c.1) is not aligned with the metronomical interpretation of the same rhythmic pattern. This suggests that deviations within a category are not confirming to Clarke's definition of timing being deviations from integer-related durations as notated in a score. Instead, it suggests that the most commonly perceived rendition of a rhythm (modal) is actually not integer-related, but contains a timing pattern (roughly a slight speeding up and slowing down), a rhythmic pattern that seems a more appropriate reference to use than the metronomical version. The latter, in fact, might well be perceived as expressive (Honing, 2013).

The results shown in Desain and Honing (2003) may explain why traditional software tools in which expressive timing is treated as a result of, e.g. a rounding-off algorithm, are often limited in expression and easily differentiated from non-machine generated rhythm (Widmer & Goebl, 2004).

Following on the categorization and representation of rhythms, Bååth, Lagerstedt and Gärdenfors (2010) showed how the perceptual rhythm categories could be effectively approximated based on the resonance theory of rhythm perception. Using around 400 Hopf oscillators and the implementation by Large (2010), Bååth effectively

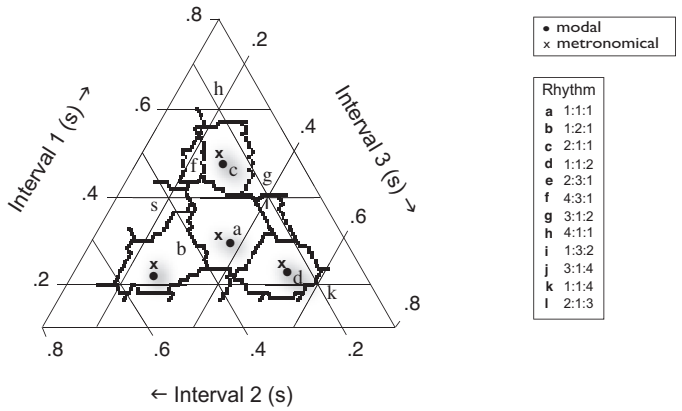


Figure (c.1) Rhythmic categories, demarcated by black lines in a chronotopological map. Each point in the map is a rhythm of four onsets, i.e. three IOIs with a total duration of one second; Perceived (modal) and integer related (metronomical) centroids are marked by dots and crosses, respectively; Letters refer to rhythmic categories annotated in the legenda.

reproduced many of the chronotopological categories obtained on Desain and Honing (2003).

Finally, in a recent study by Jacoby and Mcdermott (2017), the universal cross-cultural properties of rhythm have been questioned supporting instead the hypothesis that rhythm categories might emerge as a combination of a cultural bias combined with some biological universals. In their study, they tested a group of native Amazonians and a group of North American listeners. The results in their study indicate that while both groups share those perceptual categories in rhythms containing integer ratios, the native Amazonians present some rhythm categories specifics on their group and suggested in some specific timing deviations.

Assuming that the rhythm categories found by Desain and Honing (2003) and reproduced by Bååth et al. (2010) are, if not universal, representative of a broad cultural setting, we would be able to use this information to effectively generate "perceptually constrained" rhythms for such setting. This is the main motivation of the proposal and hypothesis presented in this appendix. In the next section, I will outline how by using the topos in the chronotopological map, in combination with a L-systems model, we may be able to generate perceptually constrained expressive timing.

C.3 LINDENMAYER SYSTEMS

Finding a relation between formal grammars and music syntax has been researched since the publication of the General Theory of Tonal Music (Lerdahl & Jackendoff, 1983), a theory inspired by Chomsky's formalization of language (Chomsky, 1956). For a comprehensive review on some of the most notable efforts in relating the cognition of language and music, the reader is referred to Patel (2008). One of the main advantages

of Chomsky's formalization is that its approach to the grammar is semantically agnostic and its production (development of its symbols) is sequential. In it, a generative grammar G is defined by the 4-tuple:

$$G = (N, S, \omega, p) \tag{21}$$

- N being a finite of nonterminal symbols (or variables) that can be replaced.
- S being a set of terminal symbols (constants) that is disjoint from N.
- ω being the initial axiom, is a string of symbols from N that defines the initial state of the system.
- p being a set of production rules that define how variables can be replaced by variables and/or constants having the axiom as the initial state and applying the productions in iterations.

Lindenmayer (1968) proposed a mathematical formalism originally conceived to model cell development and plant growth in which its symbolic structure develops in time. Being a semantically agnostic algorithm, this approach has been applied in many different fields such as artificial life, architecture, data compression or music.

Lindenmayer's systems (L-systems) permit the development of a structure of any kind being represented by a string of symbols within an alphabet. This development is done in a declarative manner according to a set of rules (defined or inferred), each of them taking care of a separate step of the process. A great difference from Chomsky's approach is that the production of L-systems in comparison to Chomsky's grammar is parallel, not sequential; consequently, a word (representing *e.g.* a state or action) might have all letters replaced at once (Prusinkiewicz & Lindemayer, 1990).

In musical L-systems we can differentiate among three different steps or types of rules (Manousakis, 2006):

- Production rules: The symbols are to be interpreted according to production rules that determine the structural development of the model. The production rules are the key to the development of the string and, thus, the richness and variety of the system depends on them. Choosing, therefore, a set of rules or another will define the type and output of the L-system being used.
- Decomposition rules: Having discrete time steps, each step applies the production rules transforming the string of symbols into a new one. The decomposition rules allow then to have a parallel development of the production rules and apply them to the resulting string of each of the derivations. Decomposition rules are always context-free and, effectively, Chomsky productions.
- Interpretation rules: On each derivation some interpretation rules must be applied to be able to parse and translate the string output to the desired field and parameter being studied. This parsing and translation will be done, as in

context-free Chomsky productions, recursively after each derivation. As we will see, a great benefit for the expressiveness generative model will focus on these interpretation rules; the mapping of them is what will allow the versatility and richness, at the same time than allowing simplicity, of the system.

As an example, we can study a simple implementation of a Fibonacci sequence (or *algae* growth model) using context free L-Systems.

- Axiom:
  ω : A
- Production rules:
  p1 : "A" $\rightarrow$ "B",
  p2 : "B" $\rightarrow$ "AB"
- Decomposition rules: We will obtain the following result during six derivations:
  n = 0: B
  n = 1: AB
  n = 2: BAB
  n = 3: ABBAB
  n = 4: BABABBAB
  n = 5: ABBABBABABBAB

L-systems are categorized depending on the type of grammar being used. These can be classified according to the appliance of the production rules, but each of the grammars can be combined with others. According to Manousakis (2006) grammars in musical L-systems can be: *context-free* (OL systems), *context-sensitive* (IL systems), *deterministic* (DL systems), *non-deterministic* (stochastic) *NDL*, *bracketed*, *propagative* (PL systems), *non-propagative*, *with tables* (TL system), parametric *parametric* or *with extensions* (EL system).

Originally conceived as a formal theory of development, L-systems were extended to describe higher plants and complex branching structure by Prusinkiewicz and Lindemayer (1990), who also worked on the implementation of them to fractal and living organisms graphical representations. Depending on the dimensions treated, the complexity of the rules and grammar models being used, L-systems might require different graphical abstractions.

Prusinkiewicz's approach was based on a graphical interpretation of L-systems built on top of the logo-style turtle (Francis, Abelson & DiSessa, 1983). The turtle movement in a two dimensions map interpretation consists on a triplet (x, y, α) that includes the Cartesian coordinates (x, y) and the angle (α) that directs the facing head of it. Once the step size (d) and the angle (α) are given, the turtle is directed by following rules such as:

- F : Move forward and draw a line. The line should be drawn between (x, y) and (x', y'). (x', y') is defined then by: $x' = x + d\cos\alpha$ and $y' = y + d\sin\alpha$
- f : Move forward without drawing a line
- + : Turn left by angle δ. The turtle should point then according to $(x, y, \alpha + \delta)$
- - : Turn right by angle δ. The turtle should point then according to $(x, y, \alpha - \delta)$

C.4 USING L-SYSTEMS TO GENERATE EXPRESSIVENESS

Prusinkiewicz (1986) proposed a musical application of L-systems and, since then, several related approaches have been proposed with different purposes such as *e.g.* compose music (Supper, 2001), generate real time evolving audio synthesis and music structures (Manousakis, 2009; Mason & Saffle, 1994) or parsing music structure from scores (Nevill-manning & Witten, 1997). However, to our knowledge, L-systems have not been approached yet in combination with perceptual constraints.

The main advantage of incorporating L-systems into a perceptual model of expressiveness is that since it's semantic relation to the modeled structure is symbolic, there is no topological similarity or contiguity between the sign and the signifier, but only a conventional arbitrary link (Manousakis, 2006). The associations to the rules and symbols can, hence, be modeled according to any of the methods mentioned. In addition, the fact that in L-systems the development of its symbols (production) is parallel, makes it very convenient to generate music expressivity due to the versatility in the production or mapping levels within different expressive categories, in any structural or generative level.

C.4.1 *Implementation*

The purpose of this implementation is verifying that the hypothesis proposed can be empirically validated as an exploratory framework and a computational model to produce expressive performance (or musical composition). Due to the versatility of the different steps of L-systems several approaches can be further developed. In the following subsections the current proposed implementation is presented within the different phases necessary to attend a possible generative system:

C.4.1.1 *Geometrical approximation*

A first challenge when using data from perceptual rhythm categories, is how to approach the complex geometrical shapes of each category. A simple solution can be the approximation of the complex geometrical shapes to congruent ones, in which all of them have the same shape and different dimensions. Due to the shape of the rhythm categories the simplest geometrical forms that we can visually approximate to them are the circumference or the ellipse (a triangle has also been considered). Since we aim

to hoard as much space of each category as possible, an ellipse seems like the best approximation. Obtaining measurements manually from the graphical representations of the categories presented in Desain and Honing (2003), we can define the position in the geometrical space as well as dimensions (axis lengths) and angle inclination of each of the ellipses being used. The result of this hand-aligned approximation to ellipses for all rhythms with a duration of one second (cf. 60 BPM) can be observed in Figure c.2. In order to simplify our proof of concept study, for the current we will focus solely on the temporal aspects of rhythm (e.g., the rhythmic pattern "a" in Figure c.2).

c.4.1.2 *Object Mapping*

Each rhythmic category can be represented by a letter of the L-system dictionary and this abstraction can be used simultaneously with different production rules attending, for instance, other expressive aspects in addition to rhythm. From a generative perspective of a compositional system, once we have mapped the different rhythm categories we can define some production rules to alternate (or "jump") from a rhythm category to another one rendering different rhythmic patterns.

c.4.1.3 *Modular Mapping*

L-systems allow approaching expressiveness from the richness of a generative symbolic interpretation that the organicity of its parallel development provides. Consequently, once each of the categories has been assigned to a dictionary, we can map the movements to be produced by the turtle to other letters of the dictionary. The modular mapping initially proposed consists of three different rules assigned to the following movements. Lets illustrate this with an example:

- Alphabet:
  $V : A$
- Production rules:
  $p1 : A \rightarrow B$,
  $p2 : B \rightarrow AB$
- Axiom:
  $\omega : A$
- Decomposition rules for three derivations:
  $n = 0 : B$
  $n = 1 : AB$
  $n = 2 : BAB$
  $n = 3 : ABBAB$
- Interpretation rules:
  A: move forward with a distance x
  B: Turn right by a random angle δ between $0°$ and $360°$

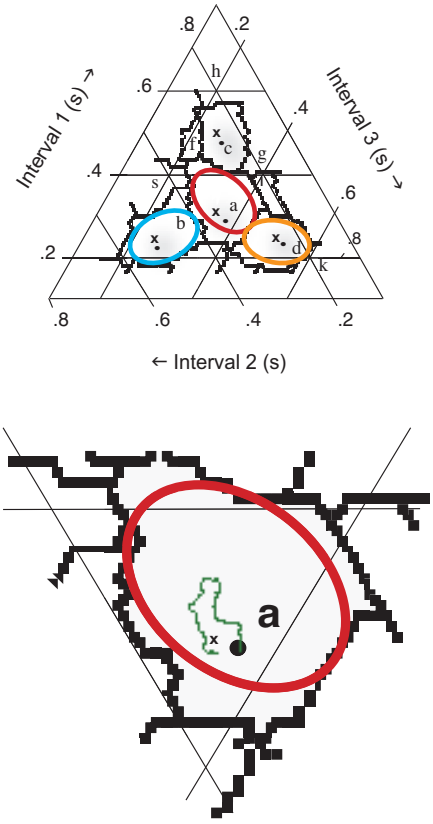


Figure (c.2) Category "a" (The full rhythm map (top panel) and zoomed-in version (bottom panel) showing category "a" with an elliptical approximation of its perceptual boundaries. The yellow line marks a potential path on that map using an L-system

According to the example presented, the represented turtle will advance three steps and move according to a random angle four times, as a result of the four derivations done. In order to warranty that the turtle will respect the size of the category approximation (ellipse in this case) a normalization of the distance from the center of the ellipse to the perimeter of it is being applied. The distance advanced on each step might be determined by the degree of expressive deviation we might want our system to produce. Accordingly, the production possibilities are greatly determined by the rules and derivations and the expressiveness and musical style coherence, will depend on the interpretation rules being used.

Expressiveness should not be modeled as a random walk across categories. The rules assignment (or inference) is, thus, a prudent choice of the model to be applied. That is, allowing the system to *e.g.* have large turtle movements within a category would only be sensible if the music to be performed would intentionally have much random rubato.

In Figure c.2 it is shown a reduction to ellipses and an example of a hypothetical trajectory of expressiveness generation (by using the turtle) through different points of a rhythmic category.

c.4.2 *Evaluation*

As explained in c.2, the perceptual categories in which the expressive timing is generated were obtained through empirical experiments. From this perspective we have a ground to understand that the material over which the expressiveness will be generated should be perceptually valid. Yet, since the use of L-systems can vary much depending on the manner on which the different rules and alphabets are being developed, in order to validate the hypothesis presented in this appendix further experiments with human listeners should be carried for each of the alternative systems being developed.

c.4.3 *Practical and conceptual challenges in the implementation of the model proposed*

Some pitfalls from turning a reductionist approach into a microworld have been previously addressed by Honing (1993). As it has been mentioned, in this microworld abstraction of music and rhythm, the formulation of the rules and assignment of their production properties will need to attend a perceptual scenario also coherent with the music theory grounds and style specifics that our generative model is dealing with.

In the current microworld, two main challenges have yet to be addressed and tested to arrive at an exploratory model of expressive timing.

- The first main issue to be explored is that the study presented on Desain and Honing (1989) only contains the rhythm categories of two duple and triple meters and within three different tempo values. This is challenging, since it is questionable whether tempo and perceptual categories could be scaled proportionally while keeping a centroid relation derived from a morphological inference between categories. Having the results of centroids and categories for the BPM values of 40, 60 and 90, we could define an optimization of the model to infer shapes and sizes of rhythmic categories belonging to other BPMs.

  However, any kind of geometrical inference or interpolation between chronotopos at different tempi may be misleading. According to Sadakata, Desain and Honing (2006) the hypothesis is that while score intervals scale proportionally with global tempo, the deviation of the performed onsets with respect to the score is invariant of that. A possible solution to complete the data of other tempo values on the rhythm categories, in case empirical data is not available, could as well approximate the different perceptual categories by using the approach researched by Bååth et al. (2010).

- The second issue to be addressed is concerned with how to correlate positions of the turtle movement within the rhythm perceptual spaces being explored. Solving

this issue is essential when applying this model in a real scenario, since music often has different rhythmic patterns to be alternated and combined, and the expressive deviations of one rhythmic category should be consistent with the deviations of the category following or preceding it. This can be done by locating these deviations according to the neighboring of their categories. The trajectory of the turtle, defined also by the length of the step, should be coherent with the rhythmic category in which it is being developed. Even when expressive timing is often oscillating between interpretations within an average of 50 to 100 ms, there is evidence that timing varies depending on tempo (Desain & Honing, 1994). Having then a bigger or smaller definition of the path of the turtle might mainly make sense to be able to define concentrically its movements around the centroid, to avoid great deviations at the same time that we are aiming to achieve variation.

Scaling the modal centroid to a fitted or approximated area of the category, will allow the turtle to jump in a continuous music line from a category to another one ("mirroring" these positions), being coherent with the degree of expressiveness among them; also when approaching expressiveness complexity in musical passages in which variation is needed. Considering the continuity and progression of time in music being produced by the model, we can establish mirror positions of the turtle within different categories that would follow the turtle positions within the ellipse, depending on the predetermined context (music style, performer). In the case of representing scored music, it would be determined by a score follower that may preselect the category and placement of the turtle before jumping between categories (different rhythmical patterns).

In Desain and Honing (1992), the boundaries of each rhythm category represent the amount of entropy in relation to the modal centroid, in matters of the difficulty to tag a rhythm and the ambiguity implications of the boundaries of each category. We can use the same idea, in this case, as the amount of degraded expressiveness before the boundary of a category is reached and might be confused with another category due to the ambiguity present on this boundary. Like this, we can relate the position of the turtle between categories to the entropy of the location in which it is placed within the category. This scaling implies the need to discretizing the category being represented, but from an implementation perspective this is a straightforward relation to the different grid samples established in the categories. Using entropy for each category as a measure to be able to compare categories seems as an optimal solution.

To simplify the definition of the areas among rhythm spaces we make use of a geometrical simplification of the categories, such as the one previously explained. Another solution would be using a curve fitting function that would approximate the differences between each of the vectors that conform the different categories.

A more complete study of the relation of centroids to absolute tempos would be to fit a Bayesian model to the data, separating the probability of identifying a performance as a certain rhythm (score), into the prior distributions of score rhythms, and a Gaussian (normal) distribution of a performance given a score. However, the last distri-

bution is expected to be off-center by an amount which is independent of global tempo (Sadakata et al., 2006).

Finally, moving through each of the rhythmic categories (*e.g.* using just the first three IOI in a 4/4 bar) implies the necessity of defining a model to estimate the duration of a fourth IOI. As of in the current approach, its duration is constrained by the duration of the bar defined by the BPM.

Being consistent with our definition of expressiveness we must reckon that listeners exposure to music can be represented by using a scheme of weighting in the bars such as the one proposed by Pearce and Wiggins (2004). Then, we can infer a distribution of the weights over the first three IOI, *e.g.*, using some type of probabilistic model to estimate the duration of the 4th IOI. A parsing method to extract the weights according to a structure could work efficiently by associating it to a grammar of symbols that could generate different degrees of expression. A technique that gives us a possibility to do this parsing using L-systems is Sequiturs algorithm (Nevill-manning & Witten, 1997).

C.5 CONCLUSIONS

Despite much research has been done in the field of expressiveness generation little attention has been paid to the possibility of using data from perceptual experiments to generate expressiveness. In order to embrace the necessary versatility to produce expressiveness in music, in this appendix I have presented a novel approach to modeling expressive timing performance by combining cognitive symbolic and graphic representations of rhythm spaces with Lindenmayer systems.

In c.1 an approach to understanding expressiveness as deviations within different perceptual categories has been presented. In c.2 the study done by Desain and Honing (2003) to collect the data empirically and the formation of the rhythmic categories has been presented. c.3 introduced a resume on what Lindenmayer systems are and what the state of the art on musical applications is. In addition, it has been described how by means of a symbolic abstraction we can construct rules, dictionaries or axioms using different L-systems types depending on the requirements of the music that wants to be generated. In c.4 a preliminary implementation of the system has been presented together with a solution for further validation of the system being implemented.

The current proposal requires much further development and evaluation. It remains a challenge to scale from a microworld approach (as was presented in this work) to a more realistic model of expressive performance that will attend to different music and performance styles. The initial steps done on this expressive cognitive model seem promising to develop better performance systems as well as to understand the cognitive aspects being involved in expressiveness perception and generation. This is a topic of future research.

REFERENCES

Askenfelt, A. & Jansson, E. V. (1988). From touch to string vibrations - the initial course of the piano tone. *Stl-Qpsr*, *29*(1), 31–109. doi: 10.1121/1.2024316

Bååth, R., Lagerstedt, E. & Gärdenfors, P. (2010). An Oscillator Model of Categorical Rhythm Perception. In M. Knauff, M. Pauen, N. Sebanz & I. Wachsmuth (Eds.), *35th annual conference of the cogni- tive science society* (pp. 1803–1808).

Bach, C. (1752). *Versuch {ü}ber die wahre Art das Clavier zu spielen*.

Bailes, F., Dean, R. T. & Pearce, M. T. (2013, jan). Music cognition as mental time travel. *Scientific reports*, *3*, 2690. doi: 10.1038/srep02690

Bates, D., Maechler, M. & Bolker, B. (2012). *lme4: Linear mixed-effects models using S4 classes. R package version 0.999999-0.*

Benetti Jr., A. (2013). Expressividade e performance: estratégias práticas aplicadas por pianistas profissionais na preparação de repertório. *Opus*, 149–172.

Bengio, Y., Simard, P. & Frasconi, P. (1994). Learning Long Term Dependencies with Gradient Descent is Difficult. *IEEE Transactions on Neural Networks*, *5*(2), 157–166. doi: 10.1109/72.279181

Bergstra, J. & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, *13*, 281–305. doi: 10.1162/153244303322533223

Bhatara, A., Tirovolas, A. K., Duan, L. M., Levy, B. & Levitin, D. J. (2011). Perception of emotional expression in musical performance. *Journal of experimental psychology. Human perception and performance*, *37*(3), 921–934. doi: 10.1037/a0021922

Bigand, E. & Poulin-Charronnat, B. (2006). Are we "experienced listeners"? A review of the musical capacities that do not depend on formal musical training. *Cognition*, *100*(1), 100–130. doi: 10.1016/j.cognition.2005.11.007

Bishop, C. M. (2013). Model-based machine learning Author for correspondence :. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *371*(1984), 20120222.

Böck, S., Krebs, F. & Schedl, M. (2012). Evaluating the Online Capabilities of Onset Detection Methods. *ISMIR*(Ismir), 49–54.

Brattico, E. & Pearce, M. (2013). The neuroaesthetics of music. *Psychology of Aesthetics, Creativity, and the Arts*, *7*(1), 48–61. doi: 10.1037/a0031624

Bresin, R. (1998). Artificial neural networks based models for automatic performance of musical scores. *Journal of New Music Research*, *27*(3), 239–270. doi: 10.1080/09298219808570748

Butt, J. (2002). *Playing with history: the historical approach to musical performance*.

Cancino-Chacón, C., Grachten, M., Sears, D. R. W. & Widmer, G. (2017). What were you expecting? Using Expectancy Features to Predict Expressive Performances of Classical Piano Music. , 1–6.

Cemgil, A. T. & Kappen, B. (2003). Monte Carlo Methods for Tempo Tracking and Rhythm Quantization. , *18*, 45–81.

Chanan, M. (1994). *Musica practica: The social practice of Western music from Gregorian chant to postmodernism*. Verso.

Cheng, E. & Chew, E. (2008a). Quantitative Analysis of Phrasing Strategies in Expressive Performance: Computational Methods and Analysis of Performances of Unaccompanied Bach for Solo Violin. *Journal of New Music Research*, *37*(March 2015), 325–338. doi: 10.1080/09298210802711660

Cheng, E. & Chew, E. (2008b, dec). Quantitative analysis of phrasing strategies in expressive performance: computational methods and analysis of performances of unaccompanied bach for solo violin. *Journal of New Music Research*, *37*(4), 325–338. doi: 10.1080/09298210802711660

Cherla, S., Tran, S. N., Garcez, A. D. A. & Weyde, T. (2015). Discriminative learning and inference in the Recurrent Temporal RBM for melody modelling. *Proceedings of the International Joint Conference on Neural Networks*, *2015-Septe*. doi: 10.1109/IJCNN.2015.7280691

Chew, E. (2012). About Time: Strategies of Performance Revealed in Graphs By. *Visions of Research in Music Education*(1).

Chollet, F. (2015). *Keras.* \url{https://github.com/fchollet/keras}. GitHub.

Chomsky, N. (1956). Three models for the description of language. *Information Theory, IRE Transactions on*, *2*(3), 113–124. doi: 10.1109/TIT.1956.1056813

Clarke, E. (1987). Categorical rhythm perception: an ecological perspective. *Action and perception in rhythm and music*, *55*, 19–33.

Clarke, E. (1995). Expression in performance: generativity, perception and semiosis. In J. Rink (Ed.), *The practice of performance: Studies in musical interpretation* (pp. 21–54). Cambridge University Press. doi: 10.1017/CBO9780511552366.003

Clarke, E. (1999). Rhythm and Timing in Music. In D. Deutsch (Ed.), *The psychology of music* (pp. 473–500). Academic Press.

Clarke, E. (2002). Listening to performance. In J. Rink (Ed.), *Musical performance: A guide to understanding* (pp. 185 – 196). Cambridge University Press.

Cobussen, M., Frisk, H. & Weijland, B. (2010). The Field of Musical Improvisation. *Konturen*, 1–18.

Collins, N. (2005). A Comparison of Sound Onset Detection Algorithms with Emphasis on Psychoacoustically Motivated Detection. , 1–12.

Cook, N. (2013). *Beyond the score: Music as performance*. Oxford University Press.

Cox, G. (2010). On the Relationship Between Entropy and Meaning in Music : An Exploration with Recurrent Neural Networks. *Annual Conference of the Cognitive Science Society*, 429–434.

Cross, I. (2009). The evolutionary nature of musical meaning. *Musicae Scientiae*, *13*(2 Suppl), 179–200. doi: 10.1177/10298649090130020911

Cuthbert, M. S. & Ariza, C. (2010). music21: A toolkit for computer-aided musicology and symbolic music data.

Desain, P. & Honing, H. (1989). The quantization of musical time: A connectionist

approach. *Computer Music Journal*, *13*(3), 56–66.
Desain, P. & Honing, H. (1992). The quantization problem: traditional and connectionist approaches. In &. O. L. M. Balaban, K. Ebcioglu (Ed.), *Understanding music with ai: Perspectives on music cognition* (pp. 448–463). MIT Press.
Desain, P. & Honing, H. (1993). Tempo curves considered harmful. *Contemporary Music Review*, *7*(March 1968), 1–22.
Desain, P. & Honing, H. (1994). Does expressive timing in music performance scale proportionally with tempo? *Psychological Research*, 285–292.
Desain, P. & Honing, H. (2002). Rhythmic stability as explanation of category size. *Conference on Music Perception and Cognition*(1), 1–4.
Desain, P. & Honing, H. (2003). The formation of rhythmic categories and metric priming. *Perception*, *32*(3), 341–365. doi: 10.1068/p3370
Devaney, J. (2016). Inter- Versus Intra-singer Similarity and Variation in Vocal Performances. *Journal of New Music Research*, *8215*(July), 1–13. doi: 10.1080/09298215.2016.1205631
Dixon, S., Goebl, W. & Widmer, G. (2002). The Performance Worm: Real Time Visualisation of Expression based on Langner 's Tempo-Loudness Animation. *Austrian Research Institute for Artificial Intelligence*, 361–364.
Dogantan-Dack, M. (2014). Philosophical reflections on expressive music performance. *Expressiveness in music performance: Empirical approaches across styles and cultures*, 1–21.
Dolmetsch, A. (1915). The interpretation of the Music of the 17th and 18th Century revealed by contemporary evidence. , 493.
Donington, R. (1963). *The Interpretation of Early Music*. New York.
Drake, C. & Palmer, C. (1993). Accent Structures in Music Performance. *Music Perception: An Interdisciplinary Journal*, *10*(3), 343–378.
Eck, D. & Schmidhuber, J. (2002). Finding temporal structure in music: Blues improvisation with LSTM recurrent networks. *Neural Networks for Signal Processing - Proceedings of the IEEE Workshop*, *2002-Janua*, 747–756. doi: 10.1109/NNSP.2002.1030094
Eerola, T. & Toiviainen, P. (2004). *MIDI Toolbox: MATLAB Tools for Music Research.* University of Jyväskylä: Kopijyvä, Jyväskylä, Finland.
Efron, B. & Gong, G. (1983). A Leisurely Look at the Bootstrap , the Jackknife , and Cross-Validation Author (s): Bradley Efron and Gail Gong Source : The American Statistician , Vol . 37 , No . 1 (Feb ., 1983), pp . 36-48 Published by : Taylor & Francis , Ltd . on behalf of the. *The American Statistician*, *37*(1), 36–48.
Ewert, S., Müller, M. & Grosche, P. (2009). High resolution audio synchronization using chroma onset features. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 1869–1872. doi: 10.1109/ICASSP.2009.4959972
Fabian, D., Timmers, R. & Schubert, E. (2014). Expressiveness in music performance: Empirical approaches across styles and cultures. , 383.
Francis, G. K., Abelson, H. & DiSessa, A. (1983). *Turtle Geometry. The Computer as a Medium for Exploring Mathematics.* (Vol. 90) (No. 6). MIT Press. doi: 10.2307/2975591
Friberg, A., Bresin, R. & Sundberg, J. (2009, jan). Overview of the KTH rule system

for musical performance. *Advances in Cognitive Psychology*, *2*(2), 145–161. doi: 10.2478/v10053-008-0052-x

Gabrielsson, A. (1973). Adjective ratings and dimension analyses of auditory rhythm patterns. *Scandinavian Journal of Psychology*, *14*(1), 244–260.

Gabrielsson, A. (1974). Performance of rhythm patterns. *Scandinavian Journal of Psychology*, *15*(1), 63–72. doi: 10.1111/j.1467-9450.1974.tb00557.x

Gabrielsson, A. (2003, jul). Music Performance Research at the Millennium. *Psychology of music*, *31*(3), 221–272. doi: 10.1177/03057356030313002

Geoffroy-Dechaume, A. (1964). *Secrets de la musique ancienne*. Ed. du Sagittaire.

Gers, F. A., Schmidhuber, J. & Cummins, F. (2000). Learning to Forget: Continual Prediction with LSTM. *Neural Computation*, *12*(10), 2451–2471. doi: 10.1162/089976600300015015

Gingras, B. (2014). *Individuality in music performance*. doi: 10.3389/978-2-88919-307-3

Gingras, B., Asselin, P. & McAdams, S. (2013, jan). Individuality in harpsichord performance: disentangling performer- and piece-specific influences on interpretive choices. *Frontiers in psychology*, *4*(November), 895. doi: 10.3389/fpsyg.2013.00895

Gingras, B., Lagrandeur-Ponce, T., Giordano, B. L. & McAdams, S. (2011). Perceiving musical individuality: Performer identification is dependent on performer expertise and expressiveness, but not on listener expertise. *Perception*, *40*(10), 1206–1220. doi: 10.1068/p6891

Gingras, B., Pearce, M., Goodchild, M., Dean, R., Wiggins, G. & McAdams, S. (2015). Linking Melodic Expectation to Expressive Performance Timing and Perceived Musical Tension. *Journal of Experimental Psychology: Human Perception and Performance*.

Glasberg, B. R. & Moore, B. C. J. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, *47*(1-2), 103–138. doi: 10.1016/0378-5955(90)90170-T

Goebl, W. (2001). Melody lead in piano performance: expressive device or artifact? *The Journal of the Acoustical Society of America*, *110*(July 2000), 563–572. doi: 10.1121/1.1376133

Goebl, W., Dixon, S. & Poli, G. D. (2005). Sense in expressive music performance: Data acquisition, computational studies, and models. *Sound to Sense â Sense to Sound: A State of the Art in Sound and Music Computing*, 195–242.

Golos, G. S. (1960). Some Slavic Predecessors of Chopin. *The Musical Quarterly*, *46*(4), 437–447.

Goodfellow, I., Bengio, Y. & Courville, A. (2016). *Deep Learning*. MIT Press.

Gould, C. S. & Keaton, K. (2000). The Essential Role of Improvisation in Musical Performance. *The Journal of Aesthetics and Art Criticism*, *58*(2), 143–148.

Grachten, M. & Widmer, G. (2009). Who is who in the end? Recognizing pianists by their final ritardandi. In *International conference on music information retrieval* (pp. 51–56).

Grachten, M. & Widmer, G. (2012). Linear basis models for prediction and analysis of musical expression. *Journal of New Music Research*, 1–21.

Graves, A., Mohamed, A. & Hinton, G. (2013). Speech Recognition with Deep Recurrent Neural Networks. (3). doi: 10.1109/ICASSP.2013.6638947

Graves, A., Wayne, G. & Danihelka, I. (2014). Neural Turing Machines. , 1–26. doi: 10.3389/neuro.12.006.2007

Grindlay, G. & Helmbold, D. (2006). Modeling, analyzing, and synthesizing expressive piano performance with graphical models. *Machine Learning*, *65*, 361–387. doi: 10.1007/s10994-006-8751-3

Haynes, B. (2010). *The End of Early Music: A Period Performer's History of Music for the Twenty-First Century*. Oxford University Press. doi: 10.1093/acprof:oso/9780195189872.001.0001

Hiraga, R., Bresin, R., Hirata, K. & Katayose, H. (2004). Rencon 2004: Turing Test for Musical Expression. In *Proceedings of the 2004 conference on new interfaces for musical expression* (pp. 120–123). Singapore, Singapore: National University of Singapore.

Hochreiter, S. & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Honing, H. (1993, jan). A microworld approach to the formalization of musical knowledge. *Computers and the Humanities*, *27*(1), 41–47. doi: 10.1007/BF01830716

Honing, H. (2006a). Computational modeling of music cognition: A case study on model selection. *Music Perception: An Interdisciplinary Journal*, 365–376.

Honing, H. (2006b). Evidence for tempo-specific timing in music using a Web-based experimental setup. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(3), 780–786. doi: 10.1037/0096-1523.32.3.780

Honing, H. (2007). Is expressive timing relational invariant under tempo transformation? *Psychology of Music*, *35*(2), 276–285. doi: 10.1177/0305735607070380

Honing, H. (2012). Without it no music: Beat induction as a fundamental musical trait. *Annals of the New York Academy of Sciences*, *1252*(1), 85–91. doi: 10.1111/j.1749-6632.2011.06402.x

Honing, H. (2013). Structure and Interpretation of Rhythm in Music. In D. Deutsch (Ed.), *The psychology of music* (3rd ed., pp. pp. 369–404). Academic Press / Elsevier.

Honing, H. (2018). On the biological basis of musicality. *Annals of the New York Academy of Sciences*, 1–6. doi: 10.1111/nyas.13638

Honing, H. & Ladinig, O. (2009). Exposure influences expressive timing judgments in music. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(1), 281–288.

Huron, D. (2006). *Sweet anticipation: Music and the psychology of expectation* (Vol. 443). The MIT Press.

Isikhan, C. & Ozcan, G. (2008). A survey of melody extraction techniques for music information retrieval. *Proceedings of 4th Conference on Interdisciplinary Musicology*(October 2016).

Jackendoff, R. & Lerdahl, F. (2006, may). The capacity for music: what is it, and what's special about it? *Cognition*, *100*(1), 33–72. doi: 10.1016/j.cognition.2005.11.005

Jacoby, N. & Mcdermott, J. H. (2017). Integer Ratio Priors on Musical Rhythm Revealed

Cross-culturally by Iterated Reproduction Article Integer Ratio Priors on Musical Rhythm Revealed Cross-culturally by Iterated Reproduction. *Current Biology*, 1–12. doi: 10.1016/j.cub.2016.12.031

Juslin, P. & Laukka, P. (2004, sep). Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. *Journal of New Music Research*, *33*(3), 217–238. doi: 10.1080/0929821042000317813

Juslin, P. & Västfjäll, D. (2008, oct). Emotional responses to music: the need to consider underlying mechanisms. *The Behavioral and brain sciences*, *31*(5), 559–75; discussion 575–621. doi: 10.1017/S0140525X08005293

Karpathy, A., Johnson, J. & Fei-Fei, L. (2015). Visualizing and Understanding Recurrent Networks. , 1–12. doi: 10.1007/978-3-319-10590-153

Kendall, R. a. & Carterette, E. C. (1990). The Communication of Musical Expression. *Music Perception: An Interdisciplinary Journal*, *8*(2), 129–163. doi: 10.2307/40285493

Kingma, D. P. & Ba, J. L. (2015). Adam: a Method for Stochastic Optimization. *International Conference on Learning Representations 2015*, 1–15.

Koch, G. (2015). Siamese neural networks for one-shot image recognition..

Koren, R. & Gingras, B. (2011). Perceiving individuality in musical performance : Recognizing harpsichordists playing different pieces. (1987).

Koren, R. & Gingras, B. (2014, jan). Perceiving individuality in harpsichord performance. *Frontiers in psychology*, *5*(February), 141. doi: 10.3389/fpsyg.2014.00141

Kosta, K., Bandtlow, O. F. & Chew, E. (2014). Practical Implications of Dynamic Markings in the Score: is Piano always Piano? *Proceedings of Audio Engineering Society 53rd International Conference*, 1–13.

Kosta, K., Bandtlow, O. F. & Chew, E. (2018). Dynamics and relativity: Practical implications of dynamic markings in the score. *Journal of New Music Research*, 1–24. doi: 10.1080/09298215.2018.1486430

Krizhevsky, A., Sutskever, I. & Hinton, G. (2012). Imagenet. *Advances in Neural Information Processing Systems 25 (NIPS2012)*, 1–9. doi: 10.1109/5.726791

Ladinig, O. & Honing, H. (2006). The effect of exposure and expertise on timing judgments in music : Preliminary results. *Science*(May), 80–85.

Lake, B. M., Lee, C.-y., Glass, J. R. & Tenenbaum, J. B. (2014). One-shot learning of generative speech concepts. *{Proceedings of the 36th Annual Conference of the Cognitive Science Society}*, 803–808.

Lambert, A., Weyde, T. & Armstrong, N. (2014). Beyond the beat: towards metre, rhythm and melody modelling with hybrid oscillator networks. *Icmc 2014*(September 2014), 485–490. doi: 10.1111/j.1471-0528.2004.00303.x

Large, E. (1996). Modeling beat perception with a nonlinear oscillator. In *Proceedings of the eighteenth annual conference of the cognitive science society: July 12-15, 1996, university of california, san diego* (p. 420).

Large, E. (2010). *Neurodynamics of Music* (Vol. 36; M. Riess Jones, R. R. Fay & A. N. Popper, Eds.). New York, NY: Springer New York. doi: 10.1007/978-1-4419-6114-3

Large, E. & Palmer, C. (2002, jan). Perceiving temporal regularity in music. *Cognitive Science*, *26*(1), 1–37. doi: 10.1207/s15516709cog2601$_1$

Lerdahl, F. & Jackendoff, R. (1983). *A generative theory of tonal music* (Vol. 7) (No. 1). MIT Press. doi: 10.1525/mts.1985.7.1.02a00120

Lindenmayer, A. (1968). Mathematical models for cellular interaction in development, Parts I and II. *Journal of Theoretical Biology*, *18*(3), 280–315.

Lipton, Z. C., Berkowitz, J. & Elkan, C. (2015). A Critical Review of Recurrent Neural Networks for Sequence Learning. , 1–38.

Maas, A. L., Hannun, A. Y. & Ng, A. Y. (2013). Rectifier Nonlinearities Improve Neural Network Acoustic Models. *Proceedings of the 30 th International Conference on Machine Learning*, *28*, 6.

Mackay, D. J. C. (2003). *Information Theory , Inference , and Learning Algorithms*.

Madsen, S. T. & Widmer, G. (2006). *Exploring pianist performance styles with evolutionary string matching* (Vol. 15) (No. 04). doi: 10.1142/S0218213006002795

Malik, I. & Ek, C. H. (2017). Neural Translation of Musical Style.

Manousakis, S. (2006). *Musical L-systems* (Master Thesis). Koninklijk Conservatorium, Institute of Sonology, The Hague.

Manousakis, S. (2009, dec). Non-Standard Sound Synthesis with L-Systems. *Leonardo Music Journal*, *19*, 85–94. doi: 10.1162/lmj.2009.19.85

Marblestone, A. H., Wayne, G., Kording, K. P. & Scholte, H. S. (2016). Toward an Integration of Deep Learning and Neuroscience. , *10*(September), 1–41. doi: 10.3389/fncom.2016.00094

Mason, S. & Saffle, M. (1994). L-systems, melodies and musical structure. *Leonardo Music Journal*, *4*(1), 31–38.

Mazzola, G. & Zahorka, O. (1991). The Rubato Performance Workstation on Nextstep. *Mathematica*.

McComb, K., Shannon, G., Sayialel, K. N. & Moss, C. (2014). Elephants can determine ethnicity, gender, and age from acoustic cues in human voices. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(14), 5433–8. doi: 10.1073/pnas.1321543111

McCulloch, W. S. & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115–133. doi: 10.1007/BF02478259

Mellenbergh, G. J. (1989). Item bias and item response theory. *International journal of educational research*, *13*(2), 127–143.

Meyer, L. (1956). Emotion and meaning in music.

Meyer, L. (1957). Meaning in Music and Information Theory. , *4*(4), 412–424.

Miranda, E. R., Kirke, A. & Zhang, Q. (2010, mar). Artificial Evolution of Expressive Performance of Music: An Imitative Multi-Agent Systems Approach. *Computer Music Journal*, *34*(1), 80–96. doi: 10.1162/comj.2010.34.1.80

Mitchell, T. (1997). *Machine Learning*. McGraw-Hill.

Molina-Solana, M., Lluís Arcos, J. & Gomez, E. (2010). Identifying violin performers by their expressive trends. *Intelligent Data Analysis*, *14*(5), 555–571. doi: 10.3233/IDA-2010-0439

Mozart, L. (1756). *A treatise on the fundamental principles of violin playing* (Vol. 6). London,

Oxford U. P, 1951. Oxford University Press.

Mozer, M. C. (2007). Neural Net architecture for temporal sequence processing. *Predicting the future and understanding the past*, 243–264.

Nakata, T. & Trehub, S. E. (2011). Expressive timing and dynamics in infant-directed and non-infant-directed singing. *Psychomusicology: Music, Mind and Brain*, *21*(1-2), 45.

Narmour, E. (1992). *The Analysis and Cognition of Melodic Complexity: The Implication-Realization Model* (Vol. 50). doi: 10.2307/898334

Nevill-manning, C. & Witten, I. (1997). Identifying Hierarchical Structure in Sequences: A linear-time algorithm. *Journal of Artificial Intelligence Research*, *7*(1), 67–82.

Ng, A. (2012). Supervised learning. In *Machine learning course cs229 stanford* (pp. 1–30).

Oore, S., Simon, I., Dieleman, S. & Eck, D. (2017). Learning to Create Piano Performances. *Workshop on Machine Learning for Creativity and Design (NIPS 2017)*(Figure 1), 1–3.

Pachet, F. & Roy, P. (2011). Markov constraints: Steerable generation of Markov sequences. *Constraints*, *16*(2), 148–172. doi: 10.1007/s10601-010-9101-4

Paderewski, I. J., Bronarski, L. & Turczynski, J. (2011). *Fryderyk Chopin, Complete works, Mazurkas,* (29th ed.). Fryderyka Chopina, Polskie Wydawnictwo Muzyczne SA.

Palmer, C. (1989). Mapping Musical Thought to Musical Performance. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(2), 331–346. doi: 10.1037/0096-1523.15.2.331

Palmer, C. (1996). Anatomy of a Performance : Sources of Musical Expression. *Music Perception*, *13*(3), 433–453.

Palmer, C. (1997, jan). Music performance. *Annual review of psychology*, *48*, 115–38. doi: 10.1146/annurev.psych.48.1.115

Pampalk, E. (2004). A Matlab Toolbox to Compute Music Similarity from Audio. *Proceedings of the 2004 International Conference on Music Information Retrieval (ISMIR '04)*, 254–257. doi: 10.1.1.68.7606

Pampalk, E., Goebl, W. & Widmer, G. (2003). Visualizing changes in the structure of data for exploratory feature selection. In *Proceedings of the ninth acm sigkdd international conference on knowledge discovery and data mining* (pp. 157–166).

Parncutt, R. (2007). Systematic musicology and the history and future of western musical scholarship. *Journal of interdisciplinary music studies*, *1*(1), 1–32.

Pascanu, R., Mikolov, T. & Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In S. Dasgupta & D. McAllester (Eds.), *Proceedings of the 30th international conference on machine learning* (Vol. 28, pp. 1310–1318). Atlanta, Georgia, USA: PMLR.

Patel, A. (2008). *Music, language, and the brain* (Vol. 100) (No. 20). Oxford University Press.

Patel, A. (2012). *Language, music, and the brain: a resource-sharing framework* (&. I. C. P. Rebuschat, M. Rohrmeier, J. Hawkins, Ed.). Oxford University Press.

Pearce, M. (2011). Time-series analysis of Music : Perceptual and Information Dynamics. , *6*(2), 125–130.

Pearce, M., Conklin, D. & Wiggins, G. (2005). Methods for Combining Statistical Models of Music. *Computer Music Modeling and Retrieval*, *3310*, 295–312.

Pearce, M. & Wiggins, G. (2004). Improved methods for statistical modelling of monophonic music. *Journal of New Music Research*.

Peirce, J. W. (2007). PsychoPy-Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1-2), 8–13. doi: 10.1016/j.jneumeth.2006.11.017

Penttinen, H. (2006). On the Dynamics of the Harpsichord and its Synthesis. *Proceedings of the 9th International Conference on Digital Audio Effects*, 115–120.

Podos, J. (2010). Acoustic discrimination of sympatric morphs in Darwin's finches: a behavioural mechanism for assortative mating? *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *365*(1543), 1031–1039. doi: 10.1098/rstb.2009.0289

Prusinkiewicz, P. (1986). Score generation with L-systems. In *International computer music conference* (pp. 455–457).

Prusinkiewicz, P. & Lindemayer, A. (1990). *The algorithmic beauty of plants* (Vol. 31; P. Prusinkiewicz, Ed.) (No. 2). Springer-Verlag. doi: 10.1177/0149206304271602

Qian, N. (1999). On the Momentum Term in Gradient Descent Learning Algorithms The Momentum Term in Gradient Descent. *Neural Networks: The Official Journal of the International Neural Network Society*, *5213*(12(1)), 145–151.

Quantz, J. J. (1752). Versuch einer Anweisung die Flöte traversiere zu spielen. *On Playing the Flute*, 423. doi: 10.2307/3390883

Rabiner, L. & Juang, B. (1986). An introduction to hidden Markov models. *IEEE ASSP Magazine*, *3*(January), 4–16. doi: 10.1002/0471250953.bia03as18

Ramirez, R., Maestre, E. & Serra, X. (2010, sep). Automatic performer identification in commercial monophonic Jazz performances. *Pattern Recognition Letters*, *31*(12), 1514–1523. doi: 10.1016/j.patrec.2009.12.032

Ramirez, R., Maestre, E. & Serra, X. (2012, feb). A Rule-Based Evolutionary Approach to Music Performance Modeling. *IEEE Transactions on Evolutionary Computation*, *16*(1), 96–107. doi: 10.1109/TEVC.2010.2077299

Ramirez, R., Perez, A., Kersten, S. & Rizo, D. (2010). Modeling violin performances using inductive logic programming. *Intelligent Data Analysis*, *14*, 573–585. doi: 10.3233/IDA-2010-0440

Raphael, C. (2010). Symbolic and Structural Representation of Melodic Expression. *Journal of New Music Research*, *39*(3), 245–251. doi: 10.1080/09298215.2010.512978

Repp, B. (1990). Patterns of expressive timing in performances of a Beethoven minuet by nineteen famous pianists. *The Journal of the Acoustical Society of America*, *88*(2), 622–641. doi: 10.1121/1.410391

Repp, B. (1992). Diversity and Commonality in Music Performance : An Analysis of Timing Microstructure in Schumann 's Traumerei ". *The Journal of the Acoustical Society of America*, *92*(5), 2546–2568.

Repp, B. (1994). Relational invariance of expressive microstructure across global tempo changes in music performance: An exploratory study. *Psychological Research*, *56*(4), 269–284. doi: 10.1007/BF00419657

Repp, B. (1996). Detectability of duration and intensity increments in melody tones: a partial connection between music perception and performance. *Perception & Psychophysics*, *57*(8), 1217–32. doi: 10.3758/BF03208378

Repp, B. (1997). Variability of timing in expressive piano performance increases with interval duration. *Psychonomic Bulletin & Review*, *4*(4), 530–534. doi: 10.3758/BF03214344

Repp, B. & Knoblich, G. (2004). Perceiving action identity: How pianists recognize their own performances. *Psychological Science*, *15*(9), 604–609. doi: 10.1111/j.0956-7976.2004.00727.x

Rink, J., Spiro, N. & Gold, N. (2011). Motive, Gesture and the Analysis of Performance. *New Perspectives on Music and Gesture*(13), 267–292.

Ruder, S. (2016). An overview of gradient descent optimization algorithms. , 1–14. doi: 10.1111/j.0006-341X.1999.00591.x

Sadakata, M., Desain, P. & Honing, H. (2006). The Bayesian way to relate rhythm perception and production. *Music Perception: An Interdisciplinary Journal*, *23*(3), 269–288.

Sapp, C. S. (2007). Comparative Analysis of Multiple Musical Performances. In *International society for music information retrieval* (pp. 497–500).

Schuster, M. & Paliwal, K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, *45*(11), 2673–2681. doi: 10.1109/78.650093

Seashore, C. (1936). *Objective analysis of musical performance* (Vol. 4). Dover Publications.

Seashore, C. (1938). *Psychology of Music*. McGraw-Hill Book Company.

Seashore, C. & Metfessel, M. (1925). Deviation from the regular as an art principle. *Proceedings of the National Academy of Sciences of the United States of America*, *m*, 538–542. doi: 10.1073/pnas.11.9.538

Serrà, J., Özaslan, T. H. & Arcos, J. L. (2013, jan). Note onset deviations as musical piece signatures. *PloS one*, *8*(7), e69268. doi: 10.1371/journal.pone.0069268

Sloboda, J. A. (1983). The communication of musical metre in piano performance. *The Quarterly Journal of Experimental Psychology Section A*, *35*(2), 377–396. doi: 10.1080/14640748308402140

Sloboda, J. A. (1985). Expressive skill in two pianists: Metrical communication in real and simulated performances. *Canadian Journal of Psychology/Revue canadienne de psychologie*, *39*(2), 273.

Sloboda, J. A. (2000). Individual differences in music performance. *Trends in Cognitive Sciences*, *4*(10), 397–403. doi: 10.1016/S1364-6613(00)01531-X

Snyder, B. (2000). *Music and memory: An Introduction* (Vol. 98) (No. 2). MIT Press. doi: 10.1037/015227

Stamatatos, E. & Widmer, G. (2005, jun). Automatic identification of music performers with learning ensembles. *Artificial Intelligence*, *165*(1), 37–56. doi: 10.1016/j.artint.2005.01.007

Stanislaw, H. & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, {&} Computers*, *31*(1), 137–149. doi: 10.3758/BF03207704

Supper, M. (2001, mar). A few remarks on algorithmic composition. *Computer Music Journal*, *25*(1), 48–53. doi: 10.1162/014892601300126106

Tidhar, D., Dixon, S., Benetos, E. & Weyde, T. (2014). The temperament police. *Early Music*, *42*(4), 579–590. doi: 10.1093/em/cau101

Timmers, R. (2005). Predicting the similarity between expressive performances of music from measurements of tempo and dynamics. *Journal of the Acoustical Society of America*, *117*(1), 391–399. doi: 10.1121/1.1835504

Timmers, R. & Honing, H. (2002). On music performance, theories, measurement and diversity. *Cognitive Processing (International Quarterly of Cognitive Sciences)*, *1*(2), 1–19.

Todd, N. (1992). The dynamics of dynamics: A model of musical expression. *The Journal of the Acoustical Society of America*, *91*(6), 3540. doi: 10.1121/1.402843

Trehub, S. E., Plantinga, J., Brcic, J. & Nowicki, M. (2013). Cross-modal signatures in maternal speech and singing. *Frontiers in psychology*, *4*, 811.

Tsay, C.-J. (2013, aug). Sight over sound in the judgment of music performance. *Proceedings of the National Academy of Sciences*, *2013*, 1–6. doi: 10.1073/pnas.1221454110

Uitdenbogerd, A. & Zobel, J. (1998). Manipulation of music for melody matching. In *Proceedings of the sixth acm international conference on multimedia - multimedia '98* (pp. 235–240). doi: 10.1145/290747.290776

van der Weij, B., Pearce, M. T. & Honing, H. (2017). A probabilistic model of meter perception: Simulating enculturation. *Frontiers in Psychology*, *8*(MAY), 1–18. doi: 10.3389/fpsyg.2017.00824

Vaquero, C. (2015, nov). A quantitative study of seven historically informed performances of Bach's bwv1007 Prelude. *Early Music*, *43*(4), 611–622. doi: 10.1093/em/cav091

Vaquero, C. & Honing, H. (2014). Generating expressive timing by combining rhythmic categories and Lindenmayer systems. *perception*.

Vera, B. & Chew, E. (2014). Towards seamless network music performance: prediction an ensemble's expressive decisions for distributed performance. In *International society for music information retrieval conference* (pp. 489–494).

Werbos, P. J. (1990). Backpropagation Through Time: What It Does and How to Do It. *Proceedings of the IEEE*, *78*(10), 1550–1560. doi: 10.1109/5.58337

Widmer, G. (2002). Machine discoveries: A few simple, robust local expression principles. *Journal of New Music Research*, *31*(1), 37–50. doi: 10.1076/jnmr.31.1.37.8103

Widmer, G. (2003, jun). Discovering simple rules in complex data: A meta-learning algorithm and some surprising musical discoveries. *Artificial Intelligence*, *146*(2), 129–148. doi: 10.1016/S0004-3702(03)00016-X

Widmer, G., Flossmann, S. & Grachten, M. (2009). YQX plays Chopin. *AI Magazine*(August 2008), 35–48.

Widmer, G. & Goebl, W. (2004, sep). Computational models of expressive music performance: The state of the art. *Journal of New Music Research*, *33*(3), 203–216. doi: 10.1080/0929821042000317804

Widmer, G. & Tobudic, A. (2003). Playing Mozart by Analogy: Learning Multi-level

Timing and Dynamics Strategies. *Journal of New Music Research*, *32*(3), 259–268. doi: 10.1076/jnmr.32.3.259.16860

Wilcoxon, F. (1946). Individual comparisons of grouped data by ranking methods. *Journal of economic entomology*, *39*(6), 269. doi: 10.2307/3001968

Zwicker, E. & Fastl, H. (1999). *Psychoacoustics - Facts and Models, Second Edition* (Vol. 1). Springer-Verlag.

SAMENVATTING

WAT MAAKT EEN UITVOERDER UNIEK?
Idiosyncrasieën en overeenkomsten in de expressieve uitvoering van muziek

Dit proefschrift onderzoekt de productie en perceptie van idiosyncratische expressiviteit in de uitvoering van muziek. Het onderzoekt vooral hoe de expressie van uitvoerders wordt bepaald en begrensd door hun eigen idiosyncratische speelstijl en door de partituur.

In het eerste hoofdstuk wordt het werkveld van de "expressive performance modeling" binnen het kader van systematische muziekwetenschap geïntroduceerd en ook worden er alternatieve definities van expressiviteit gegeven. Verder wordt in dit hoofdstuk het onderzoek van dit proefschrift gemotiveerd.

Het tweede hoofdstuk bevat een studie die inzicht geeft in de muzikale expressiviteit, daarbij rekening houdend met "timing" en geluidssterkte zoals elke individuele uitvoerder die gebruikt. Ook bevat dit hoofdstuk alternatieve methodologieën waarmee expressie in uitvoeringen geanalyseerd kan worden.

In hoofdstuk drie worden verschillende sleutelconcepten bestudeerd op het terrein van "machine learning" en "computational modeling" toegepast op "music performance analysis and modeling".

Hoofdstuk vier bevat een studie naar zowel de analyse als de modeling van expressief gebruik van tempo en geluidssterkte bij voorschriften in de partituur, als mogelijke interactieve functies binnen het raamwerk van idiosyncratische expressiviteit. De resultaten laten zien dat zowel tempo als geluidssterkte beter voorspeld worden door de gebruikte modellen als deze getraind worden op verschillende uitvoerders die allen hetzelfde stuk spelen (d.w.z. gebruik makend van "piece-based models") dan wanneer deze modellen getraind worden op een enkele uitvoerder die verschillende stukken speelt (d.w.z. gebruik makend van "performer-based models"). De resultaten laten ook zien dat zowel tempo als geluidssterkte in de meeste gevallen beter worden voorspeld wanneer de tempo en geluidssterkte voorafgaand aan de voorschriften als "predictors" zijn opgenomen.

De bevindingen suggereren dat binnen de context zoals gedefinieerd door de dataset, uitvoerders hun expressiviteit met betrekking tot tempo en geluidssterkte inrichten naar de gegeven voorschriften in de partituur en deze mogelijk gebruiken als vaste structuur voor de frasering. Ook wordt vastgesteld dat tempo en geluidssterkte in veel gevallen geen invloed op elkaar lijken uit te oefenen door de voorschriften in de partituur. Dit geldt zowel bij gebruik van "piece-based models" als "performer-based models". Deze bevinding ontkracht de hypothese die ervan uitgaat dat tempo en gel-

uidssterkte elkaar beïnvloeden bij de voorschriften in de partituur, althans binnen de context van de hier gepresenteerde experimenten.

Het vijfde hoofdstuk bestudeert of het gebruik van tempo en geluidssterkte (en de interactie tussen deze twee variabelen door individuele uitvoerders) beperkt worden door metrum of door melodisch ritme zoals die zijn aangegeven in de betreffende partituur. Aanvullend volgt een behandeling van de vraag of het idiosyncratische gebruik van expressief tempo en geluidssterkte beter voorspeld kan worden door "piece-based models" of door "performer-based models". Bovendien geeft dit hoofdstuk een verantwoording voor het gebruik van "sequential models", met name van "Long Short-Term Memory networks" om het wezen en de aard van het musiceren en luisteren naar muziek beter te benaderen. De resultaten tonen aan dat "score-based models" tot betere voorspellingen leiden dan "perfomer-based models". Dit houdt in dat de expressiviteit van de uitvoerder binnen een gegeven aanduiding sterker beperkt wordt door de muzikale structuur inherent aan de partituur, dan door de stilistische idiosyncratische expressiviteit van de uitvoerder zelf. De resultaten duiden er bovendien op dat de metrische structuur de expressiviteit beperkt voor wat betreft tempo en geluidssterkte als interactieve functies met betrekking tot lange en korte termijn frasering.

Daarenboven staat de wisselwerking tussen tempo, geluidssterkte en metrum een betere voorspelling van zowel tempo als geluidssterkte toe, dan wanneer tempo en geluidssterkte worden gebruikt als gecombineerde functies. Echter, als metrum niet wordt meegenomen, blijkt de wisselwerking tussen tempo en geluidssterkte niet waarneembaar. Deze resultaten suggereren dat de metrische structuur de expressiviteit aangaande de geluidssterkte en tempo beperkt, en hoe de metrische structuur en deze expressiviteit elkaar beïnvloeden. In dit verband verduidelijkt het resultaat hoe de verwachtingen met betrekking tot tempo en geluidssterkte van een luisteraar mogelijk worden beperkt door de structuur van de partituur, het maatdeel en de perceptie van het metrum.

Hoofdstuk zes behandelt een perceptueel experiment over de rol van expressief tempo en geluidssterkte (zowel als onafhankelijke als interactieve functies) bij het onderscheiden van twee verschillende uitvoerders. Ook wordt nagedacht over uitdagingen met betrekking tot ontwerp en methodologie die hierbij naar voren kwamen.

Het zevende hoofdstuk ten slotte is een samenvatting van de bijdragen aan deze dissertatie.

SUMMARY

WHAT MAKES A PERFORMER UNIQUE?
Idiosyncrasies and commonalities in expressive music performance

This thesis investigates the production and perception of idiosyncratic expressiveness in music performance. In particular, it studies how performers' expressive use of tempo and loudness is constrained by their idiosyncratic style and by the score.

The first chapter introduces the subfield of expressive performance modeling under the umbrella of systematic musicology, outlines alternative definitions of expressiveness, and motivates the research presented in the dissertation.

The second chapter presents a proof of concept study showing insights in the expressiveness consistency of timing and loudness of individual performers, as well as alternative methodologies to analyze expressive performances.

The third chapter offers a review and introduces several key concepts on the field of machine learning and computational modeling with applications to music performance analysis and modeling.

The fourth chapter contains a study concerning the analysis and modeling of expressive tempo and loudness at score markings, as either independent or interacting features, within the frame of idiosyncratic expressiveness. The results show that both tempo and loudness were better predicted by the models used when these were trained on several performers playing a given piece (i.e., using piece-based models) than when the models were trained per performer playing various other pieces (i.e., using performer-based models). The results also indicate that both tempo and loudness were mostly better predicted at score markings when tempo or loudness values prior to the marking were also included as predictors. These findings suggest that, within the context defined by the dataset, performers plan and direct their expressive gestures on tempo and loudness towards the score marking, possibly as a structural phrasing inflexion. Yet, it was also observed that tempo and loudness did not - in most cases - seem to interact among the score indications tested, in either piece- or performer-based approaches. This finding invalidates the hypothesis that tempo and loudness interact at score markings, at least within the context of experiments presented.

The fifth chapter studies whether individual performers' use of tempo and loudness - and the interactions between them - is constrained by the meter or the melodic rhythm specific to the score. In addition, an assessment as to whether the idiosyncratic use of expressive tempo or loudness is better predicted by performer-based models, as opposed to piece-based models, is offered. Furthermore, this chapter motivates the use of sequential models, and in particular, Long Short-Term Memory networks, to better approximate the nature of music making and listening. The results indicate that

piece-based models lead to better predictions compared with performer-based models. This suggests that performer expressiveness within a given piece is more strongly constrained by musical structure inherent to the score than by stylistic idiosyncratic expressiveness of a given performer per se. The results furthermore indicate that metrical structure constrains the expressive gestures of loudness and tempo as interacting features on long and short-term expressive phrasing. Moreover, the interactions between tempo, loudness and meter allow for predicting both tempo and loudness better than when using tempo and loudness as combined features. However, when meter was not accounted for, interactions between tempo and loudness were not discernible. Such results suggest that metrical structure constrains the expressive gestures of loudness and tempo and how these two aspects of performance interact. In this regard, it elucidates on how a listener's expectations of tempo and loudness might be constrained by salient structure of the score, beat position and meter perception.

The sixth chapter includes a perceptual experiment on the role of expressive tempo and loudness (as independent or interacting features) when discriminating between two different performers and reflects on design and methodological challenges encountered. Finally, the seventh chapter recapitulates the contributions of this dissertation.

SUMARIO

QUÉ HACE ÚNICO A UN INTÉRPRETE?

Idiosincrasias y características compartidas en la expresividad interpretativa musical

Esta tesis investiga la idiosincrasia expresiva en la interpretación y percepción musical. En concreto, estudia cómo la expresividad en el uso de tempo y dinámica puede estar restringida por la idiosincrasia estilística del intérprete o por la partitura.

El primer capítulo presenta el área de estudio del modelado computacional de la expresivividad en la interpretación, enmarcado dentro del campo de la musicología sistemática. Así, presenta diferentes definiciones de expresividad y argumenta los motivos de la investigación presentada en esta tesis.

El segundo capítulo expone un estudio de "prueba de concepto" que analiza la consistencia de la expresividad individual en el "timing" y volumen de los intérpretes, así como distintas metodologías para el análisis de la expresividad interpretativa.

El tercer capítulo expone distintos conceptos clave en el campo del aprendizaje automático ("machine learning") con aplicaciones en el análisis y modelado computacional de la expresividad musical.

El cuarto capítulo incluye un estudio sobre el análisis y modelado del tempo y el volumen en las indicaciones expresivas de la partitura (como variables interactivas o independientes) dentro del marco de la expresividad idiosincrática. Los resultados muestran que tanto las predicciones de tempo como de volumen son mejores cuando los modelos predictivos han sido entrenados con diferentes intérpretes interpretando la misma pieza (i.e. modelos basados en la partitura) que cuando los modelos han sido entrenados con el mismo intérprete (i.e. modelos basados en el intérprete). Los resultados también indican que las predicciones de tempo y volumen mejoran cuando los modelos incorporan información referente a los dos compases anteriores a la notación expresiva de la partitura. Estos resultados sugieren que, dentro del contexto definido por el conjunto de datos disponible para este experimento, los intérpretes planifican y dirigen los gestos expresivos de tempo y volumen hacia la indicación expresiva definida en la partitura, posiblemente tratando la misma como una inflexión estructural del fraseo. No obstante, también se observó que, en la mayoría de los casos, el tempo y el volumen no parecían interactuar (como gesto expresivo) en las indicaciones de partitura estudiadas en ninguno de los dos tipos de modelos, ya sea basados en la partitura o en los interpretes. Estos resultados invalidan la hipótesis de que el tempo y el volumen interactuan en las indicaciones de la partitura, al menos dentro del contexto de los experimentos presentados.

El quinto capítulo estudia si el uso individual de tempo y volumen - y las interacciones entre los mismos - en los intérpretes, se ve restringido por la métrica específica del ritmo melódico. Adicionalmente, evalúa si el uso idiosincrático expresivo de tempo y volumen se predice mejor con los modelos obtenidos basados en la partitura o los modelos basados en los intérpretes. Además, argumenta los motivos para el uso de modelos secuenciales y, en particular, de redes neuronales "Long Short-Term Memory", para una mejor aproximación a la naturaleza secuencial de la ejecución y escucha de la música. Los resultados de los experimentos indican que la estructura métrica restringe la expresividad del tempo y el volumen y su interacción como variables expresivas. Asimismo, los modelos que combinan las variables de tempo, volumen y métrica musical predicen mejor el tempo y el volumen que aquellos en los que estas variables no se combinan. Sin embargo, cuando no se tenía en cuenta la métrica, las interacciones entre tempo y volumen no eran apreciables. Esto sugiere que la estructura métrica condiciona los gestos expresivos de tempo y volumen, así como la interacción entre los mismos. A este respecto, dilucida cómo las expectativas de tempo y volumen de un oyente podrían estar condicionadas por la estructura de la partitura, la posición del "beat" y la percepción métrica.

El sexto capítulo incluye un experimento realizado con el objetivo de entender mejor la percepción de expresividad en el uso del tempo y el volumen (como variables independientes o que interactúan) al diferenciar entre dos intérpretes. Asimismo, reflexiona sobre los retos metodológicos y de diseño encontrados en el proceso. Finalmente, el séptimo capítulo resume las aportaciones de la tesis.

TITLES IN THE ILLC DISSERTATION SERIES

ILLC DS-2009-01: **Jakub Szymanik**
Quantifiers in TIME and SPACE. Computational Complexity of Generalized Quantifiers in Natural Language

ILLC DS-2009-02: **Hartmut Fitz**
Neural Syntax

ILLC DS-2009-03: **Brian Thomas Semmes**
A Game for the Borel Functions

ILLC DS-2009-04: **Sara L. Uckelman**
Modalities in Medieval Logic

ILLC DS-2009-05: **Andreas Witzel**
Knowledge and Games: Theory and Implementation

ILLC DS-2009-06: **Chantal Bax**
Subjectivity after Wittgenstein. Wittgenstein's embodied and embedded subject and the debate about the death of man.

ILLC DS-2009-07: **Kata Balogh**
Theme with Variations. A Context-based Analysis of Focus

ILLC DS-2009-08: **Tomohiro Hoshi**
Epistemic Dynamics and Protocol Information

ILLC DS-2009-09: **Olivia Ladinig**
Temporal expectations and their violations

ILLC DS-2009-10: **Tikitu de Jager**
"Now that you mention it, I wonder…": Awareness, Attention, Assumption

ILLC DS-2009-11: **Michael Franke**
Signal to Act: Game Theory in Pragmatics

ILLC DS-2009-12: **Joel Uckelman**
More Than the Sum of Its Parts: Compact Preference Representation Over Combinatorial Domains

ILLC DS-2009-13: **Stefan Bold**
Cardinals as Ultrapowers. A Canonical Measure Analysis under the Axiom of Determinacy.

ILLC DS-2010-01: **Reut Tsarfaty**
Relational-Realizational Parsing

ILLC DS-2010-02: **Jonathan Zvesper**
Playing with Information

ILLC DS-2010-03: **Cédric Dégremont**
The Temporal Mind. Observations on the logic of belief change in interactive systems

ILLC DS-2010-04: **Daisuke Ikegami**
Games in Set Theory and Logic

ILLC DS-2010-05: **Jarmo Kontinen**
Coherence and Complexity in Fragments of Dependence Logic

ILLC DS-2010-06: **Yanjing Wang**
Epistemic Modelling and Protocol Dynamics

ILLC DS-2010-07: **Marc Staudacher**
Use theories of meaning between conventions and social norms

ILLC DS-2010-08: **Amélie Gheerbrant**
Fixed-Point Logics on Trees

ILLC DS-2010-09: **Gaëlle Fontaine**
Modal Fixpoint Logic: Some Model Theoretic Questions

ILLC DS-2010-10: **Jacob Vosmaer**
Logic, Algebra and Topology. Investigations into canonical extensions, duality theory and point-free topology.

ILLC DS-2010-11: **Nina Gierasimczuk**
Knowing One's Limits. Logical Analysis of Inductive Inference

ILLC DS-2010-12: **Martin Mose Bentzen**
Stit, Iit, and Deontic Logic for Action Types

ILLC DS-2011-01: **Wouter M. Koolen**
Combining Strategies Efficiently: High-Quality Decisions from Conflicting Advice

ILLC DS-2011-02: **Fernando Raymundo Velazquez-Quesada**
Small steps in dynamics of information

ILLC DS-2011-03: **Marijn Koolen**
The Meaning of Structure: the Value of Link Evidence for Information Retrieval

ILLC DS-2011-04: **Junte Zhang**
System Evaluation of Archival Description and Access

ILLC DS-2011-05: **Lauri Keskinen**
Characterizing All Models in Infinite Cardinalities

ILLC DS-2011-06: **Rianne Kaptein**
Effective Focused Retrieval by Exploiting Query Context and Document Structure

ILLC DS-2011-07: **Jop Briët**
Grothendieck Inequalities, Nonlocal Games and Optimization

ILLC DS-2011-08: **Stefan Minica**
Dynamic Logic of Questions

ILLC DS-2011-09: **Raul Andres Leal**
Modalities Through the Looking Glass: A study on coalgebraic modal logic and their applications

ILLC DS-2011-10: **Lena Kurzen**
Complexity in Interaction

ILLC DS-2011-11: **Gideon Borensztajn**
The neural basis of structure in language

ILLC DS-2012-01: **Federico Sangati**
Decomposing and Regenerating Syntactic Trees

ILLC DS-2012-02: **Markos Mylonakis**
Learning the Latent Structure of Translation

ILLC DS-2012-03: **Edgar José Andrade Lotero**
Models of Language: Towards a practice-based account of information in natural language

ILLC DS-2012-04: **Yurii Khomskii**
Regularity Properties and Definability in the Real Number Continuum: idealized forcing, polarized partitions, Hausdorff gaps and mad families in the projective hierarchy.

ILLC DS-2012-05: **David García Soriano**
Query-Efficient Computation in Property Testing and Learning Theory

ILLC DS-2012-06: **Dimitris Gakis**
Contextual Metaphilosophy - The Case of Wittgenstein

ILLC DS-2012-07: **Pietro Galliani**
The Dynamics of Imperfect Information

ILLC DS-2012-08: **Umberto Grandi**
Binary Aggregation with Integrity Constraints

ILLC DS-2012-09: **Wesley Halcrow Holliday**
Knowing What Follows: Epistemic Closure and Epistemic Logic

ILLC DS-2012-10: **Jeremy Meyers**
Locations, Bodies, and Sets: A model theoretic investigation into nominalistic mereologies

ILLC DS-2012-11: **Floor Sietsma**
Logics of Communication and Knowledge

ILLC DS-2012-12: **Joris Dormans**
Engineering emergence: applied theory for game design

ILLC DS-2013-01: **Simon Pauw**
Size Matters: Grounding Quantifiers in Spatial Perception

ILLC DS-2013-02: **Virginie Fiutek**
Playing with Knowledge and Belief

ILLC DS-2013-03: **Giannicola Scarpa**
Quantum entanglement in non-local games, graph parameters and zero-error information theory

ILLC DS-2014-01: **Machiel Keestra**
Sculpting the Space of Actions. Explaining Human Action by Integrating Intentions and Mechanisms

ILLC DS-2014-02: **Thomas Icard**
The Algorithmic Mind: A Study of Inference in Action

ILLC DS-2014-03: **Harald A. Bastiaanse**
Very, Many, Small, Penguins

ILLC DS-2014-04: **Ben Rodenhäuser**
A Matter of Trust: Dynamic Attitudes in Epistemic Logic

ILLC DS-2015-01: **María Inés Crespo**
Affecting Meaning. Subjectivity and evaluativity in gradable adjectives.

ILLC DS-2015-02: **Mathias Winther Madsen**
The Kid, the Clerk, and the Gambler - Critical Studies in Statistics and Cognitive Science

ILLC DS-2015-03: **Shengyang Zhong**
Orthogonality and Quantum Geometry: Towards a Relational Reconstruction of Quantum Theory

ILLC DS-2015-04: **Sumit Sourabh**
Correspondence and Canonicity in Non-Classical Logic

ILLC DS-2015-05: **Facundo Carreiro**
Fragments of Fixpoint Logics: Automata and Expressiveness

ILLC DS-2016-01: **Ivano A. Ciardelli**
Questions in Logic

ILLC DS-2016-02: **Zoé Christoff**
Dynamic Logics of Networks: Information Flow and the Spread of Opinion

ILLC DS-2016-03: **Fleur Leonie Bouwer**
What do we need to hear a beat? The influence of attention, musical abilities, and accents on the perception of metrical rhythm

ILLC DS-2016-04: **Johannes Marti**
Interpreting Linguistic Behavior with Possible World Models

ILLC DS-2016-05: **Phong Lê**
Learning Vector Representations for Sentences - The Recursive Deep Learning Approach

ILLC DS-2016-06: **Gideon Maillette de Buy Wenniger**
Aligning the Foundations of Hierarchical Statistical Machine Translation

ILLC DS-2016-07: **Andreas van Cranenburgh**
Rich Statistical Parsing and Literary Language

ILLC DS-2016-08: **Florian Speelman**
Position-based Quantum Cryptography and Catalytic Computation

ILLC DS-2016-09: **Teresa Piovesan**
Quantum entanglement: insights via graph parameters and conic optimization

ILLC DS-2016-10: **Paula Henk**
Nonstandard Provability for Peano Arithmetic. A Modal Perspective

ILLC DS-2017-01: **Paolo Galeazzi**
Play Without Regret

ILLC DS-2017-02: **Riccardo Pinosio**
The Logic of Kant's Temporal Continuum

ILLC DS-2017-03: **Matthijs Westera**
Exhaustivity and intonation: a unified theory

ILLC DS-2017-04: **Giovanni Cinà**
Categories for the working modal logician

ILLC DS-2017-05: **Shane Noah Steinert-Threlkeld**
Communication and Computation: New Questions About Compositionality

ILLC DS-2017-06: **Peter Hawke**
The Problem of Epistemic Relevance

ILLC DS-2017-07: **Aybüke Özgün**
Evidence in Epistemic Logic: A Topological Perspective

ILLC DS-2017-08: **Raquel Garrido Alhama**
Computational Modelling of Artificial Language Learning: Retention, Recognition & Recurrence

ILLC DS-2017-09: **Miloš Stanojević**
Permutation Forests for Modeling Word Order in Machine Translation

ILLC DS-2018-01: **Berit Janssen**
Retained or Lost in Transmission? Analyzing and Predicting Stability in Dutch Folk Songs

ILLC DS-2018-02: **Hugo Huurdeman**
Supporting the Complex Dynamics of the Information Seeking Process

ILLC DS-2018-03: **Corina Koolen**
Reading beyond the female: The relationship between perception of author gender and literary quality

ILLC DS-2018-04: **Jelle Bruineberg**
Anticipating Affordances: Intentionality in self-organizing brain-body-environment systems

ILLC DS-2018-05: **Joachim Daiber**
Typologically Robust Statistical Machine Translation: Understanding and Exploiting Differences and Similarities Between Languages in Machine Translation

ILLC DS-2018-06: **Thomas Brochhagen**
Signaling under Uncertainty

ILLC DS-2018-07: **Julian Schlöder**
Assertion and Rejection

ILLC DS-2018-08: **Srinivasan Arunachalam**
Quantum Algorithms and Learning Theory

ILLC DS-2018-09: **Hugo de Holanda Cunha Nobrega**
Games for functions: Baire classes, Weihrauch degrees, transfinite computations, and ranks

ILLC DS-2018-10: **Chenwei Shi**
Reason to Believe

ILLC DS-2018-11: **Malvin Gattinger**
New Directions in Model Checking Dynamic Epistemic Logic

ILLC DS-2018-12: **Julia Ilin**
Filtration Revisited: Lattices of Stable Non-Classical Logics

ILLC DS-2018-13: **Jeroen Zuiddam**
Algebraic complexity, asymptotic spectra and entanglement polytopes

Institute for Logic, Language and Computation