JOHANNES KEPLER
UNIVERSITÄT LINZ | JKU

TNF

Technisch-Naturwissenschaftliche
Fakultät

Expressive Performance Rendering With Probabilistic Models

Creating, Analyzing, and Using the Magaloff Corpus

DISSERTATION

zur Erlangung des akademischen Grades

Doktor

im Doktoratsstudium der

Technischen Wissenschaften

Eingereicht von:
Sebastian Flossmann

Angefertigt am:
Department of Computational Perception

Beurteilung:
Prof. Dr. Gerhard Widmer (Betreuung)
Prof. Dr. Richard Parncutt

Linz, März, 2012

Kurzfassung

Im Jahr 1989 entstanden im Mozartsaal des Wiener Konzerthauses im Rahmen eines Konzertzyklus des russischen Pianisten Nikita Magaloff eine einzigartige Sammlung von Aufnahmen: Das komplette, zu Lebzeiten des Komponisten veröffentlichte Solo-Klavierwerk Frédéric Chopins, gespielt nicht auf einem gewöhnlichen Konzertflügel, sondern auf einem Bösendorfer SE Computer Controlled Grand Piano, ein Flügel, der alle Tasten- und Pedalbewegungen als Liste von exakt vermessenen Ereignissen zur Verfügung stellt. Im "Magaloff Corpus" ist jede Note, die Magaloff in diesen Konzerten gespielt hat, verknüpft mit ihrer Entsprechung im Notentext, was die Aufnahmen zu einem unvergleichlichen Hilfsmittel für alle Anwendung und Fragen macht, die die Untersuchung von Musik zum Inhalt haben.

Der Anfang der vorliegenden Dissertation ist der Entstehung des Corpus gewidmet: Zusätzlich zu den Aufnahmen der Stücke müssen die Notentexte in ein computer–lesbares und symbolisches Format gebracht werden. Diverse Probleme, die bei den Scan-, Transformations- und Codierungsvorgängen auftreten, und mögliche Lösungen werden diskutiert; ebenso die Alignment-Software, die eigens zu dem Zweck geschaffen wurde, Notentext und Aufführung möglichst effizient einandern zuordnen zu können, und bestehende Zuordnungen überprüfen und korrigieren zu können.

Nachfolgend beschreibe Ich erste Analysen des musikalischen Inhaltes des Corpus: Ein wesentlicher Teil der Analysen betrifft die Fehler, die Magaloff beim Spielen der Stücke unterlaufen sind. Quantitative und qualitative Aspekte werden untersucht, die die Häufigkeit von Fehlern in Zusammenhang setzen mit ihrer Offensichtlichkeit für ein Konzertpublikum. Ergänzend wird im Anschluss daran eine Klassifikation von Fehlern und Fehlergruppen vorgestellt, ein Belege dafür, dass Fehler meistens weit mehr sind, als zufällige Ausrutscher. Dem Komplex der Fehleranalysen schließt sich eine Studie über die Effekte des Alters von Pianisten an und in wie weit sie auf Magaloff zutreffen. Eine Untersuchung von zeitlichen Asynchronizitäten zwischen der linken und der rechten Hand Magaloff's beschließt das Kapitel.

Das Hauptaugenmerk der Dissertation gilt dem Themenbereich "Expressive Perfor-

mance Rendering", dem Versuch, automatisch eine ausdrucksvolle, möglichst natürlich und menschlich klingende Aufführung eines Musikstücks zu generieren. Grundlage des hier vorgestellte Models ist ein graph-basiertes Wahrscheinlichkeitsnetzwerk (graphical probabilistic model). Auch hier spielt das Magaloff Corpus eine zentrale Rolle, in diesem Fall als Trainingsdatensatz für das Netzwerk. Basierend auf einem ersten, sehr einfachen Model stelle Ich zwei Erweiterungen vor: Ein neuer Algorithmus für die Vorhersage ermöglicht es, bereits vorhergesagte Passagen des Stücks in den aktuellen Vorhersagen zu berücksichtigen und dadurch zeitlichen Abhängigkeiten zu modellieren. Des weiteren sorgt ein Aufteilen der Ausdrucksdimensionen Lautstärke und Tempo in unterschiedliche Komponenten für eine adäquatere Repräsentation, die durch das Netz besser gelernt und wiedergegeben werden können. Sowohl das einfache System, als auch das Erweiterte wurden in einem internationalen Wettbewerb für Performance Rendering Systeme (2008 und 2011) mit Preisen ausgezeichnet.

Abstract

The Magaloff corpus is a collection of on-stage recordings of essentially the complete works of F. Chopin by a world-class pianist. The recordings were made on a Bösendorfer SE computer controlled grand piano and contain precise information about all played notes and pedal movements. This thesis concentrates on three topics that are closely related to the corpus: Creating the corpus, assessing performance-related questions, and – the main focus – rendering expressive performances with probabilistic models using the Magaloff corpus as training data.

Creation of the corpus consists of first preparing the scores of all played pieces (scanning the scores, transforming the images into symbolic scores) and then aligning the scores with the performances. I discuss the steps and problems involved in the process, and present a software designed to facilitate the large-scale alignment process.

Analysis of the corpus touches on three aspects related to music performance: performance errors, the effects of age on a performer, and between-hand asynchronies. Performance errors are investigated from a quantitative and a qualitative perspective, assessing possible relations between error frequency and obviousness. A categorization of performance errors in abstract groups and patterns depicts a phenomenon that is much more elaborate than simple "accidents". Possible effects of age on a performer are discussed and assessed in Magaloff's Chopin performances. A discussion of expressive asynchronies between the Magaloff's left and right hand concludes the analysis section.

Expressive performance rendering, the endeavor to generate a human-like, naturally sounding, expressive performance automatically, is the focus of the remainder of the thesis. I propose a probabilistic graphical model as learning and prediction mechanism. It is both a rendering system in its own right (it won an international competition for computer rendering systems in 2008) and serves as a basis for two extensions also presented in this thesis: the first addresses long-term dependencies in the predicted performance parameters (performance context), the second proposes a decomposition of performance aspects into separate, independent components. The extended system won the competition again three years later, in 2011.

Acknowledgments

Like it is with all the big moments in life, many people contributed to this thesis in their own way, some directly, some in a oblique and round-about way, and I want to take this opportunity to thank some of them for it.

It is customary to light the first beacon of gratitude for one's supervisor. I also want to do just that, not out of custom, but out of great respect and gratefulness. Thank you, Gerhard, for giving me the opportunity to work with something I really care about, for being such an inspiration and support over the last few years, for always taking the time to listen and, if needed, having a piece of advice. Working with you has been an honor.

My family has been a huge support over the years, morally, financially, and through endless discussions. They also have subtly prevented me from trying to be a concert pianist, for which I have eventually become grateful.

My colleagues in Linz all have played their part well and deserve mentioning, especially Andreas Arzt, Bernhard Niedermayer, Peter Knees, Reinhard Sonnleitner, and Sebastian Böck (in alphabetical order). Thanks guys, it really has been a laugh. Thank you Claudia, you are the soul of CP! I also want to thank my friend and former co-worker Werner Goebl for many inspiring discussions, collaborations, and numerous shared lunch and coffee breaks.

I also want to express my gratitude to Mme Irene Magaloff for her generous permission to use this unique data resource for my research. On the financial side all credit goes to the Austrian National Research Fund (FWF) (grants P19349-N15, Z159 ("Wittgenstein Award"), and TRP 109-N23).

Of all the people that are not part of the aforementioned, but still have, in some way or other, helped and/or contributed I want to highlight a few: Roman, for being the best friend one could possibly imagine; Ingrid, for crossing my i's and dotting my t's and somehow always being there when I needed you; Birgit, Inna, and Clara, for having the patience of introducing me to Schubert, Rachmaninoff, and (surprisingly) Mozart. Last, but in no way least, thank you, Brigitte, for all the patience, understanding, and motivation to finish the thesis in the end.

Contents

1 Introduction **1**
- 1.1 Data Acquisition . . . 4
- 1.2 Data Collections and Performance Research . . . 5
- 1.3 Expressive Performance Modelling & Rendering . . . 6
- 1.4 Contributions and Organisation . . . 9

2 The Magaloff Corpus **13**
- 2.1 Nikita Magaloff . . . 14
- 2.2 The Magaloff Concerts . . . 15
- 2.3 The Bösendorfer SE . . . 15
- 2.4 Preparation of the Scores . . . 18
  - 2.4.1 From Score Sheets to Images . . . 19
  - 2.4.2 From Images to MusicXML – Optical Music Recognition . . . 20
  - 2.4.3 From MusicXML to Extended MusicXML . . . 26
- 2.5 Score-Performance Matching . . . 29
  - 2.5.1 Graphical Matching . . . 31
- 2.6 Overview of the corpus . . . 33
- 2.7 Applications of the Corpus . . . 33

3 Inspecting Magaloff **37**
- 3.1 Performance Errors . . . 37
  - 3.1.1 Definition of Performance Errors . . . 38
  - 3.1.2 Related Work . . . 39

3.1.3 Quantitative Results . . . 41
3.1.4 Qualitative and Perceptual Results . . . 43
3.1.5 Error Categorization . . . 45
3.1.6 Conclusion . . . 52
3.2 Performer Age . . . 54
3.2.1 Performance Tempo in Chopin's Etudes . . . 55
3.2.2 Age effects and tempo contrast in a Nocturne . . . 57
3.2.3 Conclusion . . . 57
3.3 Between-hand Asynchronies as Expressive Device . . . 58
3.3.1 Conclusion . . . 60

4 YQX – Expressive Performance Rendering 61
4.1 Related Work . . . 62
4.1.1 Score Models . . . 62
4.1.2 Performance Models . . . 64
4.1.3 Learning and Prediction Models . . . 65
4.1.4 Rendering of expressive annotations . . . 67
4.2 Score Model . . . 68
4.2.1 Rhythmic features . . . 69
4.2.2 Melodic features . . . 69
4.2.3 Harmonic features . . . 73
4.2.4 Phrase related features . . . 74
4.2.5 Narmour's Implication-Realization (IR) model . . . 75
4.3 Performance Model . . . 76
4.3.1 Articulation . . . 76
4.3.2 Loudness . . . 77
4.3.3 Tempo . . . 77
4.4 YQX - the First Step . . . 78
4.5 Introducing performance context - Step Two . . . 80
4.5.1 YQX with local maximisation . . . 81
4.5.2 Global Optimization . . . 81
4.6 Composite Performance Dimensions - Step Three . . . 87
4.6.1 Loudness & Performance Directives . . . 87
4.6.2 Tempo as a composite phenomenon . . . 89

5 Evaluation and Experiments **93**
5.1 Data and Experiment Setup 94
5.2 Problems of Automatic Evaluation 96
5.3 Score model evaluation 97
5.3.1 Separation Qualities of Different Score Features 97
5.3.2 Feature Selection 98
5.4 Articulation Prediction 101
5.4.1 Quantitative Evaluation 102
5.4.2 Qualitative Evaluation 103
5.5 Loudness Prediction 104
5.5.1 Complete Loudness Curve 105
5.5.2 Local Loudness 106
5.5.3 Qualitative Evaluation 108
5.6 Tempo Prediction 115
5.6.1 Complete Tempo Curve 116
5.6.2 Tempo as a composite phenomenon 117
5.6.3 Qualitative Evaluation 121
5.7 Listener Evaluation – the Rendering Contest RENCON 129
5.7.1 Putting it all together 129
5.7.2 RENCON 2008 130
5.7.3 YQX 0.1 - the RENCON 2008 model 131
5.7.4 RENCON 2011 133
5.7.5 YQX 0.2 Featuring the BasisMixer - the RENCON 2011 model . 134
5.8 Summary 136

6 Conclusions and Future Work **139**
6.1 Main Contributions and Results 139
6.2 Future Directions... 142
6.2.1 ... concerning the state of the corpus 142
6.2.2 ... concerning performance rendering 142

Bibliography **146**

A Graphical Probabilistic Networks **161**
A.1 Basic statistical concepts 162
A.1.1 Discrete Random Variables 162
A.1.2 Continuous Random Variables - Gaussian Distributions 164
A.2 Bayesian Networks 167
A.2.1 Inference in Bayesian Networks 171
A.2.2 Types of CPDs 173
A.2.3 Learning in Bayesian Networks 175
A.3 Dynamic Bayesian Networks 177

B Appendix B **179**
B.1 The Magaloff Corpus 179

C RECON 2008 & 2011 – Pieces and Awards **185**
C.1 Rencon Set Pieces 2008 185
C.2 Rencon Awards 2008 190
C.3 Rencon Set Pieces 2011 193
C.4 Rencon Awards 2011 198

D Tables **201**

Chapter 1

Introduction

It's easy to play any musical instrument: all you have to do is touch the right key at the right time and the instrument will play itself.

J. S. Bach

Everyday and sublime. That's what it is.

Stephen Fry about classical music

Music is part of everyday life, it is one of the elemental forms of human communication. Making music is an expression of the most basic human feelings, desires, hopes, and wishes. Listening to music can stir emotions, titillate our intellect, excite memories, manipulate moods. It can be entertainment, background, working companion, pacemaker for daily workouts, and it can be art, with beauty in every tiny detail, every facet scrutinized, overwhelming. As Stephen Fry puts it: "Everyday and sublime. That's what it is."[39]

However, there's a long way to go from the mind of a composer creating music – often in so very everyday environments like Beethoven's cluttered, even squalid study – to the sublime experience to be had in a concert hall or in the comfort of one's living room. Much research has already been invested at several stops along the way: The study of acoustics has lead to concert halls being designed and built specifically to ensure, that not only does the music indeed reach every part of the room, but also does so pure and undistorted. Acoustics and sound engineering provide recording and playback equipment to also enjoy the rich details of music performance at home. Hundreds of years of experience and development in crafting have created music instruments that give musicians the means to express their musical ideas in seemingly infinitely many different ways. At the other end of the chain, musicological research investigates the theory and structure of music, the art of composition.

Bach's very tongue-in-cheek quotation that opens this thesis puts the finger on the keystone of the whole complex: Music Performance, the production of music, both craft and art, which brings written music to life, with the performer as a bridge between the composer and the audience. Setting aside the aspect of music performance as a craft, a feat of astounding motor control, precision, velocity, and memory, which in itself takes years and years of training, there is much we *think* we know about music performance and interpretation: how Mozart should be played; the correct way of phrasing Chopin; what particular kind of ritardando to use at the end of a Bach Fugue. However, music performances are as varied as their performer, are (among many other factors) product of their performer's experiences, social background, taste, and current mood. And still, performances are far from arbitrary: for every "right" way to perform a piece there are infinitely many ways to do it "wrong" – all of them immediately detected by a listener. The human mind is on the one hand incredibly good at perceiving minute differences or irregularities in tempo, loudness, or pitch, but on the other hand hopeless at measuring and remembering absolutes[1]. We are fast to spot anything that might sound inappropriate to us at a particular moment in a particular context, but often are not able to justify, let alone quantify, this reaction.

The field of expressive performance research investigates how performers use the means available to them – be they as generic as *loudness* and *tempo*, or as specific to an instrument as the different lip and tongue techniques used for playing the flute – to play a piece in a way that explains the piece's structure to the audience, conveys its character, the intended mood, and, on top of that, is unique and distinctive. Expressive performance research tries to learn what sounds "right", tries to find the commonalities of all those "correct" Chopin phrases, and what shape a ritardando usually has. What makes a performance expressive?

A common and sensible approach to answering such questions is by examining a large number musical performances. However, listening to recording after recording of the same piece will not help to explain commonalities and differences between the recordings in a tangible, quantitative way. The help that computers have to offer regarding this kind of analysis is limited. Extensive human involvement is required to prepare audio recordings in a way that facilitates quantitative and objective comparisons and analyses. And still, any manually prepared data is constrained by human perception.

At the heart of this thesis lies the Magaloff corpus, a huge collection of precisely mea-

[1] This obviously excludes people with perfect pitch or rhythm.

sured recordings of a world-class pianist playing the entire Œvre of Frédéric Chopin. In the context of expressive performance research this resource can serve multiple purposes and gives rise to the goals of this thesis:

Create the corpus: Preparing the data, in itself an enormous task, requires new software and techniques to cope with the size of the corpus, and will precipitate the know-how to deal with collections of this kind.

Analyze the corpus: The collection facilitates exact, representative analyses of expressive music performance, unaffected by the limits of human perception. Studies of aspects of performance that are impossible to investigate properly through audio recordings, will shed more light on the elusive complex of music performance.

Use the corpus: The corpus fuels the research strand of (computational) *Expressive Performance Modeling*, "an attempt at formulating hypotheses concerning expressive performance in such a precise way that they can be empirically verified (or disproved) on real measure performance data" [129]. Such hypotheses find a voice in *Expressive Rendering Systems*, which sonify the hypotheses on pieces of music, and thereby create an expressive performance automatically. Often used to test the general validity of existing hypotheses about expression in music, rendering systems can also serve as an alternative to analyzing audio recordings in generating and exploring new hypotheses. The most obvious goal, to be able to generate a profoundly musical performance of a hitherto unknown piece of music automatically, is hardly ever reachable. However, even crude predictive models of expression can be of extreme help on the way to "synthetic and automatic expression". A possible scenarios could be a small theater that cannot afford a real orchestra and needs to resort to digitally synthesized music. A basic, expressive version provided by a rendering system, which is then refined manually, could be of invaluable help. In this respect, the main objective of the thesis is to explore the use of probabilistic models for the purpose of expressive performance rendering. In addition, by gradually improving a very basic model, we hope to gain more insights into how expressive performance works.

The remainder of this chapter is organized as follows: Section 1.1 illustrates one of the main problems in expressive performance research – the acquisition of suitable data – and the difficulties associated with it. To put the Magaloff corpus into context, I describe some existing data collections in section 1.2. An introduction to the field of Expressive

Performance Rendering, is given in section 1.3. An outline of the thesis and overview of the main contributions (section 1.4) concludes the chapter.

1.1 Data Acquisition

Data is the crucial element in empirical analyses of musical expression. Audio recordings exist by the thousands, but with current technology there are severe limitations as to what can be extracted from the audio signal. Techniques like Nonnegative Matrix Factorization [80] or Neural Networks [6] can already gain considerable information regarding pitch and onset, especially if they know what to expect (e.g. guided by the score of the piece). To a certain extent, this permits investigations into tempo of performances. Success depends on the instrument and musical content, as much as on the required accuracy: Dealing with instruments with fixed pitches, like the piano, is easier than dealing with instruments like the violin, where pitches that are subject to continuous expressive variation; accuracy is much higher on monophonic then on polyphonic instruments; music where each note is articulated clearly and not played too softly, like for instance Bach and Mozart, is much easier to process than pieces with lots of sustain pedal, and silent legato playing, where part of the notes just set a background harmonic texture, like sometimes found in Chopin's Nocturnes. To study articulation, it is essential to determine when a note is terminated, which is much harder to extract from audio recordings. As of now, extraction of loudness of individual notes is next to impossible.

Digital pianos and, even more so, computer controlled pianos, bypass all the problems above: they record precisely and in a symbolic way everything that happens. Instead of audio waves, they produce a list of time coded events describing the played notes and the pedal movements. The main problem that arises is *Availability*. It almost never is a problem, to collect a significant number of professional audio recordings of a (moderately well-known) piece, or, in extension, of the complete works of a composer. Nor is it often difficult to collect a substantial number of audio recordings of one specific pianist. Professional recordings on digital pianos or computer controlled grand pianos, however, are few and far between. It is therefore difficult to study (1) the idiosyncrasies of one performer, (2) the stylistic and interpretational characteristics of a composer, or (3) the differences between the interpretational strategies of different performers. Answering those questions would require several recordings of the same pianist, recordings of different pieces of the same composer, or several recordings of the same piece by different

performers.

Central to this thesis is the "Magaloff Corpus", a unique collection of recordings, made on a Bösendorfer SE, a computer controlled grand piano. It comprises the complete works for solo piano by F. Chopin that was published in his lifetime, played and recorded on stage by Nikita Magaloff, a world-class pianist.

1.2 Data Collections and Performance Research

Several collections exist that have been used very successfully to study aspects of musical performance and expression in music. Prominent and prototypical among those are the two collections created by Bruno Repp: (1) A set of 28 performances of Schumann's *Träumerei* by 24 professional pianists (Op. 15, No. 7), where the times of all half-beats have been marked manually [98, 99]; (2) a set of 4 complete pieces[2] recorded by 10 graduate students in the Yale School of Music on a Yamaha Disklavier (in MIDI format), manually matched to symbolic representations of the scores [100].

Several collections exist that are similar in kind to (1): One or several pieces, played by different pianists, with tempo information extracted manually from commercial audio recordings. Tobudic and Widmer prepared a collection of 15 different sections of Mozart piano sonatas by 6 different pianists (G. Gould, D. Barenboim, A. Schiff, M.J. Pires, M. Uchida, and R. Batik) [118]; for the study on the effect of age on performance tempo ([27] and section 3.2) W. Goebl and I prepared a collection of 289 recordings of 18 Études by Chopin played by 14 different pianists. The latter further enriched the pool of over 500 manually beattracked recordings of pianists such as M. Argerich, V. Horowitz, G. Gould playing pieces by various composers that had already been collected and prepared by Gerhard Widmer. While a critical resource for all investigations into tempo and timing, this kind of collection does not contain any information pertaining to loudness of individual notes or articulation. Also, precision is of course limited by human perception.

The CrestMuse Database[3] is somewhere in-between (1) and (2). In its current state it contains over a hundred scores and recordings of different pieces and performers, manually transcribed from audio to MIDI, providing note on- and offset times as well

[2] Chopin Prelude Op. 28 No. 15 (*Raindrop*), Grieg Lyric Piece Op. 43 No. 5 (*Erotik*), Schumann Op. 15 No. 7 (*Träumerei*), and Debussy Prélude No. 8 (Book I) (*La fille aux cheveux de lin*).

[3] `www.crestmuse.jp/pedb/`

as loudness information for all notes. As all transcriptions are made by human experts, precision of tempo and reliability of loudness and note offset times is limited.

A collection of 13 complete Mozart sonatas, recorded by the Viennese pianist R. Batik on a Bösendorfer SE290 computer controlled grand piano, was prepared by Widmer in a way similar to the Magaloff corpus [126]. By nature, the performance information is as detailed, complete, and rich as in the Magaloff corpus. The score information, however, is limited to the note content of the score and some information concerning key and meter of the pieces. No performance annotations were included.

Apart from that, specialized datasets for psychological and psycho-acoustical experiments have been prepared, containing recordings of artificial musical stimuli, designed for specific purposes. For instance, Palmer's study on the way music is organized and stored in memory is based on such a collection [84].

Instigated by the nature of the available collections and technology, data oriented performance research focussed on aspects of tempo: Clarke inspected meter and rhythm in Satie's *Gnossienne* No. 5 [14, 107], Repp analyzed expressive timing in Schumanns *Träumerei* [98, 99], recent studies concern different models of final ritards [47]. The Magaloff corpus and other collections like it open up new strands of performance research: Repp's study on performance errors [100], which is seminal to our own investigations into the phenomenon, is only possible on precisely measured and score-linked data; in addition to precision, Goebl's study of between-hand asynchronies [45] builds on the detailed score information only found in the Magaloff corpus.

1.3 Expressive Performance Modelling & Rendering

The possibilities of music notation are limited: the formalism is constrained to represent only a basic skeleton of the composer's intention. The musician is an integral part of the system, shaping the piece, explaining its structure to the audience, making the music engaging and effective. The result is an *expressive performance* of the piece, music brought to life. Performing a music piece expressively is a highly creative process and the number of fundamentally different performances of the classical repertoire is vast. Performances differ in ways significant enough for humans to be able – to a certain degree – to recognize artists by their style of playing and for computers to able to tell pianists apart in pairwise comparisons [131].

However different performances of the same piece may be, artistic freedom still has

its limits: an epoch's mentality shapes the stylistic boundaries of the composition; performance traditions evolve and shape what a concert audience expects; perception, both the musician's and the audiences', also puts limits on what is possible to be distinguished and, in consequence, produced. Most important in demarcating artistic freedom is the structure of the composition itself. It is generally agreed that one of the central elements of music performance is to explain to the audience how the piece is put together. This usually is what hypotheses concerning expressive performance put into words: How are structural elements of a composition represented in the performance of the piece? Central to those hypotheses can be any element of music, be it phrasing, harmonic tension and relief, dynamic or agogic changes (*crescendo/decrescendo* or *accelerando/ritardando* respectively). Computational models try to formulate those hypothetical relations precisely so that they can be tested on real performance data. Many of those models have been established over the last years. *Kinematic models*, for instance, relate dynamic and agogic change to the laws of physical movement – among those, Todd's phrase model [120], which couples of tempo and loudness, and proposes a quadratic relation to score time. *Rule systems* formulate basic "if-then" relationships between score and expressive variation – as for example the KTH rule system by Friberg et al. [35]. A very thorough overview of the different types of models can be found in [129].

All those models try to explain music performance, or at least certain parts or elements. *Expressive performance rendering* gives computational expression models a voice: The proposed hypotheses and models, formulated based on real performances, are used to generate new performances with the aim of giving the outcome the "human-like" characteristics of a real performance.

Systems for expressive performance rendering usually come in two different forms: *autonomous systems* and *interactive systems*. Interactive systems offer the user an alternative medium to change and shape a musical piece, other than the instrument(s) for which the piece was intended. Limited by the system's range of possibilities, the user can impose their "interpretation" on a mechanical rendering of a piece. The levels of abstraction are as varied as the devices and interfaces through which users interact with the system. On one end of the scale is the very simple, but effective "Air-Worm" [22], where a two-dimensional visualisation of performance tempo and loudness, the "performance worm" [21], is reversed: instead of displaying the performance trajectory of a piece, the user can actively manipulate the trajectory, which is then applied to a mechanical rendering of a music score. The means of manipulating the trajectory can basically be

every device capable of capturing position in two dimensions: a computer mouse moving over a surface or a hand tracked by a MIDI Theremin. A little closer in used imagery to the musical world is the system by T. Baba that uses an enhanced conductors baton as controlling device. The movement of the baton is mapped to expressive dimensions of the music [55]. An approach, that takes the medium through which the interpretation is transferred far away from the musical world, is taken by Chew in the *Expression Synthesis Project* [55]. The score of the piece is analyzed and transformed into a road, the curves of which correspond to musically important events. The user then, like in a computer game, drives a car along the road, the speed and turns of which are translated into expressive variations of loudness and tempo. Both of the latter participated in the International Rendering Contest 2008 (RENCON08), held in Japan (see section 5.7.2 for further details). Canazza and Rodà [12, 11] use a user controlled case-based approach (CaRo, presented and entered into Rendering Contest 2011 (RENCON11), see 5.7.4), with focus on emotions: musical parameters, such as tempo and loudness, are associated with performance styles or intended emotional content, like happy, sad, or angry. The user can then modify the interpretation of the piece by navigating between the different emotional states.

In interactive rendering systems the user is indispensable. The focus is on developing new interfaces to music, musical expression, and music interpretation. The systems act as proxy between user and music, and replace the original instruments with an abstraction. The range of interpretational possibilities of course is limited compared to the original instrument, but the set of skills necessary to operate the interface is proportionally easier to acquire. Navigating a computer mouse between four preset emotional states, like in CaRo, only requires basic musicality and no technical skills in order to produce a musical interpretation of a piece (within the limits of the systems). The "Air-Worm" offers more or less direct access to two very profound expressive dimensions of musical performance, loudness and tempo, but is much harder to control.

Autonomous rendering systems take the user out of the equation and try to come up with a musical interpretation of a given piece without any interaction. The systems work on abstract representations of the score, the *score model*. In most cases one component of the system is learned from real performances and connects the score model to the *performance model*, the expressive dimensions along which the expressive performance is shaped. The expressive performance of a piece is then built by feeding the score representation of a piece into the learned model and applying the output of the model,

the predicted performance parameters, to a mechanical rendering of the piece. Hence, instead of offering new ways to interact with music and transfer the user's interpretation, autonomous systems try to make the computer able to recognize, and react to, musically important aspects of the piece and build an expressive interpretation of a piece of music that ideally bears characteristics of a human performance.

1.4 Contributions and Organisation

The remainder of the thesis comprises five chapters, which are outlined in the following. The Magaloff Corpus is a large collection of performances both precise in measurement and representative. It comprises the complete works for solo piano by F. Chopin played by a world-class pianist, Nikita Magaloff, on a computer-controlled grand piano. However, the real potential of the data for performance research can only be accessed provided that the performances are linked to the scores. Chapter 2 describes the process of first digitizing and then converting the sheet music into a suitable digital representation. The subsequent large-scale score-performance matching provides a first alignment between score and performances but still is very error-prone. A software is presented that facilitates the necessary extensive manual corrections, as well as certain alterations and additions to the scores. The resulting performance corpus is the largest of its kind and unique in precision, and nature: one pianist playing the complete works of one composer. Due to copyright issues the corpus itself cannot be made publicly available. The techniques and software developed to prepare the data, however, include valuable concepts for working with digitized sheet music and score/performance matching. Several possible application scenarios of the corpus are discussed as a conclusion to the chapter.

Chapter 3 presents results of exploratory research into the corpus with the main focus on performance errors. The phenomenon is ubiquitous for performers, but very hard to investigate due to the lack of data with sufficient precision. The presented studies cover quantitative, qualitative and perceptual aspects of error production. The data having been gathered from live performances, the results complement Repp's previous error study [100], which was done under laboratory conditions. From a selection of pieces in the corpus a catalogue of error groups was built, which made it possible to broaden the view of performance errors from perspective of single-note or very small context to more general groups of errors. As Magaloff was already 77 years old when recording the corpus, this also is an opportunity to inspect possible effects of age on his performances,

especially in comparison with his earlier recordings of the same pieces. We look at the corpus from the point of the *selection - optimization - compensation* (SOC) model, a general model for successful aging, and investigate to what degree it applies to the Magaloff data. Lastly, a study mainly done by W. Goebl is presented as an example of a very specialized research question that can be investigated only with data collections like the Magaloff corpus: temporal onset asynchronies between right and left hand. We look for a specific kind of tempo rubato, recommended by Chopin to his students, which manifests in subtle between-hand asynchronies rather than in tempo.

Apart from its significance for expressive performance research and investigations into performer style and idiosyncrasies, the Magaloff data can be used as training data for expressive performance rendering systems. This is the main focus of this thesis. Chapter 4 presents a novel approach to expressive performance rendering based on a graphical probabilistic model. The first, very simple implementation of the model, the system YQX, was entered to the international rendering contest RENCON 2008, in which performance rendering systems are pitted against each other. Our system won the competition, establishing a basis for us to build on. The system deals with the expressive dimensions note-by-note, without taking into account how the performance evolves. This is not the case in real performances, where large tempo changes are not made spontaneously but gradually. We extend the system to integrate the performance context into the current prediction. To accomplish this, a new algorithm is proposed that also is a closed form solution to an inference problem in a special type of dynamic bayesian network. By modifying the way performance tempo is modeled, namely treating tempo as a composite phenomenon with both slow and fast fluctuating components, tempo prediction is improved further. A similar idea can by applied to the prediction of loudness changes. Performance directives given in the score, like *crescendo* and *decrescendo*, have a considerable impact on the loudness evolution of the performance. As our score model does not include the performance directives, we try to eliminate their effect from the trainings performances, and train our model to predict the part not explained by the performance directives.

In chapter 5 I am going to present the results of experiments comparing the different algorithms and representations of performance dimensions. The problems of automatic evaluation of performance rendering systems are discussed, specifically the use of correlation as a measure of similarity between two performances. The abstraction chosen to represent the score, the score model, has a major influence on the prediction. Which score

model to use depends on the training data, chosen algorithm, and expressive dimension. An experiment is presented searching for the most suitable score representation for the different situations. After that, the merits of the different algorithms for the different expressive dimensions are discussed. As numbers cannot do justice to considerations of musicality, I present both evaluation of quantitative and qualitative aspects of the predicted performance curves.

The last chapter summarizes all presented ideas and contributions and elaborates on future work regarding expressive performance rendering systems, with special focus on their automatic evaluation.

Chapter 2

The Magaloff Corpus

At the age of seventy, I have come to the conclusion that only the sentiment and fear of death can induce an immoderate passion for life

Nikita Magaloff in an interview with Eugenio Scalfari

This chapter introduces the Magaloff Corpus, a collection of recordings of all works for solo piano by Frédéric Chopin, played by Nikita Magaloff on a Bösendorfer computer controlled Grand Piano. Initially the corpus is a collection of recordings in a symbolic form, listing all played notes with precise timing and loudness information. Apart from their artistic, and historic value, the recordings have little scientific significance as such: two notes may account for the same duration of time, but one might be a prolonged quarter note that is part of a ritardando, the other a short half note that is part of an accelerando. Without knowing the nominal description of the notes, we cannot inspect how performances deviate from that, and, after all, those deviations are what constitutes expressivity. In order to make the Magaloff Corpus the valuable resource that it can be, we have to bring the scores of the music and Magaloff's performances together.

Some introductory words on the pianist Magaloff (section 2.1), the concerts that were the source of the recordings (section 2.2), and the recording device, the Bösendorfer SE (section 2.3) precede the main parts of the chapter: preparation of the scores – digitizing and converting the sheet music into symbolic, machine readable representations – is covered in section 2.4; bringing Magaloff's performances and Chopin's scores together – a process called score-performance matching – is discussed in section 2.5. The chapter is concluded with an overview of possible applications of such a collection of recordings (section 2.7). Large parts of the chapter were published in [26].

2.1 Nikita Magaloff

Nikita Magaloff, born on February 21, 1912, in St. Petersburg, was a Russian pianist. As his family was friendly with musicians like Sergei Rachmaninov, Sergei Prokofiev and Alexander Siloti, he grew up in a very musical environment. In 1918, the family first moved to Finland and then to Paris soon after (1922), where Nikita Magaloff started studying piano with Isidore Philipp, graduating from the Conservatoire in 1929 [13, 7].

Magaloff started his professional career mainly in Germany and France, often appearing together with the violinists Józef Szigeti (whose daughter Irène he later married) and Arthur Grumiaux, and the cellist Pierre Fournier. In 1949, he took over Dinu Lipatti's piano class at the Geneva Conservatoire where he continued teaching until 1960. His pupils include Jean-Marc Luisada, Maria Tipo, Sergio Calligaris, Michel Dalberto, and Martha Argerich.

Magaloff is especially known for his performances of the complete works of Frédéric Chopin, which he usually presented live in a cycle of six recitals. He distanced himself from the sentimental interpretations of Chopin's work by the generation of his teacher I. Philipp and especially by Paderewski, who he believed had falsified Chopin. Discarding the sentimental aspects, Magaloff praised Chopin's music as being "incredibly musicianly and full of feeling" [13]. He preferred and recorded Chopin's manuscripts rather than Paderewksi's editions of the scores or the posthumously published versions of the Waltzes by J.Fontana

The first ever recording of the complete works of Chopin was made by Magaloff in the years 1954–1958 for Decca. He repeated this for Philips in 1975. Other than that, only a few studio recordings by Magaloff exist. There exist recordings of the complete etudes of Scriabin (under the Valois/Naíve label), Granados' Goyescas (Decca), several works by Mendelssohn (including Variations sérieuses, Rondo capriccioso, a selection of Songs Without Words, and the Sonata Op. 106), and a live recital from the Salzburg Festival in August 1969 (including Dallapiccola's Sonatina canonica and Ravel's Gaspard de la nuit). The record label Philips included Magaloff in their *Great Pianists of the* 20^{th} *Century* series and published 2 CDs with Liszt's Paganini Études, a sonata by Haydn, Chopin's Sonata Op. 4 in C minor, the Bolero Op. 19, and several Nocturnes and Mazurkas.

Nikita Magaloff died on 26 December 1992, at the age of 80 in Vevey, in the Canton Vaud in Switzerland [13].

2.2 The Magaloff Concerts

Between 1932 and 1991, Magaloff appeared in 36 concerts in the *Wiener Konzerthaus*, one of Vienna's most illustrious concert venues – 24 solo concerts, 10 concerts as orchestra soloist, 2 chamber recitals together with József Szigeti.[1] In 1989, he started one of his famous Chopin cycles in which he would play all Chopin's works for solo piano that were published in the composer's lifetime, essentially Op. 1 to Op. 64, in ascending order. Each of the six concerts was concluded with an encore from the posthumously published work of the composer. The concerts took place between January 16 and May 17, 1989, in the *Mozartsaal* of the *Wiener Konzerthaus*. At the time of the concerts, Magaloff was already 77 years old. Daily newspapers commenting on the concerts praise both his technique and his unsentimental, distant way of playing [108, 110]. Table 2.1 lists the programs of the six concerts.

Although the technology had only been invented a short time before (first prototype in 1983, official release 1985 [74]), all six concerts were played and recorded on a Bösendorfer SE, precisely capturing every single keystroke and pedal movement (see 2.3 for further details). This was probably the first time the new Bösendorfer SE was used to such an extent. In 1999, Gerhard Widmer received written and exclusive permission by Irène Magaloff, Nikita Magaloff's widow, to use the data for our research.

2.3 The Bösendorfer SE

The first model of Bösendorfer's computer-controlled grand pianos, the Bösendorfer SE, developed by Wayne Stahnke, was officially released in 1985 [74]. It is a combination of a standard concert grand piano, in this case a Bösendorfer Imperial, with an array of infrared sensors and motors that can be used to record and reproduce key and pedal movements with very high precision. It was originally intended to facilitate the recording process: a pianist's keystrokes and pedal movements would be recorded, basically a list of note on- and offset events and changes of pedal position. This symbolic data could then be edited much more effectively than an audio recording (selection of passages or deletion of unwanted notes). Afterwards, the reproduction unit of the grand piano would be used to replay the data and record the sound. In case of live recordings this would eliminate unwanted background noise, like coughing in the audience or ringing cell-phones (a need

[1] Information available through the program archive of the *Wiener Konzerthaus*, `http://konzerthaus.at/archiv/datenbanksuche`.

Date	Played
16 Jan	Rondo Op. 1; Piano Sonata No. 1 Op. 4; Rondo Op. 5; 4 Mazurkas Op. 6; 5 Mazurkas Op. 7; 3 Nocturnes Op. 9; 12 Études Op. 10. *Encore:* Fantaisie-Impromptus Op. posth. 66.
19 Jan	Variations Op. 12; 3 Nocturnes Op. 15; Rondo Op. 16; 4 Mazurkas Op. 17; Grande Valse Op. 18; Bolero Op. 19; Scherzo No.1 Op. 20; Ballade No. 1 Op. 23; 4 Mazurkas Op. 24; 12 Études Op. 25. *Encore:* Variations "Souvenir de Paganini" (posth.)
15 Mar	2 Polonaises Op. 26; 2 Nocturnes Op. 27; 24 Preludes Op. 28; Impromptu No.1 Op. 29; 4 Mazurkas Op. 30; Scherzo No.2 Op. 31. *Encore:* Waltz in E minor (posth.)
10 Apr	2 Nocturnes Op. 32; 4 Mazurkas Op. 33; 3 Waltzes Op. 34; Piano Sonata No.2 Op. 35; Impromptu No.2 Op. 36; 2 Nocturnes Op. 37; Ballade No.2 Op. 38; Scherzo No.3 Op. 39; 2 Polonaises Op. 40; 4 Mazurkas Op. 41; Waltz Op. 42; Tarantella Op. 43. *Encore:* Waltz E♭-Major (posth.)
13 Apr	Polonaise Op. 44; Prelude Op. 45; Allegro de Concert Op. 46; Ballade No.3 Op. 47; 2 Nocturnes Op. 48; Fantaisie Op. 49; Impromptu No.3 Op. 51; 3 Mazurkas Op. 50; Polonaise Op. 53; Scherzo No.4 Op. 54. *Encore:* Ecossaises Op. posth. 72 No.3.
17 May	2 Nocturnes Op. 55; 3 Mazurkas Op. 56; Berceuse Op. 57; Piano Sonata No.3 Op. 58; 3 Mazurkas Op. 59; Barcarolle Op. 60; Polonaise-Fantaisie Op. 61; 2 Nocturnes Op. 62; 3 Mazurkas Op. 63; 3 Waltzes Op. 64. *Encore:* Waltz Op. posth. 69 No. 1

Table 2.1: The Magaloff *Konzerthaus* Concerts in 1989

that only arose in the last decade). Performances could be re-recorded in a different environment or with different recording equipment.

Below each set of strings[2] an infrared sensor is mounted. The sensor is triggered twice by a hammer striking the strings: 5 mm before the hammer strikes and upon impact. The

[2]Depending on pitch height one to three strings are tuned to the same frequency.

difference between the two times is used to calculate the velocity with which the hammer strikes the string, and, in extension, the loudness which which the note was played. The sensor operates at 25.6 kHz, which makes 0.0391 ms the shortest measurable interval. This corresponds to a hammer velocity of 128 m/s. The interval length is limited to 40 ms, or 0.125 m/s, a velocity at which the hammer barely reaches the string. Instead of the length of the time interval, or the hammer velocity, the Bösendorfer data contain the *inverse hammer velocity* (IHV), which is defined as $\text{IHV}(h) = \frac{128}{v(h)}$. Typical IHV values for pianists range from $32 - 512$ IHV, which corresponds to $4 - 0.25$ m/s [42]. The following formula is usually used to convert hammer velocities into MIDI loudness values:

$$v(k) = 57.96 + 71.3 * \log_{10}(v(h)), \tag{2.1}$$

where $v(k)$ is the MIDI loudness value of the key, and $v(h)$ is the hammer velocity in meters per second. This relates the allowed MIDI loudness range of $[0, 127]$ to hammer velocities $[0.15, 9.3]$m/s, or IHV values $[832, 13]$. The aforementioned typical range of $32 - 512$ IHV is mapped to MIDI loudness range of $[15 - 101]$. Stahnke proposes a different mapping, that also factors in pitch, assigning a slightly higher loudness to higher pitch than to a lower pitch at the same hammer velocity:

$$v(k) = 52 + 25 * \log_2(v(h)) + (n - 60)/12, \tag{2.2}$$

where n is the MIDI note number (middle C = 60). The differences between this and the standard mapping can be seen in figure 2.1: with Stahnke's conversion map a middle C (MIDI pitch 60) with a hammer velocity below approximately 3.22 m/s is assigned a lower MIDI loudness than with the standard map. The point beyond which Stahnke's conversion assigns higher values (earlier for higher pitches, later for lower pitches) is at the upper end of the typical range. This means that pieces converted with Stahnke's map will on the whole be in a lower MIDI loudness range values than pieces converted with the standard map.

For the most part, the conversion of the Magaloff Corpus into MIDI was done using the standard mapping in equation 2.1. Both Étude collections (Opp. 10&25), the Sonatas Opp. 4&58, and the Préludes Op. 28 were converted using the mapping suggested by Stahnke (equation 2.2).

Each key is equipped with an infrared sensor that measures the time a key is pressed at least 3mm and is released. The impact time of the hammer on the string is used as *note-on* time of this event. Releasing the key registers the *note-off* event. The

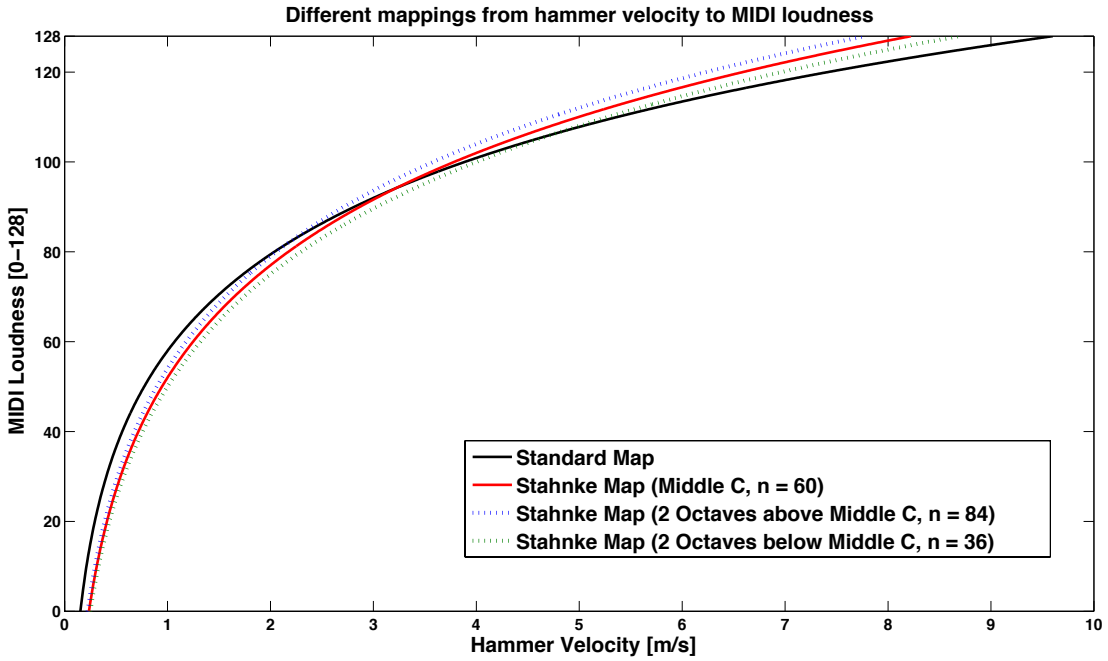


Figure 2.1: Different conversion maps from hammer velocity to MIDI loudness.

sensor resolution of 800 kHz results in a time resolution of $1.25ms$ [42, 43]. The on- and offset times of note events can be converted into MIDI time codes. Pedal positions are discretized to 256 values (8 bit) and recorded with a frequency of $100kHz$ (sustain pedal) and 50 kHz (soft pedal).

2.4 Preparation of the Scores

The recorded symbolic performance data requires careful preparation to become accessible for further investigations. Without any reference to the score, nothing can be said about how specific elements were realized. A lengthened eighth note and a shortened quarter note may account for the same amount of performed time, the former being part of a slower passage in the same piece. Without any information about the notated duration of the note, no assumption can be made about what kind of modification the performer applied to the note.

We need the final state of the corpus to be a piecewise list of all performed notes

aligned with their counterparts in the score. This requires symbolic, computer-readable representations of all scores, which are then aligned to the MIDI data of Magaloff's performances. Given the nature of Chopin's music – high note density, high degree of expressive tempo variation – automatic matching will be error-prone and accordingly, intensive manual correction of the alignment is required. The most intuitive way to view a score is the sheet music itself. Inspecting and correcting an alignment manually is therefore most efficient and tractable using a display of the score page and a piano roll representation of the performance (MIDI) joined together by the alignment. This requires a score representation that contains not only information pertaining to the musical content of the piece but also to the geometrical location of each and every element on the original printed score.

The problems involved in scanning and converting the sheet music into a symbolic, machine readable representation (a process called *Optical Music Recognition* (OMR)) are described in sections 2.4.1 and 2.4.2. Extending the results of the recognition process, which are symbolic scores in musicXML format, with the necessary geometric information is covered in section 2.4.3.

2.4.1 From Score Sheets to Images

The first step in digitising the score is to scan the sheet music. We have no information as to which editions of the scores Magaloff learned the pieces from, so we used the Henle Urtext Editions [143, 135, 140, 133, 142, 137, 139, 134, 136, 138, 141, 144, 57]. Henle does not provide the Sonata Op. 4, and the Rondos Op. 1, Op. 5, and Op. 16. In these cases we used the obsolete Paderewski editions [82, 81]. *Edition Peters* recently (2009) published a new Urtext edition of the complete Chopin, including the Sonatas and Rondos, which will replace the Paderewski Editions in future versions of the corpus.

The quality of the conversion process from image to symbolic score depends on the quality of the scan. The 930 pages of sheet music were scanned in greyscale with a resolution of 300 dpi. Additionally we tried to minimize the skewness of the scans by allowing a maximum of 15 pixels deviation from the horizontal across the complete width of the page. At the chosen resolution, this amounts to 0.5% of the roughly 3500 pixels in height, an angle of 0.006° to the horizontal.

2.4.2 From Images to MusicXML – Optical Music Recognition

Several commercial applications exist for extracting musical content from a scanned score sheet, a process called *Optical Music Recognition (OMR)*. Most prominent among those are *PhotoScore* by Neuratron[3], which is used in the Notation Software Sibelius[4], *SmartScore* by Musitek[5], the *Lite* version of which is used in the Notation Software Finale[6], and the *Liszt-music OCR engine*, which is used in *SharpEye* by Visiv[7]. All three export the results of the scanning process in musicXML format.

The musical content, as recognized by all of the above, however, is not enough. In order to make the alignment process between score and performance feasible, we also need precise layout information of all score elements. Of the three applications mentioned, SharpEye is the only one providing access to the intermediate, internal representation of the analysed page. This information, which includes the geometrical data, is stored in *mro* files, the output format of the underlying Liszt OCR engine. Access to this information was the main reason for choosing SharpEye for this task.

Figure 2.2 shows a screenshot of the program working on Chopin's Prelude Op. 28 No. 1. After the initial recognition process, SharpEye indicates bars that are rhythmically erroneous. In most cases this is a sign that either some notes were not detected or their duration was misread. However, this only covers a small percentage of the encountered problems in the recognition process, several of which are discussed in the following.

Reading and Implementation Problems

In general, the recognition in terms of pitch and accidentals is acceptable. However, situations occur where notes or whole sequences of notes are left out. Figure 2.3 shows an example: The middle voice in the second half of the bar could not be read from the scan and has to be added manually. In addition, the middle voice starting in the beginning of the bar (B♭4) is misinterpreted as a series of sixteenth notes instead of eighths, which is easy to miss both when reviewing the score as well as listening to a mechanical MIDI rendering.

[3] see `http://www.neuratron.com`
[4] see `http://www.sibelius.com`
[5] see `http://www.musitek.com`
[6] see `http://www.finalemusic.com`
[7] see `http://www.visiv.co.uk`

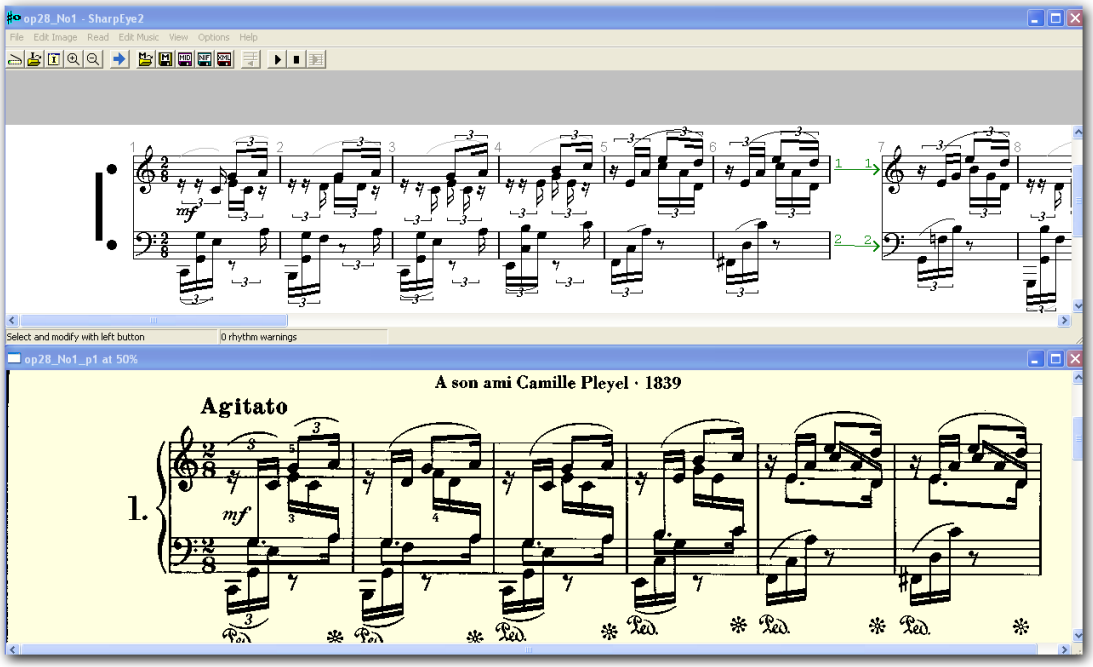


Figure 2.2: The SharpEye OMR software showing the printed score (lower panel) and the manually corrected result of the recognition process (upper panel).

Especially problematic are notes with small heads – appoggiaturas, acciaccaturas (long and short grace notes) and longer ornamentations. Figure 2.4 shows an example taken from the Nocturne Op. 27 No. 2 in D♭ Major where all notes in the upper staff had to be added manually. The sequence presents a further difficulty: the beam connecting all ornamentation notes coincides with a line in the staff. SharpEye completely fails to recognize the upper staff, and, as the software lacks the required editing capabilities, the only way to convert the page is to edit the scanned score sheet with an image processing software.

Another frequent problem are 8*va* lines, dashed lines indicating that certain notes actually have to be played one octave higher or lower (see figure 2.4). SharpEye neither recognises them, nor does it provide means to add them manually. The graphical alignment software discussed in section 2.5 was used to add those to the scores afterwards. The same holds for the brackets used to denote different endings of repeated parts of a

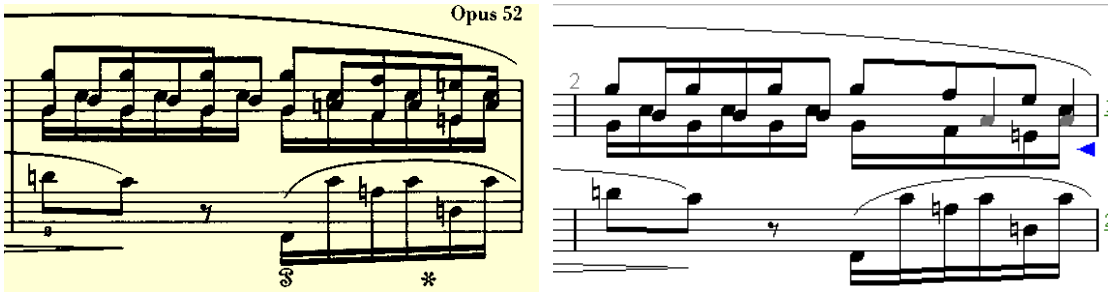


Figure 2.3: *Left*: Printed score of Chopin Ballade Op. 52, Bar 2. *Right*: First result of the recognition process by SharpEye.

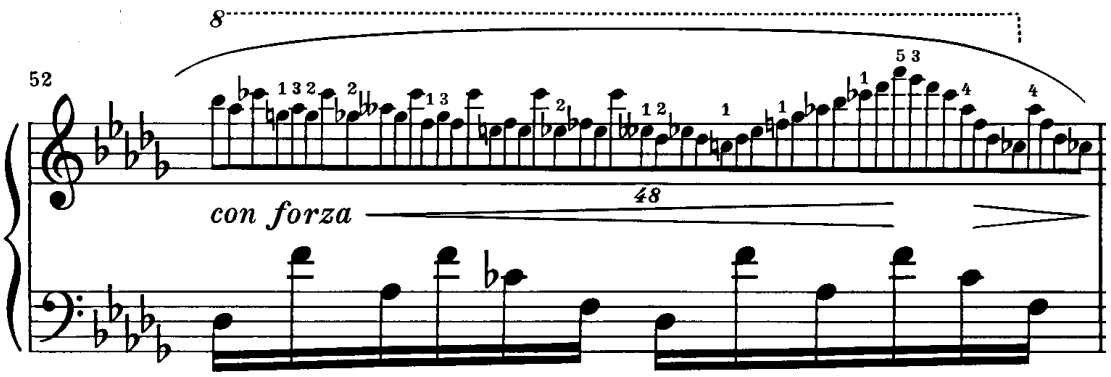


Figure 2.4: Printed score of Nocturne op. 27 No. 2 in D flat Major, Bar 52. The beam connecting the notes in the upper staff coincides with a line in the staff, making the complete staff unreadable for SharpEye.

piece. Out of the several different types of ornamentations (*trills, mordents, schleifer*, etc.) only the standard trill is recognised and provided by SharpEye. As this is not as important for Chopin as it is for other composers like Bach or Mozart, we made no differentiation between different trills in our data.

The recognition quality of expressive score annotations varies. Dynamic changes, indicated by wedges (> (decrescendo), < (crescendo)) are generally recognized well. Verbal dynamic indicators (*cresc., crescendo, dim.*) are treated, and converted into musicXML, like plain text lyrics and have to be replaced by > and < respectively. Loudness directives

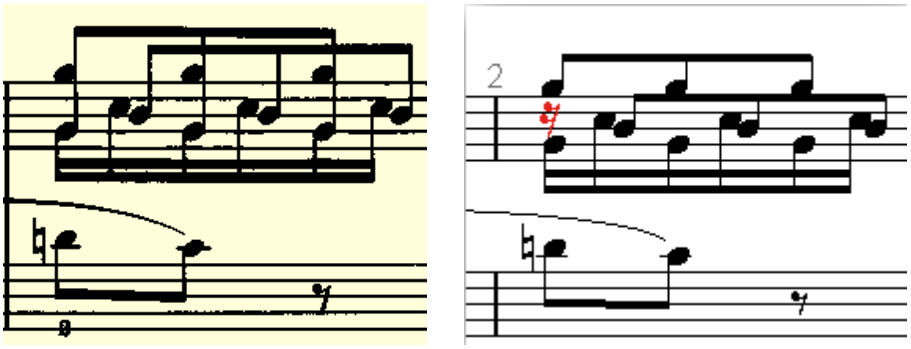


Figure 2.5: *Left*: Printed score of Ballade Op. 52, Bar 2. *Right*: Same bar with an added sixteenth rest (red) to preserve the correct onset of the middle voice.

(*f, ff, p*, etc.) are often not properly recognized, but SharpEye provides corresponding symbols, so that, once corrected, they are converted correctly into musicXML. The graphical alignment software discussed in section 2.5 provides the means to add dynamic annotations more efficiently.

Rhythmic problems

Given a perfectly read and recognized score, a correct interpretation of the content is still difficult. Especially rhythmically complex situations with different independent voices can lead to problems in the conversion. Figure 2.5 shows such a situation: the right panel (SharpEye interface) displays the middle voice in the right hand (starting with a B♭ on the second sixteenth note of the bar) on the correct onset. However, the voice is treated as rhythmically independent of the surrounding voices, which, in absence of any preceding notes, places the beginning of the voice on beat 0 in the bar. In order to have the beginning placed correctly, a sixteenth rest has to be added.

Voices that cross from one staff to the other present a similar problem, as exemplified in figure 2.6: sixteenth rests have to be added in the right hand to preserve the correct rhythmic position of the outer upper voice (𝅀- g - c). The quarter note rest in the left hand was replaced by a 3-tuplet quarter rest and serves the purpose of filling the second triplet in the left hand.

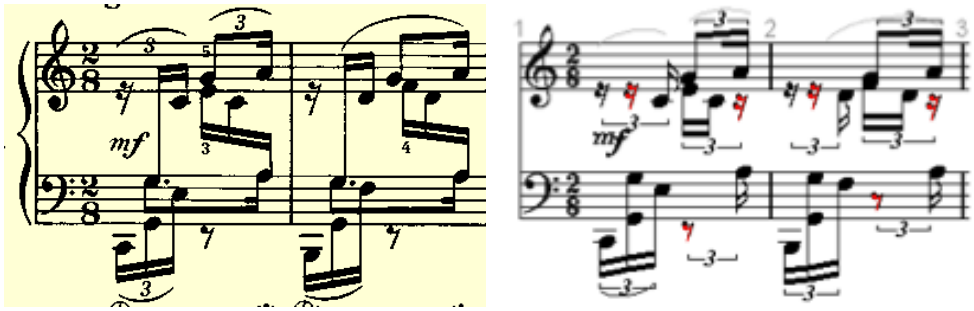


Figure 2.6: *Left*: Printed score of Prelude Op. 28 No. 1, Bars 1&2. *Right*: Same bars with added sixteenth rests (red) to preserve the correct rhythmic position of the outer upper voice. Duplicated notes have been deleted.

Encoding Problems

With our particular applications of the data in mind we made additional alterations to the scores. To emphasise a melody voice or to clarify a situation where voices cross, a note may have two stems with different durations. The sixteenth notes G4 in figure 2.3, and likewise the dotted eighth note G3 in figure 2.6, can be interpreted as expressive annotation or interpretative advice rather than actual note content. Keeping the one with the shortest duration, duplicate notes were removed, as they would bias the error statistics we carry out on the performances (see chapter 3.1).

Common in Chopin's work are figurations and ornamentations, that do not observe the rhythmic grid of the piece. Figures 2.7 shows a typical example: The excerpt from the Nocturne Op. 9 No. 3, written in a $\frac{6}{8}$ meter, is meant to be played without the typical characteristics of a $\frac{6}{8}$ meter and with a very free choice of tempo (cf. the *senza tempo e legatissimo* in the score). While this is obvious to a musically trained eye, from a numerical standpoint the notes do not fit into one bar. One possible solution is to tag all notes after the third beat, starting with the $g4\sharp - f5\sharp - g5\sharp$ chord in the right and the $b2\sharp - b3\sharp$ octave in the left hand, as *grace* notes, which are treated as having zero duration. However, this severely limits their use for investigations in tempo changes. Instead we decided to split the long bar into several bars and use their nominal score durations, as indicated by the dashed red lines in figure 2.7.

This is not always possible and additional changes may have to be introduced. Figure 2.8 shows an excerpt from Nocturne Op. 9 No. 2, where we introduced changes in meter,

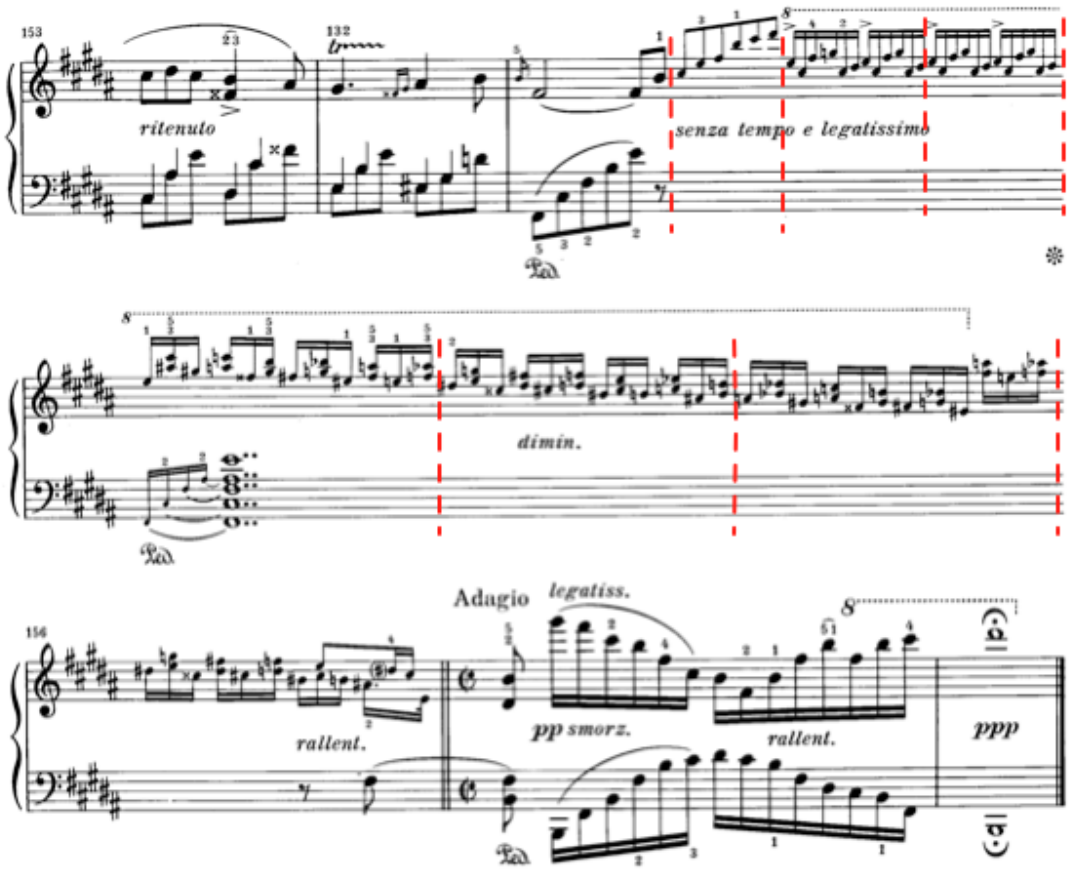


Figure 2.7: Nocturne Op. 9 No. 3 in B Major, Bars 153-158. Additional bar lines have to be introduced to fit the figurations into a rhythmic grid.

from $\frac{12}{8}$ to $\frac{10}{8}$ to $\frac{14}{8}$ and back to $\frac{12}{8}$. From a musicological and interpretational point of view, this of course changes the rhythmical characterisation, but for the purpose of tempo analysis and prediction (see chapter 4) it is more important to maintain the even distribution of note durations across the bars.

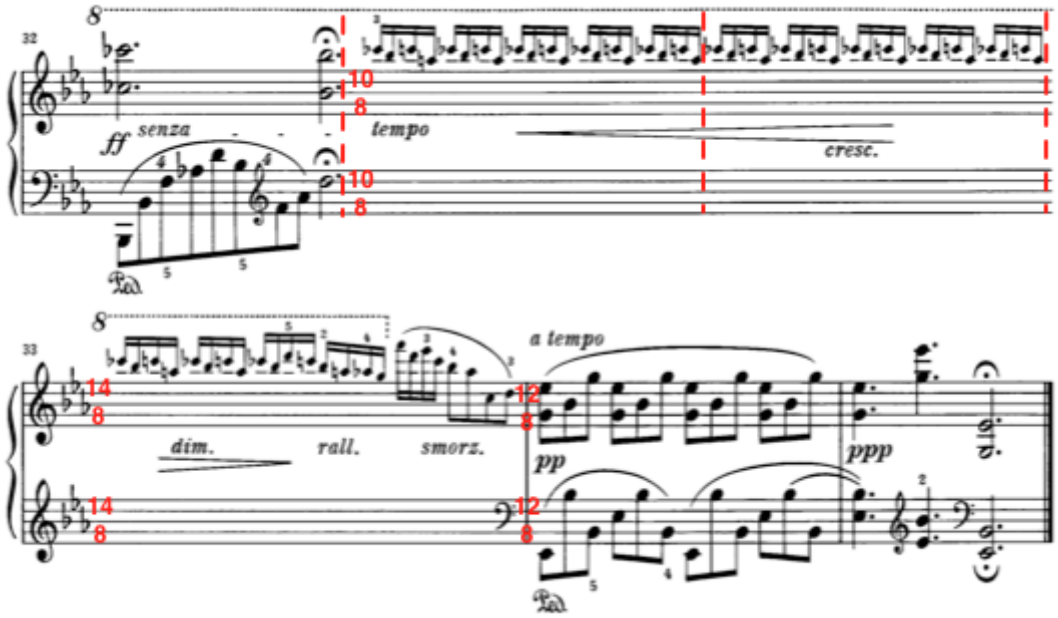


Figure 2.8: Nocturne Op. 9 No. 2 in E♭ Major, Bars 32-37. Additional bar lines and changes of meter have to be introduced to fit the figurations into a rhythmic grid.

2.4.3 From MusicXML to Extended MusicXML

SharpEye exports the recognized and corrected music into musicXML [97], an XML-based, human readable format, originally developed by Recordare LLC for the purpose of interchanging scores between composers and publishers. MusicXML is intended to describe all information – musical content, expressive annotations, editorial information – contained in a score. It does not provide any information as to the layout of the score, which is normally taken care of by rendering engines and manual editing by the music publisher.

However, as the format is text-based and human readable, it is easy to extend it with the geometrical information we need. Mainly, those extensions are:

Measure Attributes: Each `measure` is extended by 2 attributes containing the number of the page and system that hold the measure.

ID: Each `note` is given a unique ID.

Geometric Location: Each `note` element is extended by a node containing its coordinates within the scanned page.

System Location: The geometrical location of each `system` is given in terms of a rectangle with top-left corner coordinates and height and width measurements.

Beat Onset: MusicXML implicitly encodes the onset of a note through order and duration of the elements within the bar. In order to determine the onset of a note or rest all notes and rests preceding the one in question have to be considered. We precalculate and store the nominal onset of each note within the bar for easier access.

As mentioned above, an important aspect of SharpEye for the purpose of this work is that the layout information of the scanned scores is accessible. The information can be exported to a human readable, structured format called *mro*, SharpEye's own internal file format[8].

The information is stored without their musical meaning being interpreted and all recognized elements are described graphically rather than musically: instead of storing pitch names with octave numbers, as done in musicXML, the positions of note heads are relative to the middle line of the staff without taking the clef into account; the duration of a note is stored through both the shape of the head and the number of flags attached to the stem of its chord.

The elements are grouped into the following hierarchy: `page` > `system` > `staff` > `bar` > `chord` > `note`. Geometrical positions are stored for most elements: positions of `system` and `staff` are relative to the page; the position of a `bar` is determined by its right barline, the horizontal position of which is stored relative to the staff. The horizontal position of a note is stored indirectly through the position of the stem of the associated chord and the information on which side of the chord the note head is placed. The vertical position can be calculated from the position of the staff and position of the note head within the staff. Image coordinates are converted from pixel coordinates to units, relative to the size of the staves: the distance between two note lines corresponds to 16 units.

[8]For a full description of the format see `http://visiv.co.uk/tech-mro.htm`.

```
<note>
      <pitch>
             <step>G</step>
             <octave>5</octave>
      </pitch>
      <duration>840</duration>
      <voice>1</voice>
      <type>eighth</type>
      <stem>up</stem>
      <staff>1</staff>
      ...
</note>

<note>
       <chord/>
       <pitch>
             <step>G</step>
             <octave>4</octave>
       </pitch>
       <duration>840</duration>
       <voice>1</voice>
       <type>eighth</type>
       <stem>up</stem>
       <staff>1</staff>
       ...
</note>
```

```
chord {
       <attributes of the chord and the beam>
       flagposn -56,232 headend 2
       notes {
              note {
                     shape Solid staveoffset 0 p -5 accid None
                     accid_dc 0 normalside True
              }
              note {
                     shape Solid staveoffset 0 p 2 accid None
                     accid_dc 0 normalside True
              }
       }
}
```

Figure 2.9: A chord in the musicXML format (left panel) and its counterpart in the SharpEye *mro* format (right panel).

Figure 2.9 shows the same chord represented in two different formats, musicXML and mro. The chord is positioned 232 units (3712 pixel, `flagposn` $-56, 232$) to the right of the top left corner of the staff and consists of two notes: the first one being 5 positions above the middle line, sitting on top of the highest line, which puts the center of the note head 8 pixels (0.5×16) above the top of the staff. The second one is 2 positions below the middle line, on the second line, which puts the note head 48 pixels (3×16) below the top line of the staff. Neither has an accidental and both have a solid head. With the additional information in the chord that the flag count is 1, and information from the surrounding bar structure on the shape of the current clef, this translates to an octave $g4 - g5$ with the duration of an eighth note.

To add the geometric location to the musicXML score, a 1-to-1 matching to the mro-file has to be established. With very few exceptions this is straight forward. In the mro format the notes in a system are stored staff wise: upper staff, left to right, followed by lower staff, left to right. Storage in musicXML is bar wise: first bar, upper staff left to

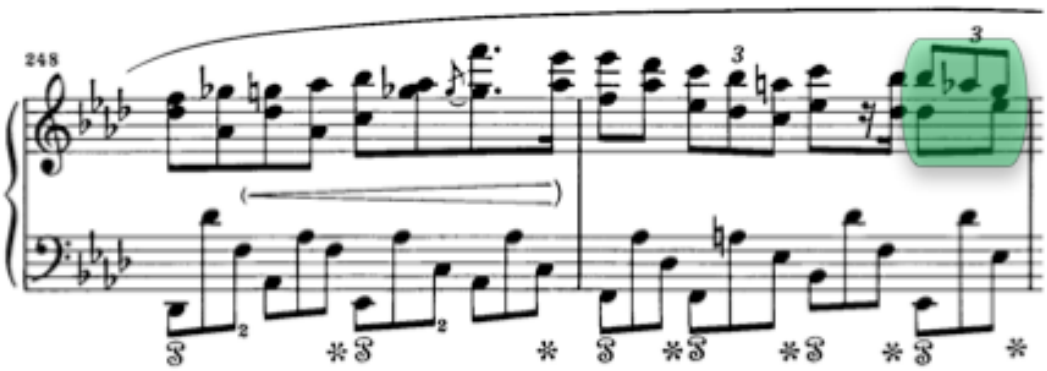


Figure 2.10: Fantasy Op. 49, Bars 248 - 249. The order in which the notes on the last onset in the upper voice (g and $e\flat$) are stored, differs between musicXML and mro format.

right, followed by lower staff, left to right, followed by the next bar. Notes on the same vertical position (mro) or onset (musicXML) are sorted top to bottom. Hence, except at the end of a bar the note to note order is the same in both formats.

Situations exists, however, where the local note order is changed through the interpretation of the graphical content: Figure 2.10 shows two bars from the Fantasy in F Minor, Op. 49. The triplet at the end of the second bar consisting of $b\flat - a\flat - g$ is accompanied by two eighth notes $d\flat - e\flat$, which are printed on the same onset as the first and last note of the triplet. In the mro file, vertically aligned elements are presented from top to bottom, which puts the g before the $e\flat$. Taking the conflicting rhythms into account, however, the $e\flat$ is played before the g which results in a reversed order in the musicXML representation.

2.5 Score-Performance Matching

Score-Performance Matching is the process of aligning score and performance of a musical piece. Depending on the representation of the performance – audio or symbolic data – approaches and techniques differ. In the case of score-audio alignment, a standard approach is to first render the score into an audio file, convert both score and performance

into a feature representation (spectral representation [23], chromagrams [16], or statistics over several frames of audio data (Chroma Energy distribution Normalized Statistics [75])), the sequences of which are then aligned using Dynamic Time Warping, Hidden Markov Models or hybrid graphical models [93]. Applications can be found in both online, real-time situations, like score following [2] and automatic accompaniment [94], and offline situations, like content-based indexing of audio files [23]. A detailed overview can be found in [2].

Aligning a score and a symbolic representation of a performance, such as MIDI, presents different problems: For each performed note the corresponding note in the score has to be marked and vice versa. However, even for expert performers, score and performance rarely match one-to-one. The main differences arise from (1) performance errors, (2) temporal deviations through expressive variations of timing, and (3) underspecified scores (especially trills and other ornamentations) [56]. An important distinction has to be made between *online* matching algorithms, mainly used for score-following purposes [15], that have to meet robustness criteria and real-time constraints, and *offline* algorithms, that need to be as accurate as possible.

The latter, offline score-performance matching with symbolic scores, reflects the situations we need to address with regard to the Magaloff data: the finished corpus should contain an alignment of all scores with their respective performances, such that (1) each score note is either marked as *matched* (and linked to the corresponding performance note) or *omitted* if the score note was not played and (2) each performed note is marked as either *matched* (and linked to the corresponding score note) or *inserted* if the played note has no counterpart in the score.

Several matching strategies have been developed and evaluated[56, 41]. They range from the *strict matcher* [59], which tries a note-by-note matching based on strict temporal constraints derived from the note order in the score, to more sophisticated approaches clustering simultaneously played notes together and trying to find a globally optimal alignment over the complete performance [68]. Gingras [41] presents a matcher that uses structural score information about ornamentations and tracks local tempo changes in the performance to map performance events to the corresponding score events.

We use the *edit-distance* paradigm that was initially invented for string comparisons [122] and has been used in different music computing applications [15, 86]. In [46], Grachten offers more detailed information on score-performance matching as an application of *edit-distance*. Since the edit-distance assumes a strict order of the elements in

the sequences to be aligned, it is not directly applicable to polyphonic music. To solve this problem, we represent polyphonic music as sequences of *homophonic slices* [90], by segmenting the polyphonic music at each note onset and offset. The segments, represented as the set of pitches in that time interval, have a strict order, and can therefore be aligned using the edit-distance. A series of edit operations – insertion, omission, match and trill operations in our case – then constitute the alignment between the two sequences. Each of the applied operations comes at a cost (the better the operation fits in a specific situation, the lower the cost), the sum of which is minimised over the two sequences – score and performance.

2.5.1 jGraphMatch - An Interface for Graphical Score-Performance Matching

Due to the complexity of the music, the large variations of tempo and timing in the performances, and the considerable amount of performance errors, automatic score-performance matching is very error-prone. As the number of notes is vast, the interface for correcting and adjusting the alignment has to be intuitive and efficient. Figure 2.11 shows a screenshot of the Java-Application developed and used for the preparation of the Magaloff corpus. The upper half of the application displays the scanned sheet music one system at a time. As the position of each note element on the page is known, each click on an element in the image can be related to the corresponding entry in the musicXML score. The lower half displays the performance of the piece as a piano roll: pitch on the vertical axes, time on the horizontal. The initial, automatic alignment provided by the edit-distance matcher is displayed and can be manipulated. Score notes without associated performance note are highlighted, as are performance notes that have no partner in the score, and aligned performance and score notes with different pitches. Apart from visual feedback, future versions will also include sonification of the alignment: Constructing a MIDI version of the alignment by playing all pairs of matched notes at the time of the nominal (score) onset but with the pitch of the matched performance note provides a further integrity check.

The software was enhanced to also provide corrections of certain limitations of SharpEye, the software used for constructing the musicXML scores, mainly adding 8*va*-lines and Volta brackets[9]. Expressive annotations (*crescendo, diminuendo, accelerando* and

[9]Volta Brackets denote that certain passage are to be played differently on different repetitions of the same segment.

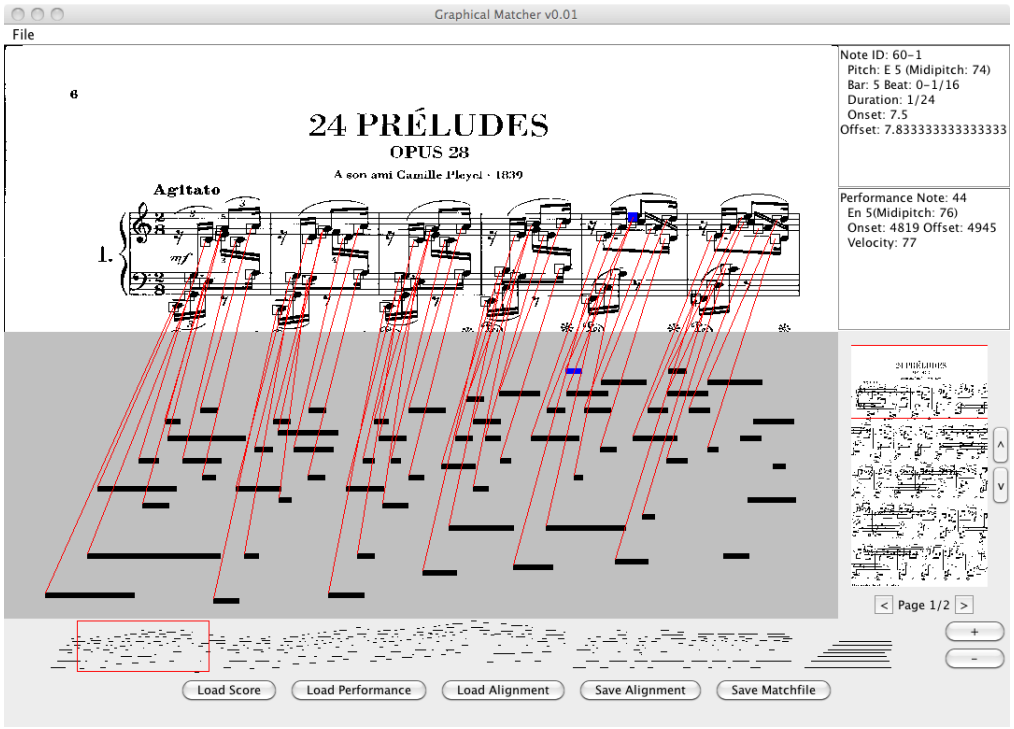


Figure 2.11: jGraphMatch: a Software tool for display and manual correction of score-performance alignments.

ritardando) and dynamics (*p,f,sfz*, etc) can be added and their on- and offsets modified. Possibilities for classifying errors into error categories (as done in the error studies presented in section 3.1.5) were also integrated into the software.

A particular problem with the matching was that in some pieces there are differences between our version of the score and the version performed by Magaloff: this ranged from small discrepancies where, e.g., Magaloff repeats a group of notes more often than written in the score (e.g., in the Nocturne Op. 9 No. 3, bar 111), to several skipped measures (e.g., Waltz Op. 18, where he omitted bars 85 to 116), to major differences that probably are the result of a different edition of the score being used by Magaloff (e.g., in the Sonata Op. 4 Mv. 1, bars 82 to 91, where the notes he plays are completely different from what is written in the Paderewski edition of the piece).

2.6 Overview of the corpus

Magaloff's performances comprise 336.581 notes. Chopin's compositions comprise 331.080 notes. From those we exclude passages where differences between our score and the score version performed by Magaloff seemed likely. To this end, we apply the a simple heuristic: sequences with more than 10 successive insertion notes and no matched notes are excluded, as are sequences of deletion notes spanning more than 4 onsets with no match notes in-between. The excluded passages amount to 2808 score notes and 1859 performed notes. Unmatched performance notes with a MIDI velocity below 10 are also excluded (731 performed notes). Table 2.2 shows a summary of the corpus. The remaining score and performance notes fall into three categories: *Matched Notes* (score note/performed note pair), *Inserted Notes* (performance notes without corresponding score notes), and *Omitted Notes* (score notes without corresponding notes in the performances). *Substituted Notes* are matched notes, where the pitches differ between score and performance. A more detailed discussion of the different kinds of errors can be found in 3.1.1. Grace notes and trills are mentioned separately: Grace notes do not have a nominal duration defined by the score. Therefore they cannot contribute to discussions of temporal aspects of the performance. As a consequence we normally exclude those from the data. Trills constitute many-to-one matches of several performance notes to a single score note. When counting the performance notes in the corpus, the number of performance notes matched to a trill have to be accounted for. Accordingly, the complete number of performed notes is composed of the number of matches, substitutions, insertions, matched grace notes, and trill notes. The complete number of score notes is composed of the number of matches, substitutions, omissions, and matched and omitted grace notes.

Table 2.3 shows the note and matching statistics according to piece categories. The generic category *Pieces* includes: Introduction and Variations Op. 12, Bolero Op. 19, Tarantella Op. 43, Allegro de Concert Op. 46, Fantaisie Op. 49, Berceuse Op. 57, Barcarolle Op. 60, and Polonaise-Fantaisie Op. 61. The encores have not yet been included in the corpus.

2.7 Applications of the Corpus

Plenty of possible applications of a corpus of this precision and dimensions exist, some of which I will glance at in conclusion.

Pieces/Movements	155
Score Pages	930
Score Notes	328.272
Performed Notes	333.991
Playing Time	10h 7m 52s
Matched Notes	307.680
Inserted Notes	10.769
Omitted Notes	10.522
Substituted Notes	5.330
Matched Grace Notes	4289
Omitted Grace Notes	451
Trill Notes	5923

Table 2.2: Overview of the Magaloff Corpus. The numbers do not include passages where differences between our score and the score version performed by Magaloff seemed likely. Unmatched performance notes with a MIDI velocity below 10 are also excluded.

First, and foremost, the corpus is a unique document of one of the great pianists of the 20^{th} century. It offers perspectives on Magaloff's style, his interpretational ideas, his image of Chopin's work. In the course of this thesis – chapter 3 investigates Magaloff's performance errors, compares his interpretations of Chopin's Études to those of other pianists in search of effects of age, and inspects how Magaloff uses asynchronies between right and left hand as an expressive device – I only scratch at the surface of what can possibly be learned from this collection.

Apart from Magaloff's own particular style, this collection is also representative of piano performance in general, and can as such be used to validate already existing computational models of expressivity. While it is possible to a certain extent to mark beat times in audio recordings manually, and with the so measured data test and develop models for tempo changes, it is impossible to do this for loudness of individual notes or even onsets. The Magaloff corpus provides loudness information for individual notes, which makes it possible to study loudness not only in its general shape over larges units, like phrases, but also in surgical detail. An example for the latter are the relative loudness values of different notes in a chord, which have substantial influence on the timbre of the chord. This also applies to offsets of individual notes which, in case of audio recordings,

Category	Pieces	Score	Played	Matched	Inserted	Omitted	Substituted
Ballades	4	19511	20160	18975	911	469	274
Etudes	24	40894	40757	38697	1449	1620	609
Impromptus	3	7216	7300	7152	81	154	67
Mazurkas	41	47156	46844	45119	963	1608	521
Nocturnes	19	31108	31974	30961	595	838	318
Pieces	7	39759	40926	38281	1497	1395	976
Polonaises	7	27875	28161	26246	1372	1103	510
Preludes	25	20067	20172	19242	586	605	344
Rondos	3	18250	18301	17491	287	433	304
Scherzi	4	21957	22543	20861	1236	670	407
Sonatas	12	38631	39886	36694	1385	1413	756
Waltzes	8	18656	18826	18173	407	678	244

Table 2.3: Overview of the Magaloff Corpus by piece category.

are in most instances next to impossible to determine automatically or manually. Most articulation-related aspects are therefore accessible only through data-collections like the Magaloff corpus.

All machine learning applications, and, for that matter, all data-driven enterprises, need representative data. One of those applications, expressive performance modeling, is the main focus of this thesis. Without data as precise and representative as the Magaloff corpus such an endeavor would not be possible. Arzt [2] uses the data as ground truth and evaluation criteria for his score following system. As soon as audio recordings are made from the symbolic part of the corpus, the data can been used as training data for machine learning algorithms to automatically transcribe audio data into symbolic music.

Audio-to-Score alignment is a form of automatic transcription of audio into symbolic music. Based on complete knowledge of the score, the task is to extract each score note's position within the audio recording [80]. The Magaloff corpus can serve as ground truth data for training as well as evaluating alignment systems. Given reliable automatic transcriptions, the long term objective is to build corpora of symbolic performance data for other pianists, in order to investigate inter-artist differences and commonalities. Therein the Magaloff corpus can play the role of a "master corpus": for most performances of the same piece the scores are practically identical. If for all bars or even onsets in a

performance the corresponding bar or onset in the Magaloff corpus can be identified, any information and manual annotation available in the Magaloff corpus can also used for the analysis.

Chapter 3

Inspecting Magaloff

If you want me to play only the notes without any specific dynamics, I will never make one mistake.

Vladimir Horowitz

The Magaloff corpus and similar collections open up possibilities for new strands of performance research. The following chapter touches on three aspects of music performance: (1) Performance errors, a ubiquitous element of a musician's life, have been studied before under laboratory conditions by Repp [100]. Section 3.1 complements the study with a perspective on the performance errors of a professional pianist during a live recital. The size of the corpus makes it possible to abstract from the view of single note errors and establish error categories. (2) Section 3.2 presents a study investigating the effect of age on a performer. (3) Section 3.3 briefly reports on a study done mainly by W. Goebl on the Magaloff corpus (originally published as [45]) that extends previous work on temporal asynchronies in music ensembles to the domain of solo piano.

3.1 Performance Errors

Musicians at all levels of proficiency must deal with performance errors[1] and have to find strategies for avoiding them. As their level of skill increases, errors occur less frequently and also mostly seem to go unnoticed by the audience. Which of the errors the audience in fact notices, depends on three main factors: (1) how exposed the note is within its context, (2) how saliently a wrong note was played, and (3) the listener's musical abilities and acquaintance with the piece. A recent neuro-imaging study discovered a correlation between the pianists' EEG patterns and their performance errors, suggesting that they know 70ms in advance when they are going to make a mistake [103]. So it seems that,

[1] As will be explained in detail in section 3.1.1, this means pitch differences between score and performance.

in theory, the pianist can influence how subtly or saliently a wrong note is played, and anecdotal evidence claims that they in fact do.

While audio recordings abound, extracting information from them related to timing, dynamics, and articulation automatically is still not possible at the level of precision required for large-scale music performance studies. Current techniques for audio transcription focus on extracting pitches and their respective onsets. Although the overall precision is promising, parts with low intensity or extensive use of the sustain pedal are still not sufficiently well recognized [80]. Extracting the dynamics of individual notes from audio recordings is virtually impossible. On the whole this makes audio recordings unusable for studying performance errors. The Magaloff corpus is an ideal source for investigations into the matter, because it contains precise information about which notes were played and which were left out. This allows an exact evaluation of pitch errors in the played pieces.

Following an assessment of quantitative aspects of Magaloff's errors, I am going to relate them to performance tempo (section 3.1.3). Section 3.1.4 then presents the results of a study published in [28], that focuses on perceptual aspects of single-note errors. Many factors contribute to producing an error, among them technical deficiencies, lack of concentration, and poor memorization. While accidentally hitting a wrong note might be a local event, technical problems of course persist and may resurface in similar situations, leading to errors that are systematic. The same holds true for problems with memorization, which may lead to similar or identical sequences containing similar or identical errors. To investigate this further, I propose a categorisation of errors (published in [32]) in section 3.1.5.

3.1.1 Definition of Performance Errors

The first question to be dealt with is what we consider an error in a performance. Deviations from the notated score come in many facets – tempo, note duration, dynamic changes, note order, pitch. However, with the exception of pitch (and even this is only true for instruments where pitches are fixed, like on the piano, and not subject to expressive variation, like on the violin), it is hard to draw a line where deviations stop being perceived as interpretation and start sounding wrong. A criterion commonly used to express what is "right" and what is "wrong" in musical performance is "faithfulness to the score", a term that combines a plethora of different aspects of performance and does not lend itself to an easy definition [114]. Pitch is the one aspect that is usually

agreed on in western piano music (at least for a given edition of the score), and can also be easily measured in our data. This makes pitch errors an objectively definable type of performance error.

Following Repp's definition [100], three types of pitch errors are distinguished: *omissions*, score notes that are not played in the performance; *insertions*, played notes that are not written in the score; and *substitutions*, notes that are played on the wrong pitch. Figure 3.1 illustrates the different situations: in the right hand the notes $f\sharp - b$, $f\sharp$ (marked as red squares in the score part of the figure) were omitted, as were g, c in the left hand. A couple of notes were inserted (red bars in the performance part of the figure) and the $c\natural$ was substituted with a $c\natural$ one octave higher (circled in green).

Taking only insertions and omissions, the issue is well-defined. However, allowing errors to be labeled as substitution, leaves certain decisions to interpretation. The $a\sharp$ that was inserted alongside the octave $b5 - b4$ in the right hand on beat 4.5 (circled in red in the lower part of figure 3.1) does not seem to be a substitution for the omitted $f\natural$ "to the same degree" as the substituted $c\natural$ in the left hand. Of course, it is not possible to say with certainty which notes were played as a substitute for something written in the score, either by accidentally hitting a wrong note or because of incorrect memorisation or reading, and where a simultaneous but independent insertion and omission took place. To assess the issue objectively, for every omitted score note s we search for a "suitable" insertion note p that satisfies the following two conditions: (1) p is played closer to the average onset of all notes with the same score onset as s than to the average onset of the notes preceding or succeeding s (maximum allowed time difference 1 second), and (2) p it is closer in pitch height to s than all other insertion notes that satisfy first condition.

3.1.2 Related Work

One of the few studies on the phenomenon of performance errors was done by Repp in 1996 [100]. The study focuses on pitch errors and their perceptual salience for listeners. The assumption, based on informal observation and performer's accounts [109], is that skilled performers avoid mistakes that are obvious, and the committed errors are difficult to hear. This leads to the following hypotheses about both perceptibility and distribution of performance errors [100]:

1. Errors in inner voices are less noticeable than in the melody voice and hence will occur more frequently.

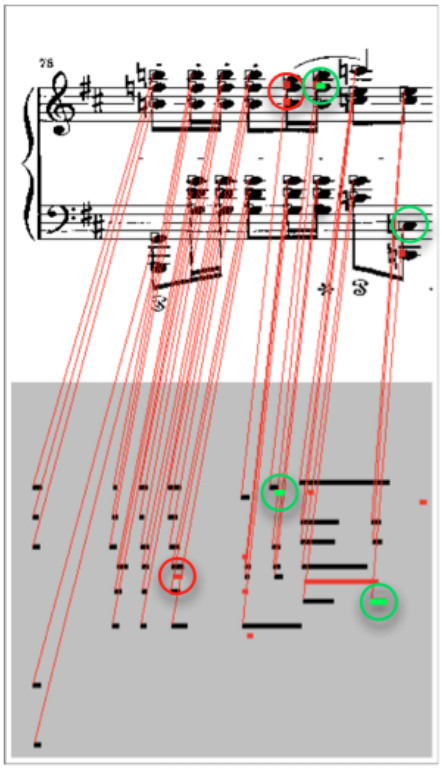


Figure 3.1: Polonaise Op. 40 No. 1, Bar 78. Different Error situations: Insertion (red circle, lower panel), Omission (red circle, upper panel), and Substitution (green circles, upper and lower panel)

2. An increasing number of simultaneously played notes makes it more difficult to hear single voices, and, consequently errors. This is will make errors more frequent in proportion to the number of simultaneous onsets.

3. Insertion and substitution errors will be less noticeable when they fit into the harmonic context than when they are in conflict with it, hence the majority of errors will be harmonically appropriate in their context.

4. Inserted and substituted notes are less likely to be noticed when they are low in relative intensity, hence erroneous notes will be lower in intensity than the correct notes in immediate vicinity.

After a short rehearsal period, ten graduate piano students played four piano pieces repeatedly on a Yamaha Disklavier:

- Op. 15, No. 7, *Träumerei*, by Robert Schumann
- Prélude No.8 (Book I), *La fille aux cheveux de lin*, by Claude Debussy
- Op. 28 No. 15, *Raindrop-Prelude*, by Frédéric Chopin
- Lyric Piece Op. 43 No. 5, *Erotik*, by Edvard Grieg

All pitch errors in the recorded MIDI data were identified and labeled. The findings support the stated hypotheses with respect to the distribution aspect: errors occurred mainly in subsidiary voices (1), where many notes coincided (2), were for the most part harmonically appropriate (3), and low in intensity (4). In a listening experiment, musicians, partly acquainted with the pieces in question, would then try to detect the errors. From the low number of actually detected errors, Repp concluded that most errors are indeed perceptually inconspicuous and only a very small fraction is likely to be noticed by a concert audience.

Several aspects of the hypotheses had been inspected and confirmed before. Especially (1) and (3) were discussed by Palmer et al. [84, 83, 85]. They focus on aspects related to the production of errors rather than the perception, and investigate how performance errors may shed light on the way performers memorize and organize music in their memory. They attribute their experimental corroboration of hypothesis (1) to the effect of conceptual prominence of melody over accompaniment: melodic elements may be retrieved first from memory, followed by all accompanimying voices together. This makes it less like to confuse the melody with other items and increases the likelihood of errors in the chordal elements. Further, the authors reason that hypothesis (3), which they also see supported by their experiments, "suggests that retrieval of musical elements from memory reflects multiple structural levels and units" as opposed to a retrieval in single note units [84].

3.1.3 Quantitative Results

We counted all instances of the different types of errors in all 153 pieces in the Magaloff corpus and computed the error rates with respect to the overall number of performance and score notes (Table 3.1): 3.22% of all performed notes are insertions, 3.21% of all score

Score Notes	328.272	
Performed Notes	333.991	
Matched Notes	307.680	
Inserted Notes	10.769	3.22%
Omitted Notes	10.522	3.21%
Substituted Notes	5.330	1.70%

Table 3.1: Overview of the errors and error rates in the Magaloff Corpus

Figure 3.2: *Left panel* Error percentages by piece category. *Right panel:* Correlation coefficients between note-density and error rate by piece category

notes were omitted, and 1.7% of all matched notes were played at a wrong pitch. This exceeds the percentages reported in (Repp, 1996) (1.08% insertions, 1.64% omissions, and 0.26% substitutions). Looking only at the Chopin piece Repp used in his study (Prelude Op.28/15, 1506 performed notes, 1521 score notes), we encounter error rates that are more similar: 0.6%/1.58%/0.54% (Magaloff) vs. 0.98%/1.48%/ 0.21% (Repp).

Figure 3.2 gives an overview of the categories of pieces with their respective sizes and error numbers. Assessing the pieces by category, the Scherzos and Polonaises stand out in terms of insertion errors (5.5% and 4.9% respectively), the Rondos, Impromptus, and Nocturnes constitute the low-insertion categories (insertion rate below 2.0%). The Impromptus are also the category with the lowest percentage of deletion errors (2.13%), while Etudes and Polonaises exhibit the highest percentage of deletions (around 4%).

It might be, that faster and denser pieces give rise to a larger error proportions then

slow pieces, suggesting that slower pieces are easier to execute. Thus, as a measure for note density, we count the number of notes per 3 second time unit and consider the number of errors in this context. We found that a high note density goes along with a higher error frequency: the more notes played per time unit, the larger the proportion of errors. This holds to a varying degree for all kinds of errors. Overall, the corpus exhibits correlation coefficients between note density and frequency of (a) insertion errors, (b) deletion errors and (c) substitution errors of 0.42, 0.27, and 0.61, respectively. Figure 3.2 shows the correlation coefficients of error frequency and note density for the respective categories of pieces. The Ballads and Polonaises both show a high error percentage as well as a high correlation of error frequency and note density. This may indicate that these are technically particularly demanding.

3.1.4 Qualitative and Perceptual Results

One of Repp's main hypotheses is that skilled musicians avoid errors that are obvious [100]. Whether an error is conspicuous is closely related to several factors:

1. How loud was an added note played in relation to the other notes in the vicinity?

2. How well does an insertion/substitution note fit into the harmonic context, or how important was an omitted note for the harmony?

3. Is the error located in a melody or an inner voice, and how many simultaneous voices surround it?

To assess the first factor, we compared the loudness of each insertion note with the loudness of the correct notes in the immediate vertical vicinity (notes on the same score event). On average wrong notes are inserted at 59% of the volume of the loudest, correct note on the same onset. Only 7% of the inserted notes are the loudest in their context, 16% are louder than the average note with the same onset. 28% are inserted at less than 40% of the maximum loudness, 19% at less than 40% of the average loudness on the same onset. In terms of absolute MIDI loudness, the average insertion loudness is 45.5. In Repp's study, insertion notes are reported to be mostly "of relatively low intensity" [100]. This seems only to be true to a certain extent in Magaloff's performances.

Considering the vast number of errors, assessing the harmonic appropriateness of substitution and insertion errors by listening, as done in [100], is not feasible. Instead,

we estimated the consonance of an insertion/substitution error with respect to the local harmony automatically. Temperley [116] derived key profiles for major and minor scales from the Essen Folksong Collection [105] that rate the probability of occurrence of pitch classes within the context of a given harmony. We used these profiles to determine automatically the most likely local harmony given the pitches that were identified as correctly played. To judge the consonance of an erroneous note within an estimated local harmony, we used the key-profiles proposed by Krumhansl and Kessler [67]. The profiles were established via probe-tone experiments and rate how well a pitch class fits into a harmonic context. The normalized values range from 0.0534 for the least consonant (minor second upwards from the tonic) to 0.1522 for the most consonant (tonic) pitch classes in a scale. Assuming that pitches with a value below 0.0603, the value of the tritone in the major scale, are perceived as harmonically inappropriate, 51% of all insertions and 44% of all substitutions are incompatible with the local harmony. This, of course, is only a crude approximation, which completely ignores that, apart from fitting the local harmony, pitches have additional functions (lead to different harmonies, act as anticipation preceding harmonic relief) which can make them absolutely appropriate regardless of the consonance value assigned above.

Repp reports lower percentages of insertions and substitutions that he judged harmonically jarring: 31% of the insertions and 16% of the substitutions in Schumann's *Träumerei*, 36% of the insertions and substitutions in the Debussy Prélude, and 40% and 25% in Grieg's *Erotik*. The numbers in the Chopin Prélude are very different between the *piano* and the *forte/fortissimo* parts of the piece: in the soft sections (74 bars in total) 34% of insertions were harmonically inappropriate, in the loud sections (15 bars in total) 84%, most of which were very low in intensity. Magaloff's errors do not follow that particular pattern, but are mainly located in the last section of the piece.

The third facet stated above remains to be assessed: Are errors mostly located in inner voices as opposed to melody voices, and do they mainly occur in situations where many notes sound simultaneously? We extracted the melody voice of the pieces, assuming that the highest pitch in the upper staff at any given time is the melody voice of the piece[2], and then calculated the error proportions for the different parts separately. The omission rate in the melody voices is 0.8% as compared to 4.1% for voices not belonging to the melodic part of the score. Table 3.2 shows the error rates by staff and

[2]In the case of Chopin, this very simple heuristic is correct often enough (though not always) to be justifiable. Not only do we not have the resources to manually identify and label all melody notes in Chopins complete piano works, but also is the concept of *melody* not unequivocally defined.

	IR[%]				**OR**[%]			
	Multi		Mono	All	Multi		Mono	All
Upper Staff	4.10	>	2.91	3.57	3.77	>	0.82	2.40
Lower Staff	3.15	>	3.04	3.10	5.46	>	2.27	4.17

Table 3.2: Insertion (IR) and Omission Rates (OR) by staff and surrounding musical texture (Staffwise without regard to musical texture (All), only one score note present in the staff (Mono), and Multi-voiced (Multi)

surrounding musical texture: The staff-wise insertion (omission) rates (columns *All* in the table) were calculated relative to the overall number of performed notes (score notes) in the respective staff. To assess errors with respect to the surrounding musical texture, we defined an onset to be mono-voiced (relative to a staff) if there is only one score note present, and multi-voiced otherwise. The texture-specific insertion (omission) rates (columns *Multi* and *Mono*) are calculated relative to all performance notes (score notes) present in the respective texture and staff. Overall, in the upper staff, insertions are more likely to occur than omissions and vice versa in the lower one. For omission errors the difference between multi-voiced and mono-voiced situations is obvious: A single note is less likely to be omitted, particularly if the note is located in the right hand. Insertion errors are slightly more likely in multi-voice situations, but the difference is less striking.

3.1.5 Error Categorization

While some errors are just local accidents – a finger "brushing" a note alongside a chord, or hitting a wrong note due to a short lack of concentration caused by an irritation in the audience – some errors show regularities, follow patterns, share commonalities. These might be caused by memory problems, such that similar passages contain similar errors, or technical deficiencies, such that certain errors become systematic.

For a part of the corpus (4 Ballades, 24 Préludes op.28, 24 Études opp.10&25, 17 Nocturnes) I categorized errors (single errors or groups of errors) manually into error patterns. Several distinct error patterns emerged, covering roughly 40% (36% of the insertion notes, 44% of the deletion notes, and 44% of all substituted notes) of the errors in the pieces examined. The remaining errors could not be distinguished further. Table 3.3 shows the error categories with their respective error counts. The different categories

Category	Insertions	Omissions	Substitutions
Omitted Inner Voice	-	630	-
Forwards Related Errors	59	9	40
Backwards Related Errors	75	8	53
Unharmonic Errors	694	-	88
Harmonic Errors	104	-	69
Tied Notes	91	294	-
Repeated Notes	123	-	-
Systematic Errors	228	555	110
Note Order Errors	-	-	261
Total	1385	1496	635

Table 3.3: Number of errors in the different categories

are explained in the following.

Forward- and Backward-Related Errors

Errors in this category have a clear forward or backward relation. Figure 3.3 shows typical examples: the insertion in the Nocturne Op. 9 No. 2 (left panel) is caused the pitch B♭ in the immediately following chord; in the right panel (Nocturne Op. 9 No. 3) the opening melody note d in the first of the displayed bars is unaccompanied by the right hand, which probably caused the omission of the right hand chords accompanying the following two melody notes $c\sharp - d\sharp$. Analogous situations occur for substitutions with both forward and backward relations. In almost all cases, the most probable cause is a memorization problem.

Repeated Notes

A special form of backward-related insertions are *repeated notes*, notes that were (most likely unintentionally) played twice, or notes that were first played too soft and then re-struck. In many cases, one of the performed notes is much softer than the other one. Possible causes include a silent finger change on the particular note where the finger was lifted too high in the transfer, thus striking the note twice. Figure 3.4 (left panel) shows a typical example.

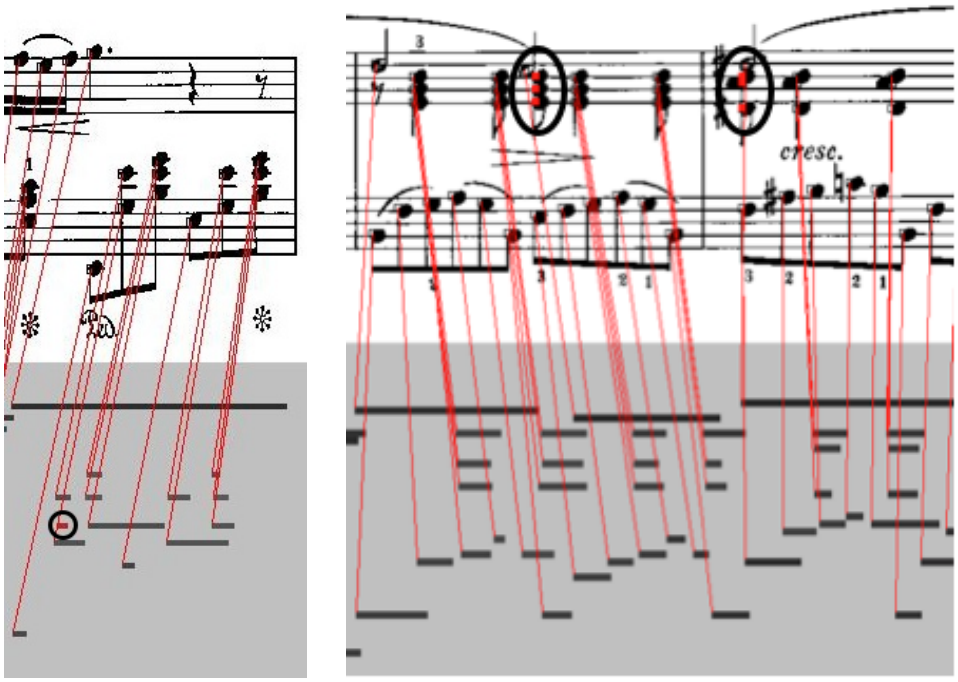


Figure 3.3: *Left panel*: Forward related insertion in Nocturne Op. 9 No. 2, Bar 29; *Right panel*: Backward related deletion in Nocturne Op. 9 No. 3, Bars 89-90

Tied Notes

Two kinds of errors are related to the concept of tied notes: (1) A tied note might be struck again, resulting in an insertion note (figure 3.4, middle panel); this is either a problem of memorization (mostly in inner voices) or done intentionally to emphasize a melody line or harmonic component that otherwise lacked continuation. (2) Two successive notes of the same pitch might be played only once, as if they were notated as tied, resulting in an omission of the second note. Especially in fast passages this suggests the need for technical simplification. Figure 3.4 (right panel) shows a very soft part in the Ballade Op. 52, where tying the $A\flat$ might help emphasizing the melody without letting the inner voice get too loud.

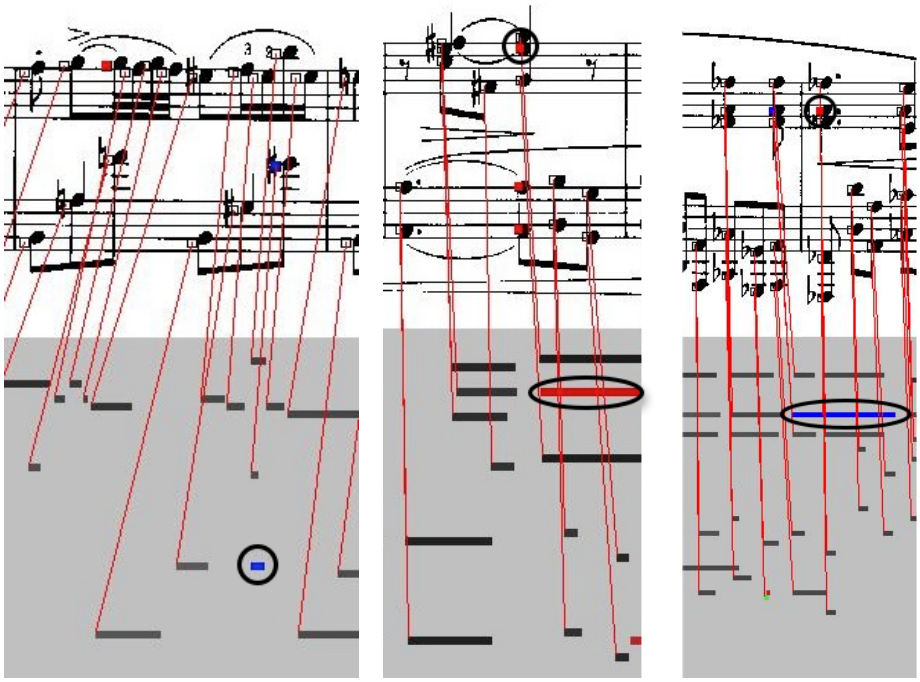


Figure 3.4: *Left panel*: Repeated Note in Nocturne Op. 9 No. 3, Bar 23; *Middle Panel*: Replayed tied note in Prélude Op. 28 No. 22, Bar 29; *Right panel*: Incorrectly tied note in Ballade Op.52, Bars 44-45.

Unharmonic Errors

Errors that obviously disrupt the harmonic context were classified as unharmonic. This mainly involves insertions one semitone above or below the notated pitch. Nearly half of those occurred in octave runs in either one or both hands. Figure 3.5 shows a passage from Nocturne Op. 48 No. 1.

Harmonic Errors

Insertion or Substitution notes associated with this category do not disrupt the harmonic context of the piece. In most cases, these are added octaves in the accompaniment or accompanying notes that were shifted by one octave. While the latter points to a memorization problem, the former could also be deliberate harmonic emphasis. Rare

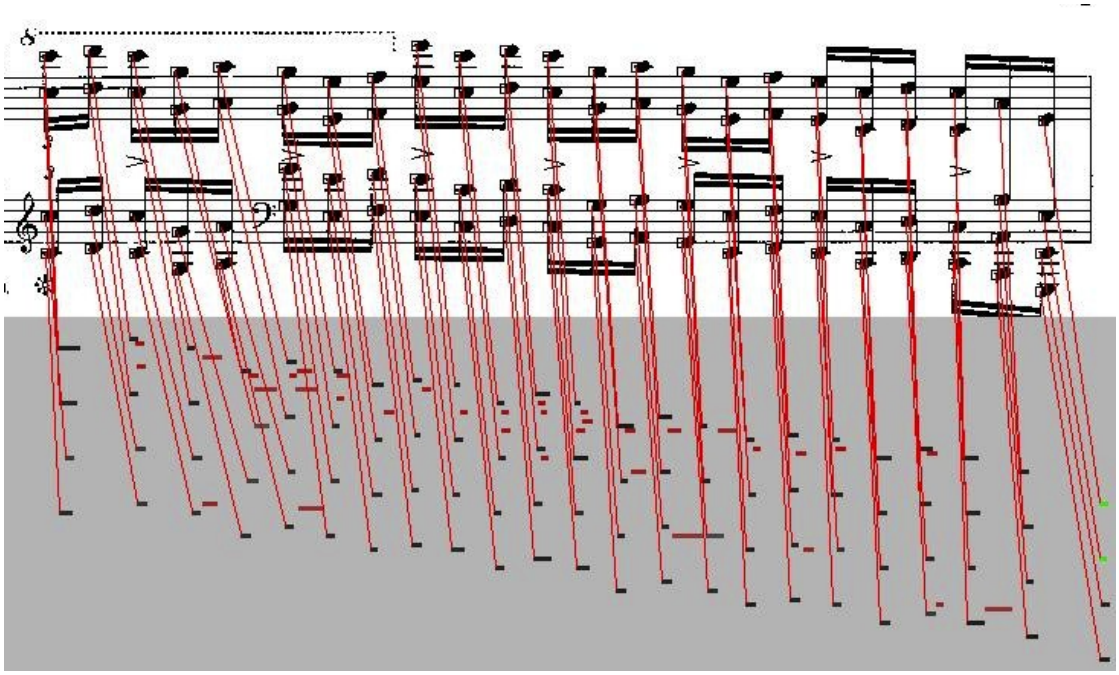


Figure 3.5: Sequence of unharmonic insertions in Nocturne Op. 48 No. 1, Bar 46.

cases involve added figurative elements, such as trills, that were not notated in the score.

Systematic Errors

I call an error systematic if it occurs in more than 60% of instances of the same or an analogous context. This covers a variety of situations. Figure 3.6 shows a systematic insertion from Ballade op.38: In almost all instances where the right hand starts with a downward run accompanied by a rising sequence of octaves in the left hand (e.g. bars 46, 48, and 50), Magaloff inserted a note shortly before or after the first octave in the left hand. This could be an indication of the Limburg-Comstock Syndrome, a condition found mainly among writers and musicians that can cause the last two phalanges of thumb and index finger to link together [123]. As an effect, when playing octaves, the index finger can not be kept straight and might accidentally hit a key. As noted before, octave runs show a particularly high number of inserted notes in both hands.

Étude op.25 No.6 contains several downward runs in thirds. In each of these runs,

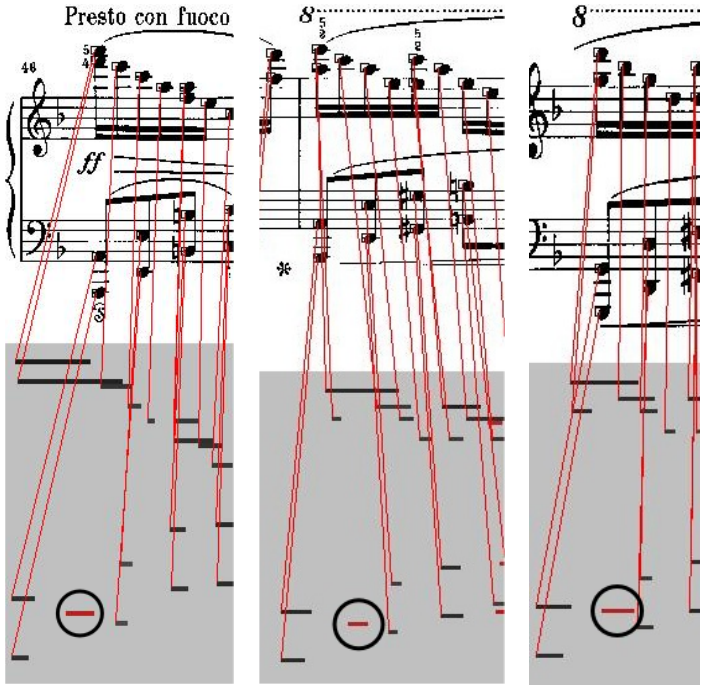


Figure 3.6: Systematic Insertions in Ballade Op. 38, Bars 46 (left), 50 (middle), and 58 (right panel).

Magaloff omitted notes from the highest voice at regular intervals (every third or fourth note). The regularity suggests a technical problem with the fingering in this passage (see Figure 3.7). In Étude op.25 No.1 Magaloff often omitted the second note of the figure in the left hand. This suggests a weak third finger and a problem covering the large span required in the left hand. A systematic substitution can, for instance, be found in Étude op.25 No.6, bars 7 and 8, where Magaloff consistently played A instead of A♯. This is probably a problem of memorization.

Omitted Inner Voice

A special case of systematic deletion is the omission of an inner voice: throughout a sequence of onsets, an inner voice is omitted partially or completely. In most instances the most likely cause is either a memorization problem (the least significant voice was simply forgotten) or the need for technical simplification, depending on the complexity

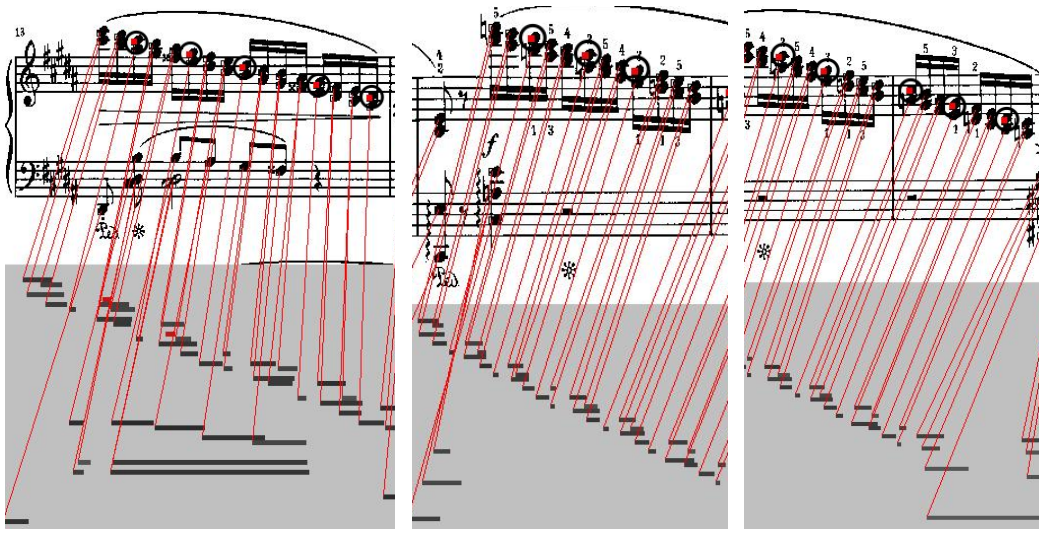


Figure 3.7: Systematic Omissions in Étude Op. 25 No. 6, Bars 13, 45, and 47.

of the passage. For instance, in Étude op.25 No.10, Bar 16 (figure 3.8, left panel), Magaloff omitted one of the two inner voices from a sequence in which the two hands move in parallel octaves. In this highly homogeneous context, the omission is obvious to the audience and clearly not a problem of memorization but a result of the technically demanding nature of the piece.

Note Order Errors

Note Order Errors are the only category that relates to timing rather than pitch: The order in which two (or more) successive notes are played is switched, resulting in two (or more) substitution notes. Instances of this pattern are mainly found in Étude op.25 No. 3 and Prélude op.28 No.8. In both pieces, the affected group of notes is a descending pattern in the left hand, consisting of 4 notes. In the Étude, the lower of the two notes at the first onset is played after the third note in the group, resulting in a downward sequence of 4 notes. In the Prélude, the affected group is very similar, with the slight difference that the first two notes are to be played successively instead of simultaneously. Again, the two middle notes are often switched, producing the same downward sequence as in the Étude. Figure 3.9 shows an excerpt from the Prélude. As the performance tempo of both pieces is very high, the change in note order is impossible to discern for

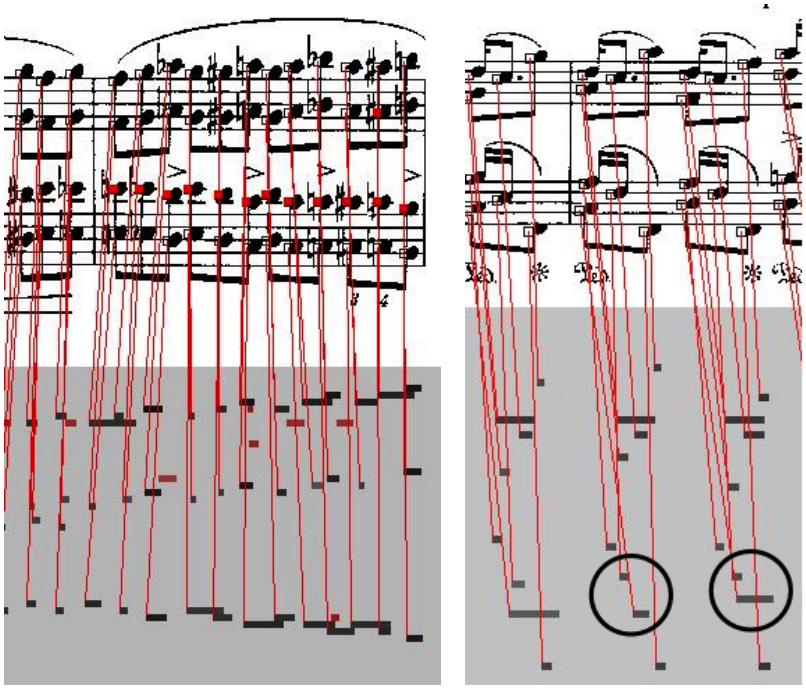


Figure 3.8: *Left panel*: Omitted Inner Voice in Étude Op. 25 No. 10, Bar 16; *Right panel*: Note Order Errors in Étude Op. 25 No. 3.

the audience. For the Étude intentional simplification is more likely to be the reason for the errors, as the figuration is characteristic for the complete piece. In the Prélude both technical simplification and memorization could be the cause. However, as the accompaniment does not change over the course of the piece, and some of the groups are played in the correct order, technical simplification seems to be more likely.

3.1.6 Conclusion

Comparing the 1989 concerts with his earlier recordings (both in the studio and on stage) it is obvious that Magaloff's age affected his playing. The chapter opens with a quote from V. Horowitz, effectively postulating that any error he makes are due to the realization of his musical ideas. Perhaps it is in this spirit, that the error analysis of Magaloff's Chopin has to be interpreted: He was not willing to sacrifice his musical ideas and ideals at the altar of pitch perfection. This seems warranted especially when

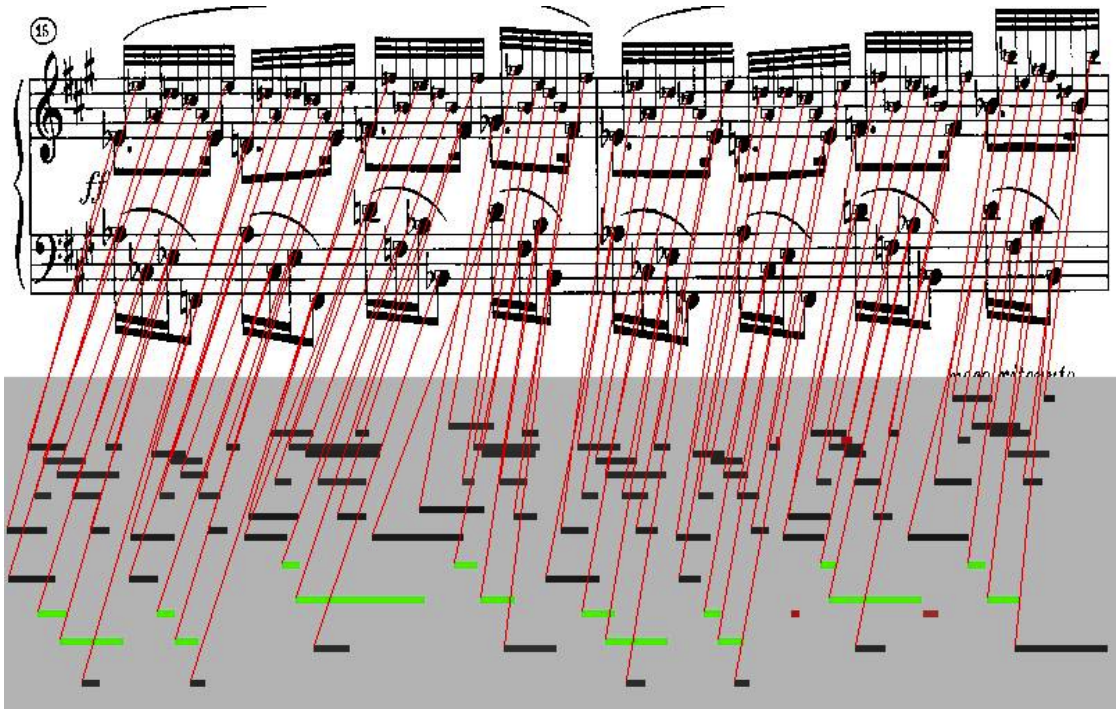


Figure 3.9: Note Order Errors in Prelude Op. 28 No. 8.

taking into account the considerations of performance tempo and performer age, that are presented in the following section 3.2.

Nevertheless, the Magaloff corpus contains performance errors in abundance and therefore offers a unique insight into the phenomenon; all the more so since the recordings have been on-stage without any editing, as would be the case for studio recordings. Overall, the results corroborate and complete the previous studies on the matter: erroneous notes have the tendency to be inconspicuous and mostly will go unnoticed by the audience. The inventory of occurring error patterns and situations can help understanding why errors are happening in the first place. Given suitable data, it would also be interesting to compare how other pianists cope with technically demanding situations: whether they share techniques to simplify passages by harmonic substitutions and whether there are pieces that all find particularly hard to memorize.

The analyses presented here are mostly based on the Henle Urtext Editions of the scores. As mentioned in section 2.1, Magaloff was known for having studied Chopin's

manuscripts extensively. Recently, a collection of all first editions of Chopin's scores was made available under the direction of John Rink at the Royal Holloway, University of London[3]. Together with the here discussed differences between Magaloff's performances and the Henle Urtext editions, this resource could be used to narrow down the editions Magaloff might have used to study the pieces in the first place. This might offer further explanation for some of the disparities.

3.2 Performer Age

One of the remarkable aspects of Magaloff's Chopin concerts is the age at which he undertook this formidable task: he was 77 years old.[4] The demands posed by performing publicly are enormous (motor skills, memory, physical endurance, stress factors, see [132]). Theories of human life-span development identify three factors to be mainly responsible for successful aging: selection, optimization, and compensation (SOC model, [3]). Applied to piano performance this would imply that older pianists play a smaller repertoire (selection), practice these few pieces more (optimization), and hide technical deficiencies by reducing the tempo of fast passages, while maintaining tempo contrasts between fast and slow passages (compensation) [121]. In [27], a study conducted in cooperation with W. Goebl, we test whether Magaloff actually used strategies identified in the SOC model.

The first aspect of the SOC model, *selection*, seems not to be supported in this case: Magaloff performed the entire piano works by Chopin within four months.[5] We cannot make a statement about *optimisation* processes due to our lack of information about his practice regime before and during the concert period. Regarding possible *compensation* strategies, we first examine Magaloff's performance tempi of the Etudes in the context of other recordings (section 3.2.1). We then compare the tempo ratios between fast and slow passages in Magaloff's performance of a Nocturne to the tempo ratios seen in recordings of several other pianists (section 3.2.2).

[3] `http://www.cfeo.org.uk`

[4] At age 77, Alfred Brendel performed one solo program and one Mozart Concerto for his last season in 2008

[5] Of course, Magaloff's repertoire might have been broader in younger years, which would then indicate otherwise. A systematic comparison of earlier concerts seasons and all concerts in 1989 would provide further insights into that particular aspect.

3.2.1 Performance Tempo in Chopin's Etudes

To assess Magaloff's compensation strategies, we examine the performance tempo of difficult pieces, such as the Etudes, and compare them to his earlier recording at the age of 63 and to recordings by several renowned pianists. These audio recordings, a total of 289 performances of 18 Études by 16 performers[6], were beat-tracked semi-automatically using the software *Beatroot* [19, 20]. A basic tempo value was estimated by the mode value, the most frequent bin of an inter-beat interval histogram with a bin size of 4% of the mean inter-beat interval.

Op.10 No.1		Op.10 No.2		Op.10 No.4		Op.10 No.5	
BI49	157	BI49	129	HA29	157	SH32	104
HA29	159	**MA77**	139	BI49	157	**MA63**	111
SH32	163	SG32	140	AR53	161	LO27	115
CO56	164	HEN	144	SC31	165	**MA77**	115
MA63	165	HA29	145	**MA63**	166	AS38	115
SC31	169	**MA63**	145	SH32	169	HEN	116
AS38	170	CO56	149	LO27	169	BI49	117
MA77	170	AR53	152	PO30	169	SC31	117
HEN	176	SC31	152	**MA77**	170	Gi33	118
PO30	178	PO30	152	GI33	174	CO56	120
LO27	179	LO27	156	AS38	174	LU27	120
BA44	179	AS38	157	CO56	175	AR53	121
LU27	180	LU27	159	HEN	176	HA29	122
GA30	190	GI33	165	LU27	179	PO30	123
GI33	191	GA30	173	BA44	191	GA30	131
AR53	196	BA44	176	GA30	197	BA44	139

Table 3.4: Tempo values for selected Etudes from Chopin Op. 10. Each performance is named by the first two letters of the pianist followed by the pianists age at the time of the recording. Columns are sorted by ascending tempo values.

[6] Arrau (recorded 1956), Ashkenazy (1975), Backhaus (1928), Biret (1990), Cortot (1934), Gavrilov (1985), Giusiano (2006), Harasiewicz (1961), Lortie (1986), Lugansky (1999), Magaloff (1975), Magaloff (1989), Pollini (1972), Schirmer (2003), Shaboyan (2007), Sokolov (1985).

Op.10 No.7		Op.10 No.8		Op.10 No.10		Op.10 No.12	
BI49	232	BI49	142	BI49	426	PO30	64
MA63	237	HA29	157	BA44	450	LO27	64
HA29	242	SH32	157	**MA63**	467	**MA63**	65
SC31	243	**MA63**	159	SC31	471	SC31	66
MA77	244	BA44	168	HEN	480	LU27	66
SH32	248	SC31	173	SH32	480	AS38	66
HEN	252	LO27	174	AR53	483	HA29	68
AR53	252	GI33	174	LU27	487	BA44	71
GA30	254	**MA77**	174	HA29	505	SH32	71
LU27	256	HEN	176	GA30	508	**MA77**	72
LO27	256	AS38	177	AS38	512	BI49	74
CO56	263	CO56	178	PO30	513	CO56	75
AS38	264	AR53	179	LO27	529	HEN	76
PO30	266	PO30	180	CO56	542	GI33	77
GI33	271	GA30	188	**MA77**	550	GA30	87
BA44	285	LU27	190	GI33	574	AR53	88

Table 3.5: Tempo values for selected Etudes from Chopin Op. 10 (continued).

Tables 3.4 and 3.5 show the tempo modes obtained for all pianists on the Etudes selected from Op. 10. Tables D.1 - D (see appendix D) contain the measurements for the remaining pieces. For the sake of comparison the metronome indications from the Henle Edition [137] of the Etudes were added (HEN). In 12 of the 18 pieces Magaloff's tempo is within a 10% range of the metronome markings of the Henle edition. Three pieces are more than 5% slower and three pieces more than 5% faster compared to the metronome markings. Compared to the performances by other pianists, Magaloff's performances of the Op. 10 Etudes are on average 1.2% slower than the average over all other recordings. The Op. 25 Etudes are on average about 5.6% slower than the average performance. Comparing Magaloff's recordings at the age of 63 and 77, the tempi vary to a surprising degree, but no systematic tempo decrease in the latter can be found. On the contrary: in 12 pieces out of 18 the recording at age 77 is faster, sometimes to a considerable degree (up to 17% in op. 10 No. 10).

On the whole, Magaloff's performances do not suggest a correlation between age and

tempo, while the tempi of the other pianists' recordings show a slight age effect (with piecewise correlations between pianist age and tempo ranging from −0.66 to 0.51, with an average of −0.17). These considerations are based on the underlying assumption that the difficulty of a piece increases with the tempo. This is not universally true. However, for the pieces in question – the fast pieces of the Études – the assumption seems warranted.

3.2.2 Age effects and tempo contrast in a Nocturne

In his interpretation of the SOC model Vitouch reasons, that *compensation* takes place by an overall decrease of performance tempo, while the contrast between fast and slow passages is kept constant [121]. To assess this we examined the tempo values in 14 performance of the Nocturne op. 15 No. 1 by other pianists[7]. The piece has an A-B-A form, with a fast middle part (*Andante cantabile* (Bars 1-24) – *con fuoco* (25-48) – *Tempo I* (49-75)). We found a significant correlation between the performance tempo of the middle section and the age of the performer (the older, the slower, see figure 3.10, left panel). However, the tempo ratios between the contrasting sections of the piece showed no overall age effect, confirming Vitouchs [121] interpretation of the SOC model. Age seemed to have no effect on Magaloff's Nocturne: he played faster than the youngest of the performers while keeping a comparable tempo ratio. The same tendency could be found in the Etude Op. 25 No. 10, however the negative correlation was not significant.

3.2.3 Conclusion

Based purely on the fact that Magaloff performed the entire piano works by Chopin, we can refute the selection part of the SOC model. Due to missing information about Magaloff's practice regime before and during the performance period, we cannot make a statement about optimization processes. Magaloff's performance tempi do not point to compensation processes, which were indeed found in recordings by other famous pianists. In sum, Magaloff's data does not seem to corroborate the SOC model.

[7] Argerich (recorded 1965), Arrau (1978), Ashkenazy (1985), Barenboim (1981), Harasiewicz (1961), Horowitz (1957), Leonskaja (1992), Maisenberg (1995), Magaloff (1975), Perahia (1994), Pires (96), Pollini (68), Richter (68), and Rubinstein (1965).

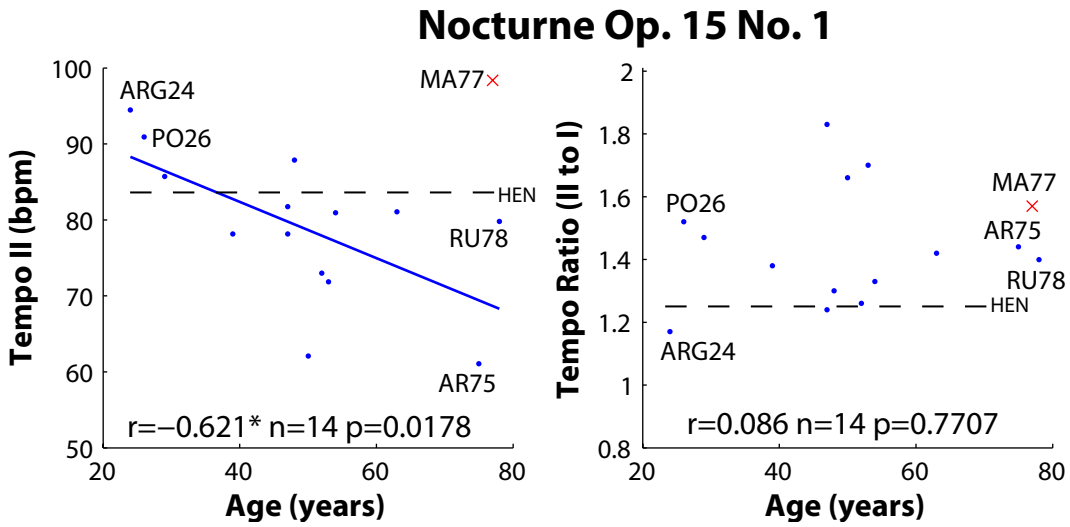


Figure 3.10: Left panel: Tempo and performer age of 14 different pianists playing the middle section of Nocturne Op. 15, No. 1, including the tempo prescribed by the Henle edition of the score. A significant effect of age (the older, the slower) can be identified. Right panel: Tempo ratio between the middle section and the first section of the Nocturne. No significant effect of age can be identified.

3.3 Between-hand Asynchronies as Expressive Device

The following is a recount of a study done mainly by W. Goebl [45]. The study discusses another example of a musicological phenomenon, the examination of which requires data as precise, and as representative for a composer and a musician as the Magaloff corpus. The summary of this study can also be found in [26].

Temporal offsets between the members of musical ensembles have been reported to carry specific characteristics that might reflect expressive intentions of the performers; for instance, the principal player in wind or string trios precedes the others by several tens of milliseconds [95], and soloists in jazz performances have been shown to synchronise with the rhythm section at offbeats [38]. As the hands of a pianist are capable of producing different musical parts independently, the temporal asynchronies between the hands may be an expressive means for the pianist. The asynchronies were computed automatically over the entire corpus based on staff information contained in the score, assuming that overall the right hand played the upper staff and the left hand the lower. For the

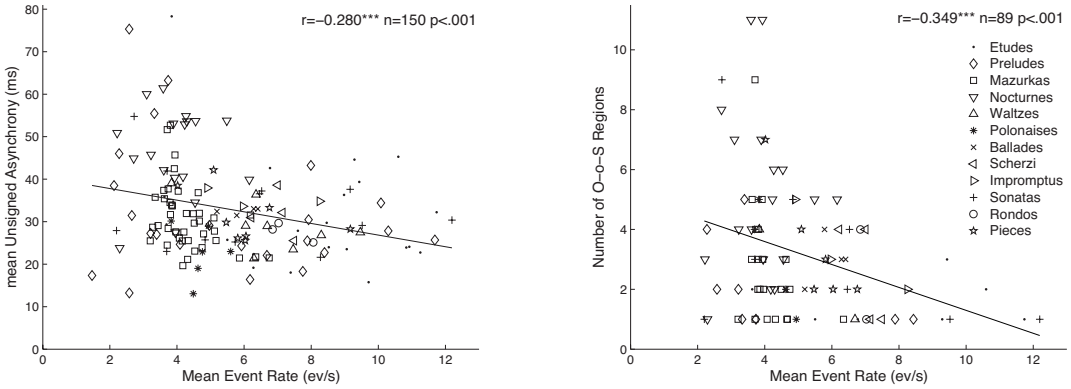


Figure 3.11: *Left panel*: Absolute asynchronies plotted against the mean event rate by piece category. *Right panel*: The number of out-of-sync regions plotted against the mean event rate plotted. (Picture from [44])

analysis of this phenomenon we excluded all onsets marked in the score as *arpeggiated*; in these cases temporal deviations are prescribed by the score rather than being part of the interpretation. The main results of this study [45, 44] are reported briefly in the following.

The analysis of over 160,000 nominally simultaneous events revealed tempo effects: slower pieces were played by Magaloff with larger asynchronies than faster pieces. Figure 3.11 (left panel) shows the correspondence between event rate and asynchrony. Moreover, pieces with chordal texture were more synchronous than pieces with melodic textures. Subsequent analyses focussed on specific kinds of between-hand asynchronies: bass anticipations and occurrences of "tempo rubato in the earlier meaning" [61].

As *bass anticipations* we consider events where a bass note precedes the other voices by more than 50 ms. They can be clearly perceived due to their large asynchronies and can be considered to be expressive decisions by the performer. Magaloff's performances contain a considerable number of these bass anticipations (about 1% of all simultaneous events). Again, higher proportions are found in slower pieces.

The *"tempo rubato in the earlier meaning"* refers to particular situations in which the right hand deviates temporally from a stable timing grid established by the left hand [61]. Chopin, in particular, recommended to his students this earlier type of rubato as opposed to the later type that refers to a parallel slowing down and speeding up of all

parts of the music (today referred to as expressive timing). We automatically identified sequences where Magaloff apparently employed an "earlier tempo rubato" by searching for *out-of-sync regions* in the pieces. An out-of-sync region is defined as a sequence of consecutive asynchronies that are larger than the typical perceptual threshold (30ms) and that comprises more events than the average event rate of that piece. On average, 1.8 such regions were found per piece (283 in total) with particularly high counts in the Nocturnes – a genre within Chopin's music that leaves most room for letting the melody move freely above the accompaniment. Figure 3.11 shows the correspondence between event density and number of "earlier tempo rubato" sequences, suggesting that slower pieces leave room for tempo rubato than faster pieces

3.3.1 Conclusion

On the whole, this is a strong indication that indeed asynchronies between a pianist's right and left hand serve an expressive purpose. Magaloff is known for having studied Chopin's manuscripts extensively, so it can be safely assumed that he was familiar with Chopin's recommendation of how to "steal time" (the literal translation of *tempo rubato*). Obviously, this also translated into his interpretations. This behavior is neither unique to Magaloff's way of playing nor particular to Chopin. Daniel Barenboim teaches this understanding of tempo rubato in his masterclasses on Beethoven's Sonatas[8]. Up to now, however, this has not been investigated and confirmed empirically. These first investigations into between-hand asynchronies illustrate how specific musicological questions may be assessed by elaborate quantitative means.

[8]Miller, A. (Director). Barenboim on Beethoven – Masterclasses. EMI Music (2005).

Chapter 4

YQX – Expressive Performance Rendering

Train your senses. Music without interpretation is music without meaning.

W. B. Bailey, "Piano Pointers"
The Etude, July 1923

The Magaloff corpus is a versatile resource. As demonstrated in chapter 3 it facilitates performance analyses at the highest level of precision, and makes it possible to examine musicological questions empirically. The following chapter shows the corpus in its role as training data for a machine learning endeavor: *Expressive Performance Rendering.*

As introduced in section 1.3, Expressive Performance Rendering systems sonify hypotheses on expressive music performance, and thereby try to create expressive performances automatically. The most obvious goal, to be able to generate a profoundly musical performance of a hitherto unknown piece of music automatically, is hardly ever reachable. However, even crude predictive models of expression can be of extreme help on the way to expressive, synthetic performances. This chapter explores the use of probabilistic models for the purpose of expressive performance rendering. By gradually improving a very basic model, we hope to additionally gain more insights into how expressive performance works.

First, the different components of our approach are put into context (section 4.1). Following a description of the score and performance models we use (sections 4.2 and 4.3), the basic model is introduced (section 4.4, first published alongside a general introduction to the field in [128]). Two extensions of the prediction algorithm are presented in section 4.5 (first published as [30]), that make the system aware of performance context. To use the proposed extensions to their advantage, adaptations of the performance model have to be made (section 4.6, published in part in [33], and as part of a general introduction to performance rendering with probabilistic models [31]).

The presented model is based on a method called graphical probabilistic models, the basics of which are outlined in appendix A.

4.1 Related Work

Systems can be compared in terms of four main components: the score representation, the learning and prediction model, which and how performance parameters are controlled, and the way expressive directives given in the score are rendered.

4.1.1 Score Models

Any representation of a music score and/or its structure can be seen as a score model. It is used to capture any aspect of music that seems important for the particular task. The most obvious and straight-forward model would be a list of all notes in the score by pitch, duration, and temporal position. This representation, however, still has a lot of information that is not made explicit, but just implicitly present: strong and weak rhythmic positions imposed by the meter, harmonic tension and relief created through the vertical and horizontal context, melodic patterns and repetitions, phrase structure and sub phrases. All of these can only be made apparent by putting the notes in context of melody and accompaniment, and viewing them as parts of larger musically meaningful units rather than just as a sequence of single notes.

Several music theories exist that provide methodology for analyzing and understanding a piece, explaining its structure, its harmonic progression (and related tension and relief) or climactic points. Most prominent among those are: (1) Schenkerian analysis [34], which shows hierarchical relationships among the pitches of a passage of music by rhythmic reduction, and infers structural cues from this hierarchy; (2) Lerdahl&Jackendoff's Generative Theory of Tonal Music (GTTM) [69], which explains the structure of a piece by applying the supposed mental procedures that lead to the insights an experienced listener forms about a piece [52]; (3) Narmous's Implication-Realization (IR) model [78], which draws conclusions about structure from the listener's melodic expectations and how they are evoked and satisfied by the composition (see section 4.2.5 for a short overview); (4) Parncutt's accent theory [87], which searches for moments that attract the listener's attention that are immanent in the score, for example through metrical, melodic, and harmonic saliences or grouping boundaries; (5) Cambouropoulos' *Local Boundary Detection Method (LBDM)* [9, 10] applies rules related to the Gestalt principles [72] to hypothesize about possible groups based on the amount of local change. The approach can be applied to arbitrary musical aspects that contribute to perceived segmentation, like melodic texture, harmonic progression, or rhythmic change.

The goal, segmenting a piece into musically meaningful and perceptually relevant units, is in some way addressed by all of the above. Phrases, an example of such units, are one of the most important and noticeable elements of human performance, and performance modeling systems can benefit greatly from such structural cues [124]. However, even this elemental aspect of music performance is subjective, unequivocal, and result of too complex a combination of factors, such as personal taste and interpretational intention, to be cast into a definite and determinate set of rules with satisfactory results. This is reflected in the generally rather low quality of the automatically generated segmentations for moderately complex music.

Of the above mentioned, only the LBDM is formulated precisely enough to be automated. All other analyses leave certain details to interpretation, and in that reflect the subjectivity and complexity of the issue. This of course makes it difficult to automate the analyses, and renders each attempt to quantify and implement them a subjective interpretation. For both the GTTM and the IR model efforts have been made to provide a faithful implementation ([52], and [46] respectively). Temperley's Melisma project [115] is an implementation of parts of the GTTM. Automatically finding accents in music according to Parncutt's theory is the focus of active research.

Score descriptors, or *features*, are a simple and (mostly) computationally inexpensive way to provide basic abstraction from the pitch/duration/position representation. They just characterize local melodic, rhythmic, and harmonic aspects of the score. Section 4.2 shows an example of such a feature based approach: notes are put into context of their immediate vicinity, describing for example melodic contour (pitch intervals to the next note), rhythmic evolution (ratio of the durations of two successive notes), and harmonic tension (dissonance with the current accompaniment) on a very local level.

Several systems exist that rely on simple score features as a score model, among them Widmer's Rule System [127] and Teramura's statistical model [117]. Both Grindlay's probabilistic approach [50, 51] and Widmer's case-based system [119, 130] use local descriptors combined with a manual segmentation of the pieces into phrases (in Widmer's case even four hierarchical phrase levels). SaxEx, the system by Arcos et al. [1], is based on a combination of local score descriptors and parts of the GTTM. As no implementation of the theory existed at that time, the analysis was done by hand. The accent-based approach presented by Bisesi et al. [5] relies on the user to manually annotate the accents in the score according to Parncutt's taxonomy.

Any manual annotation is laborious, requires musical knowledge and thus limits the

applicability of the systems to experiments on a smaller scale. Our system YQX is exceptional in using a computational implementation of a music theory, the IR model, enriched by local score descriptors. Grachten's parser, used to compute the IR analysis, is described in [46].

4.1.2 Performance Models

The performance model is the set of dimensions that is manipulated to make a performance expressive. As we deal with piano performances, the most obvious are tempo, loudness, and articulation, which are, in some way or other, addressed by all systems. Generally, there is not much variation across the different systems in the way tempo and loudness are defined. Instead of absolute values, tempo is modeled through inter-onset-intervals between successive notes, which makes it independent of the globally established tempo. Loudness in some cases is represented in absolute terms [117], which excludes a global regulative, as opposed to a model of loudness relative to a reference value (e.g. the loudness mean), which allows raising or lowering the global loudness without changing the relative intensities. Expressive rendering systems deal with articulation only in terms of gaps between notes (staccato vs. *legato*), although, from a performer's perspective, several articulation techniques and effects also involve loudness modulation (e.g. *martellato* and *marcato*). Teramura [117] models the difference between the offset prescribed by the score and the time a note is actually released. This is only one half of the picture, as it does not consider the onset of the following note. The KTH rule system [36] defines articulation, or the amount of *legato/staccato*, as a ratio between duration of a note in the performance and the inter-onset-interval to the successive note, which is the approach we take.

An effect specific to the piano is the sustain pedal, with which a pianist can raise and lower the dampers onto the strings. This affects the produced sound immensely. As it raises the dampers from all strings, they all can resonate, giving the piano a fuller sound. In addition, pianists use the pedal to generate various effects: for instance, "collect" the sounds of several consecutive notes and chords and hereby aggregate harmony, or facilitate highly legato passages. Because of the significant influence on the generated sound, the sustain pedal can ruin a lot, if applied incorrectly. As pedaling is to a large extent related to accompaniment and phrasing – the former we do not explicitly consider, the latter we cannot reliably address – we set modeling the pedal aside for the time being. To the author's knowledge *Coper* by Kenji Noike [55] is the only system

actively modeling pedal, but, due to the above mentioned risk, with rather poor results.

4.1.3 Learning and Prediction Models

Regarding the learning and prediction models used, three different categories can be distinguished [129]: Case-based Reasoning (CBR), rule extraction, and probabilistic approaches.

Case-based approaches

Case-based approaches use a database of example performances of music segments. The underlying assumption is that, given a suitable distance or similarity metric between score models, segments that have a similar score also entail a similar performance. Hence new segments are played by imitating the stored ones that have the most similar score. Prototypical case-based performance models are *SaxEx* [1] and *Kagurame Phase II* [113]. Widmer and Tubodic developed a CBR system based on a hierarchical phrase segmentation of the music score [119, 130]. The results are exceedingly good, but the approach is limited to small-scale experiments, as the problem of algorithmic phrase detection is still not solved in a satisfactory way. Dorard et al. [24] used Kernel methods to connect their score model to a corpus of performance worms, aiming to reproduce the style of certain performers.

Rule-based systems

Rule-based systems map score features directly to performance modifications, governed by a set of rules. Widmer [127] developed an inductive rule learning algorithm that extracted performance rules from piano performances; it discovered a small set of rules that accounted for a surprisingly large amount of expressivity in the data. Ramirez et al. [91] followed a similar approach using inductive logic programming to learn performance rules for Jazz saxophone from audio recordings. Perez et al. [89] used a similar technique on violin recordings. The well-known KTH rule system was first introduced in [112] and has been extended in more than 20 years of research. A comprehensive description is given in [36]. The rules in the system refer to low-level musical situations and theoretical concepts, and relate them to predictions of timing, dynamics, and articulation. In contrast to Widmer's rules, the KTH rules have not been learned automatically from real

performance data, but developed via an analysis-by-synthesis-approach, where professional musicians evaluated rules brought forward by researchers. The *Director Musices* system is an implementation of the KTH system that allows for expressive rendering of musical scores. All rules are governed by parameters set by the user. Rendering a performance involves tuning the parameters until the result complies with one's musical ideas. The rendering of a Bach Prelude won the RENCON 2004, proving the validity of the approach. A problem, however, is that one parameter controls the application of a rule for the complete piece, requiring careful tuning. Also, good parameter settings vary from piece to piece.

In a recent extension, rules have been added that implement Parncutt's accent theory [5]. The user can choose how the system will react to (manually) annotated metrical, harmonic, rhythmical and grouping accents in the score. In the RENCON 2011 the "Accent based approach to performance Rendering" achieved the third prize with a successful rendering of the third movement of Beethoven's Sonata Pathethique (see section 5.7.4).

Probabilistic approaches

In probabilistic approaches, performance and score model are regarded as a joint multivariate probability distribution, which is estimated from a large set of training performances. Based on the assumption that a new piece and an appropriate performance thereof come from the same underlying distribution as the training performances, the idea is that, as the new score is known, the performance part of the distribution can be inferred. The approaches differ in how the inference is carried out. The Naist model [117] applies Gaussian processes to fit a parametric output function to the training performances. Grindlay and Helmbold first proposed a Hidden Markov Model (HMM) [51] that they later extended to a Hierarchical HMM [50], where phrase information is coded into the structure of the model. All approaches mentioned above learn a monophonic performance model, predict the melody voice of the piece and, in the rendering, synchronize the accompaniment accordingly. Kim et al. [65] proposed a model of three sub-models: local expressivity models for the two outer voices (highest and lowest pitch of any given onset) and a harmony model for the inner voices.

Mazzola follows a different concept, building on a complex mathematical theory of musical structure [70], implemented in the *Rubato* system [71, 73].

4.1.4 Rendering of expressive annotations

To a certain degree expressivity is determined by the note content of the score, like pitch, duration, and temporal position of the notes, meter and key. This is evident from the fact that the validity of the music theories mentioned in section 4.1.1 can indeed be verified [106]. Printed musical scores usually contain much more information: depending on composer, and edition of the score, additional directives to the player are given regarding tempo, dynamics, articulation, phrasing et cetera. They constitute expressive content that is proposed or intended by the composer or the editor. The directives might underline what could be read from the score alone (through experience, performance tradition, composer style etc.), they might enhance a segment that does not by itself suggest a certain interpretation, or they might even contradict what one would expect from the note content. Musicians performing the piece should observe the directives, and arrange their interpretation around them. To quote Schulz from the "Allgemeine Theorie der schönen Künste" (General Theory of the *Beaux Arts*)[111]:

> [...] the few signs with which the composer describes the execution of single notes of phrases must be observed as exactly as possible because for certain movements they are as essential as the notes themselves.

A discussion of how a musician should realize the directives is for example given in [102]. A system striving to generate a performance that is as "human" as possible should strictly adhere to the directives given in the score.

Regarding the question of how to computationally render expressive performance directives given in the score (e.g., rit., cresc., fermatas etc.) in a natural or musical way, there is little directly relevant literature. The question of what sounds natural, and why, is discussed in research that attempts to relate musical expression to other natural or physical phenomena. Kinematic models of expressive timing (and, in part, dynamics) have been proposed most prominently by Todd [120], who uses particle movement under constant acceleration/deceleration as an analogy, and Friberg and Sundberg [37], who derive their model of constant deceleration from observing human runners coming to a halt. Honing [60] argues that kinematic models ignore certain essential characteristics of music – namely event-density, rhythmic structure and overall performance tempo – and proposes a perceptual model that puts bounds on the range of acceptable tempo curves for ritards. Nevertheless, [120] and [37] show that kinematic models can predict final ritards quite accurately when fitted to empirical data.

A very recent approach by Grachten formulates the problem as a machine learning task [48]. Each loudness directive is assigned a basis function that represents loudness as a function of time. A weighted sum of all expressive functions for one piece approximates a loudness curve for a performance. In [48] Grachten shows that a considerable amount of dynamic variation can indeed be explained by realizations of the expressive directives in the score, suggesting that those are major cues for performance musicians. There is no indication if and how competing rendering systems realize cues from the expressive annotations at all.

4.2 Score Model

For us to be able to evaluate the system on a larger scale, like the Magaloff Corpus, we have to be able to compute the score model automatically, as opposed to manually analyze and annotate the scores. Between the methodolgies that, as of now, have been implemented as computational models, the GTTM [52], the IR-model [46], and the LBDM [10], we decided to use the IR-model. In addition to an approximation of the phrase structure, it provides descriptions of melodic patterns that we can also use as score features. Moreover, the implementation by M. Grachten [46] was readily available. In addition to the IR-model, we decided on a number of local descriptors of the score, that provide some very crude abstractions regarding melodic, rhythmic, and harmonic aspects, and put score events in context of their close vicinity.

Most of the proposed features and targets require a monophonic stream of notes: trying to connect, for example, the pitch interval to the next note to a pianist's choice of tempo modification only makes sense if the next note is a successor in the same melodic idea and not part of a different voice. The same principle applies to most of the other features. Consequently, with the exception of features using harmonic aspects of the score, we only look at the "melody" voice of the pieces. For a part of our data, the Mozart sonatas (see 5.1), the melody voice was marked by hand. For the Chopin data, such an annotation is not available. Instead we apply the following, very simple heuristic: at any given time, the highest pitch in the upper staff is considered to be part of the melody voice. This certainly is not always true, but in case of Chopin, is correct often enough to be justifiable.

Some of the features produce discrete, categorical values, and are later on treated as discrete random variables, while others are continuously valued, and are later on treated

as continuous random variables.

4.2.1 Rhythmic features

Rhythmic features describe the relations of score durations of successive notes and their rhythmic context. In our system, we use:

Duration Ratio (continuous): The numeric relation between the durations of two successive score notes s_i and s_{i+1}. Let dur_i be the score duration of note s_i measured in beats; the duration ratio $durR_i$ is then defined by:

$$durR_i = \frac{dur_i}{dur_{i+1}}.$$

An example of a sequence of duration ratios can be seen in figure 4.1.

Rhythmic Context (discrete): A symbolic description of the rhythmic context of a note. The score durations of notes s_{i-1}, s_i, and s_{i+1} are sorted, compared, and assigned 3 different labels: the shortest of the three is labeled *short (s)*, the longest *long (l)*, and the third *neutral (n)*. The triplet of rhythmic labels $rl_{s_{i-1}}$, rl_{s_i}, and $rl_{s_{i+1}}$ for the three notes s_{i-1}, s_i, and s_{i+1} put together form the rhythmic context $rhyC_i$ of s_i. Examples of different rhythmic contexts and how they are labeled can be seen in Figure 4.1.

As we do not take the absolute durations into account, the triplets lsl and lnl describe the same situation. This is the case every time two of the three durations are the same. In those cases we always label the longer one l and the shorter one n. How rests are treated depends on their duration: a rest immediately before s_i shorter than half the duration of s_{i-1} is ignored and the durations of s_i and s_{i-1} are used for the triplet; a rest longer than that is dealt with by replacing the respective labels with (-).

4.2.2 Melodic features

Melodic features describe the shape of the soprano voice of the piece. We include pitch intervals between two successive notes, both as they are and clustered into groups. The *peak* - features try to locate meaningful turning-points in the melodic trend: points where absolute or average pitch reaches a local maximum or minimum. Narmour's

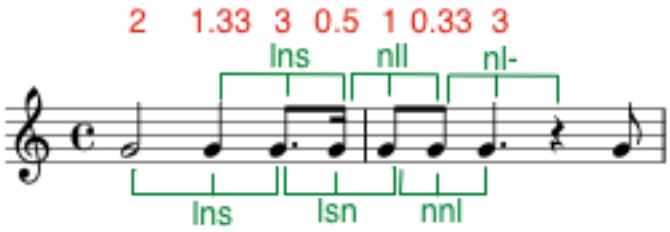


Figure 4.1: A sequence of notes with different durations illustrating two rhythmic features: *Duration Ratio* (red) is the ratio of the durations of two successive score notes; *Rhythmic Context* (green) describes the duration of note in relation to the preceding and the following note.

Implication-Realization model provides a characterization of note triplets, describing the expectations they raise in a listener.

Pitch interval (discrete): The interval to the next melody note, measured in semitones. The values are cut off at −13 and +13 so that all intervals greater than one octave are treated as identical. See figure 4.2 for an example.

Grouped pitch interval (discrete): The pitch interval pi_i to the next melody note is not used directly, but put into one of the following groups: $pi_i \leq -9$, $-9 < pi_i \leq -5$, $-5 < pi_i \leq -2$, $-2 < pi_i \leq 2$, $2 < pi_i \leq 5$, $5 < pi_i \leq 9$, and $9 < pi_i$. An example can is shown in figure 4.2.

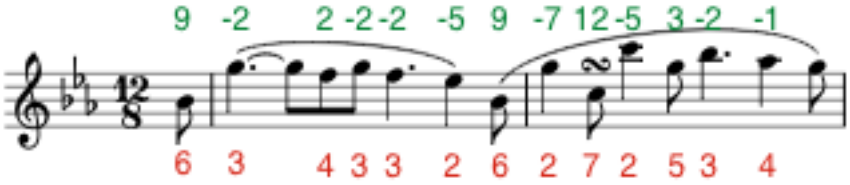


Figure 4.2: The soprano voice of the opening bars of Chopin's Nocturne, Op. 9 No. 2; shown are the features *Pitch Interval* (green) and *Grouped Pitch Interval* (red).

Melodic Max (Min) Peaks (discrete): The sequence of pitches is first smoothed by applying a moving average with a window of size 5 (2 preceding notes, the central note, and 2 subsequent notes). The resulting series of average pitches is then segmented at the points where the gradient changes – either from a rising average pitch to a falling average pitch (for max peaks) or vice versa for min peaks. The note with the maximum (minimum) pitch in the segment is called *melodic max (min) peak*. The distances of the remaining notes to the peak in the surrounding segment – negative prior to the peak, 0 at the peak, and positive subsequently – define the feature. Figure 4.3 shows both features – max and min peaks – for the opening bars of the Nocturne Op. 9 No. 2.

Max (Min) Average Peak (discrete): The sequence of pitches is smoothed with a window of size 5 (2 preceding notes, the central note, and 2 subsequent notes). The maxima (minima) of this sequence are call *max (min) average peaks*. For groups of 4 notes surrounding the peaks the distances to the respective peak defines the value for the notes. Distances are negative prior to the peak, 0 at the peak, and positive subsequently. Groups that overlap by 2 notes are merged into a group with the larger (smaller) of the two peaks as a reference. Notes not within the scope of a maximum or minimum default to the value −3. Figure 4.3 shows both features – max and min average peaks – for the opening bars of the Nocturne Op. 9 No. 2.

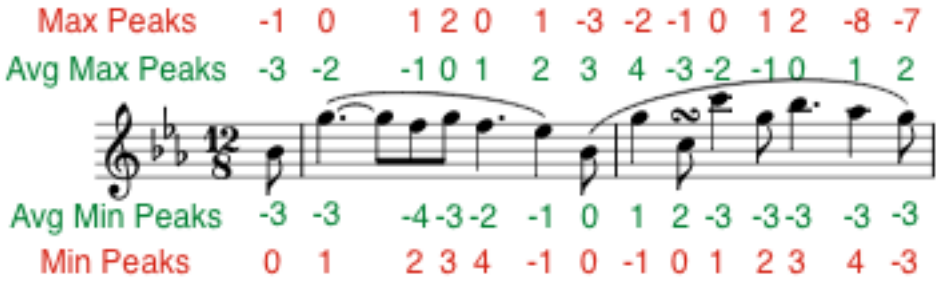


Figure 4.3: Soprano voice of the opening bars of Nocturne, Op. 9 No. 2. The numbers in red form the sequence of *Melodic Max/Min Peak* features – notes with value 0 are the melodic peaks, non-zero values are distances to the peak. The numbers in green representing the corresponding sequence for the *Max (Min) Average Peak* feature.

IR Label (discrete): Narmour's Implication–Realization model provides a categorization of triplets of notes according to the expectations they raise in the listener, see section 4.2.5 for a short description. We use the label assigned to each melody note as a score feature. Figure 4.4 shows examples for a subset of the IR categories we use.

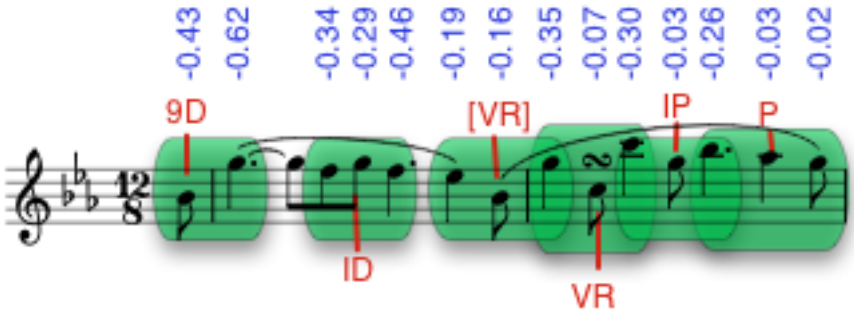


Figure 4.4: The soprano voice of the opening bars of Nocturne Op. 9, No. 2; green shapes represent a selection of the IR-structures determined for the sequence by Grachten's IR-parser; red lines indicate the note the structure is associated with. Blue numbers represent the IR-Arch features described below: the numbers indicate the position of the soprano note with respect to the next point with strong closure.

xD Dyadic Unit involving only 2 notes, spanning an interval of x semitones. $B\flat$ and G are 9 semitones apart, which makes the instance in figure 4.4 a **9D**.

ID Intervallic Duplication, a small interval followed by an identical small interval, in different registral directions.

VR Registral Reversal, a large interval followed by a large interval, different registral directions.

[VR] Retrospective Registral Reversal, a VR that was determined in retrospect: the first interval is too small to qualify as a large interval, but together with the following interval, in retrospect, VR seems the best fitting situation.

IP Intervallic Process, a small interval followed by a similar small interval in different registral directions.

P Process, two small intervals in the same registral direction.

4.2.3 Harmonic features

Harmonic features describe perceptual aspects related to consonance. Possible characterizations include harmonic progression throughout the piece, identifying harmonic tension and relief, and the consonance between the melody notes and the local harmony. Both are based on automatic harmonic analysis of the piece, a process called key-finding. Temperley developed a dynamic programming algorithm to establish the most likely sequence of keys for an entire piece [116]. The resulting sequence is locally stable: key does not change on spontaneous, local deviations from the current key and only switches when the transition is stable over a longer period. The sequence of keys can serve as a basis for a functional analysis of the local harmonies, or as part of a structural analysis of the piece. However, for our purpose, it is more useful to known how the local harmony changes from onset to onset than to know where the rather few changes in key occur: the former may give new information for each onset, the latter only for a handful of onsets in the entire piece.

We collect all pitches that occur within a beat and its predecessor and calculate which key has the highest probability given the set of pitches. The calculation is based on key profiles constructed by Temperley from the Essen Folk Song Corpus, where all keys have been manually labelled throughout the corpus. The profiles list occurrence probabilities for all steps of the diatonic scale. Figure 4.5 shows the key estimate for the opening bars of Chopin's Nocturne Op. 9 No. 2. According to the key profiles, the dominant is slightly more likely to occur than the tonic, which explains why the opening note, $B\flat$, is labelled as $E\flat$ major. The change to C minor is obviously incorrect from a musicological point of view. However, from a probabilistic point of view, the $C\flat$, which is viewed as a B, is equally likely in $E\flat$ Major as the minor sixth, as in C minor as the major seventh. The G in the melody voice, and the two $A\flat$ (one on the preceding onset), yield a higher probability as the fifth and minor sixth in C minor, than as major third and fourth in $E\flat$ major. The change to F Major in the second bar on beat 3 is equally surprising and incorrect. However, C occurs 3 times in the collected pitches from beats 3 and 4, and, apart from being the tonic in C Major (which is the correct key here), is also the dominant in F Major. In addition, $B\flat$ has a higher probability of being the subdominant to F Major than the minor seventh in C Major, which tips the scale in favor of F Major.

Even though the analysis does not work well from a musicological point of view, it is at least consistent in the kind of errors it makes, and thus still an informative score descriptor. We base the following harmonic features on the harmonic analysis:

Local Consonance (continuous): Given the most probable local harmony, the consonance of a melody note within the estimated harmony is judged using the key-profiles proposed by Krumhansl and Kessler [67].

Consonance Difference (continuous): The difference between two successive local consonance values.

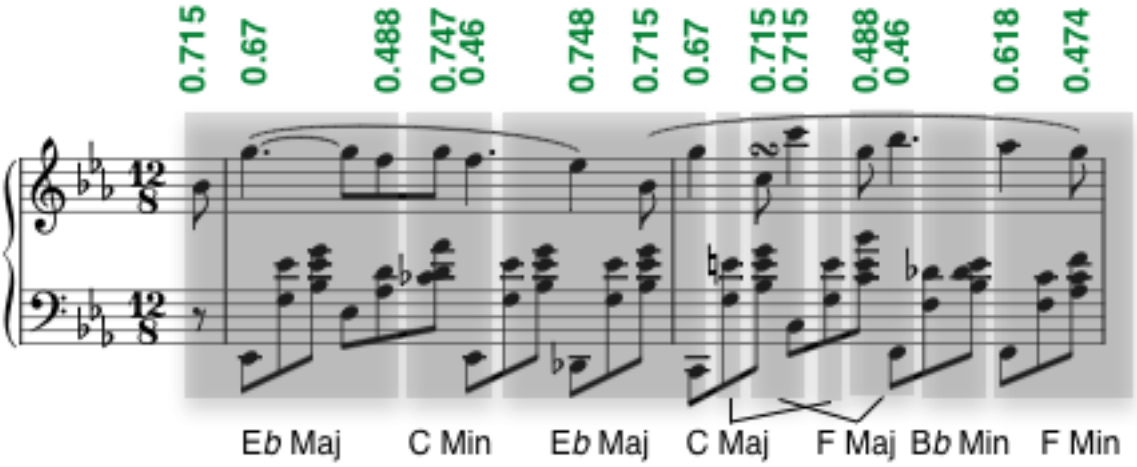


Figure 4.5: Result of an automatic harmonic analysis of the opening bars of Chopin Nocturne Op. 9, No. 2. Grey Areas indicate sequences where the key is the same, numbers in green represent the *local Consonance* feature value of the soprano note on the corresponding onset.

4.2.4 Phrase related features

The phrase structure in pieces can be coded by marking the phrase boundaries. However, much more information can be provided if instead the distance to the next boundary is used as a descriptor. Manual phrase analysis can give a more accurate segmentation than the one estimated from the Implication-Realization analysis. However, it requires

somebody with musical training and lots of experience to provide such an analysis. While for parts of our data, the Mozart sonatas, such an analysis is available (provided by Werner Goebl), for the Chopin data we have to rely on the IR approximation.

IR Arch (continuous): Points of closure – situations where listeners might expect a caesura – can be thought of as an approximation of phrase boundaries. The IR-model calculates those on the basis of changes in melodic, harmonic, and rhythmic intervals (see section 4.2.5). We use the distance to the next point of strong closure as a continuous phrase feature. See Figure 4.4 for an example.

4.2.5 Narmour's Implication-Realization (IR) model

The Implication-Realization (I-R) Model proposed by Narmour [79, 78] is a cognitively motivated model of musical structure. It tries to describe explicitly the patterns of listener expectation with respect to the continuation of the melody. It applies the principles of Gestalt theory to melody perception, an approach introduced by Meyer [72]. The model describes both the continuation implied by particular melodic intervals and the extent to which this (expected) continuation is actually realized by the following interval. Schellenberg [106] provides evidence for the validity of the model: across different levels of musical education and background listeners' expectations were predicted successfully. Grachten [46] not only provides a short introduction to the processes involved but also an implementation to automatically analyze a score accordingly.

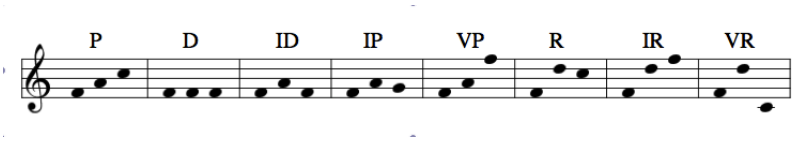


Figure 4.6: Examples of eight IR-structures (picture from [46])

.

Two main principles of the theory concern the direction and size of melodic intervals: (1) Small intervals imply a following interval in the same registral direction, and large intervals imply a change in registral direction. (2) A small interval implies a following similarly-sized interval, a large interval implies a smaller interval. Based on these two principles, melodic patterns, or *structures*, can be identified that either satisfy or violate the implications predicted by the principles. Figure 4.6 shows eight such structures:

Process (P), Duplication(D), Intervallic Duplication (ID), Intervallic Process (IP), Registral Process (VP), Reversal (R), Intervallic Reversal (IR), and Registral Reversal (VR). The Process structure, for instance, satisfies both registral and intervallic implications. Intervallic Process satisfies the intervallic difference principle, but violates the registral implication.

Another notion derived from the concept of implied expectations is *closure*, which refers to situations in which listeners might expect a caesura. In the IR model, closure can be evoked along several dimensions of the music: intervallic progression, metrical position, rhythm, and harmony. The accumulated degrees of closure in each dimension constitute the perceived overall closure at any point in the score. Occurrences of strong closure may coincide with a more commonly used concept of closure in music theory that refers to the completion of a musical entity, for example a phrase. Hence, calculating the distance of each note to the nearest point of closure can provide a segmentation of a piece similar to phrasal analysis.

4.3 Performance Model

The performance model is the interface through which a rendering system can manipulate the expressive dimensions of a performance. As we deal with piano performances we focus on *articulation*, *loudness*, and *tempo*. Following the usual nomenclature in machine learning, we call the different dimensions performance *targets*.

4.3.1 Articulation

Articulation in music refers to the transition between notes. The possibilities and techniques vary for different types of instruments. In case of the piano, the dominant aspect is the degree to which two notes are joined together: the smaller the audible gap between two successive notes, the more *legato* (it.: "joined together") the first one becomes; the larger the gap, the more *staccato* (it.: "detached"). Some articulation marks (e.g. *martellato* and *marcato*) involve in their realization also the loudness of the affected notes. This is not covered in this definition of articulation but subsumed in the loudness component.

Articulation: Let $ioi^s_{i,i+1}$ and $ioi^p_{i,i+1}$ be the score and performance IOIs between the successive notes s_i and s_{i+1}, and dur^s_i and dur^p_i the nominal score duration and

the played duration of s_i respectively. The articulation art_i of a note s_i is defined as

$$art_i = \frac{ioi^s_{i,i+1} * dur^p_i}{dur^s_i * ioi^p_{i,i+1}}. \tag{4.1}$$

4.3.2 Loudness

The loudness[1], of a performed note is characterized as the ratio between the loudness of the note and the mean loudness of the performance. Logarithms are used to scale the values to a range symmetrical around zero, with values above 0 being louder than average and those below 0 softer than average.

Loudness: Let mld_i be the midi loudness of note s_i, and n the number of score notes in the piece. The loudness mld_i is then calculated by

$$mld_i = \log \frac{mld_i}{\frac{1}{n}\sum_j mld_j}. \tag{4.2}$$

4.3.3 Tempo

We define the local tempo at a note through the notion of inter-onset-intervals (IOI), i.e., the time between two successive notes. Relating the IOI prescribed by the score (*score IOI*) and the IOI of the same two notes in the performance (*performance IOI*), determines if the second note was placed early, on time, or late[2]. The description is independent of the absolute tempo and focuses on changes.

We refer to the sequence of ratios between score IOIs and performance IOIs as *complete tempo curve*:

Complete Tempo Curve: Let s_i and s_{i+1} be two successive melody notes, p_i and p_{i+1} the corresponding notes in the performance, $ioi^s_{i,i+1}$ the score IOI, $ioi^p_{i,i+1}$ the performance IOI of the two notes[3], l_s the duration of the complete piece in beats,

[1] Computer-controlled Pianos measure loudness by measuring the velocity at which a hammer strikes a string. Especially in the context of MIDI, this lead to loudness sometimes being referred to as *velocity*.

[2] Tempo, of course, needs more than one note to manifest. When we talk about the tempo of one note we actually talk about the relative placement of the following note.

[3] The unit of the duration does not matter in this case, as it cancels out with the unit of the complete duration of the performance

and l_p the length of the performance. The IOI ratio $ioiR_i$ of s_i is then defined as:

$$ioiR_i = \log \frac{ioi^p_{i,i+1} * l_s}{ioi^s_{i,i+1} * l_p}. \tag{4.3}$$

Normalising both score and performance IOIs to fractions of the complete score and performance respectively, makes this measure independent of the actual tempo. Loosely speaking this relates the performed duration to the expected duration. The logarithm is used to scale the values to a range symmetrical around zero, where $ioiR_i > 0$ indicates a prolonged IOI, i.e., a tempo slower than notated, and $ioiR_i < 0$ indicates a shortened IOI, i.e., a tempo faster than notated.

4.4 YQX - the First Step

Our performance rendering system, called YQX, models the dependencies between the score model, the set of *features*, and the performance model, the set of *targets*, by means of a probabilistic network. For an introduction to basic concepts and notations of probabilistic networks, see appendix A. The network consists of several interacting nodes representing different features and targets. Each node is associated with a probability distribution over the values of the corresponding feature or target. A connection between two nodes in the graph implies a conditioning of one feature or target distribution on the other. Discrete score features (the set of which we call $\mathbf{Q}$) are associated with discrete probability tables, while continuous score features ($\mathbf{X}$) are modelled by Gaussian distributions. The predicted performance characteristics, the set of targets $\mathbf{Y} = \{articulation, loudness, tempo\}$, are continuously valued and conditioned on the set of discrete and continuous features. Figure 4.7 shows the general layout. The semantics is that of a linear Gaussian model [76]. This implies that the case of a continuous distribution parenting a continuous distribution is implemented by making the mean of the child distribution linearly dependent on the value of the condition.

Mathematically speaking, a target, viewed as a continuous random variable Y, is modelled as a conditional distribution $P(Y|\mathbf{Q}, \mathbf{X})$. Following the linear Gaussian model, this is a Gaussian distribution $\mathcal{N}(y; \mu, \sigma^2)$ with the mean μ varying linearly with $\mathbf{X}$, and with a fixed variance σ^2. Given specific values $\mathbf{Q} = \mathbf{q}$ and $\mathbf{X} = \vec{x}$ (treating the real-valued set of continuous score features as a vector):

$$\mu = d_{\mathbf{q}} + \vec{k}_{\mathbf{q}} \cdot \vec{x} \tag{4.4}$$

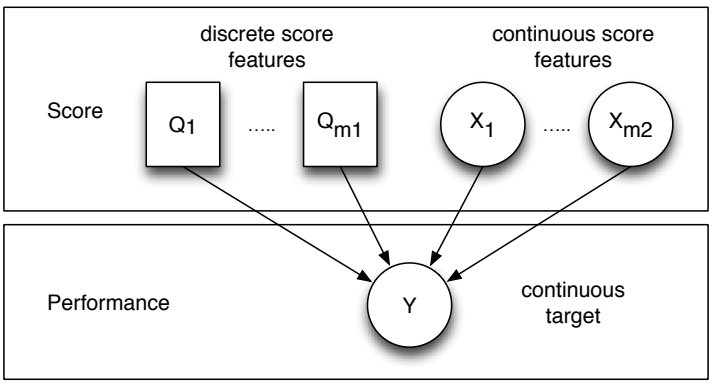


Figure 4.7: The probabilistic network forming the YQX system.

where $d_{\mathbf{q}}$ and $\overrightarrow{k}_{\mathbf{q}}$ are estimated from the data by least squares linear regression. The average residual error of the regression is the variance σ^2 of the distribution.

From a practical point of view, training this model only requires estimating the probability distributions. This is done in two steps: first, we collect all instances in the data that share the same combination of discrete feature values and build a joint probability distribution of the continuous features and targets of these instances. This implements the conditioning on the discrete features $\mathbf{Q}$. In the second step, an affine, linear function (equation 4.4) is determined via linear regression. The function relates the mean μ of the target distribution to the values $\overrightarrow{x}$ of the continuous features $\mathbf{X}$. The average residual of the regression provides the variance of the target distribution. Hence, the target distribution is a gaussian, that is conditioned on the discrete features and linearly dependent on the continuous features. Estimating $\overrightarrow{k}_{\mathbf{q}}$ and $d_{\mathbf{q}}$ for all possible $\mathbf{q} \in \mathcal{D}(\mathbf{Q})$ constitutes the training phase.

Performance prediction is done note by note. The score features of a note are entered into the network as evidence $\overrightarrow{x}$ and $\mathbf{q}$. The instantiation of the discrete features determines the appropriate probability table and the parameterisation $d_{\mathbf{q}}$ and $\overrightarrow{k}_{\mathbf{q}}$, and the continuous features are used to calculate the mean μ of the target distribution. This value is used as the prediction for the specific note. We ignore all dependencies and interactions that may exist between tempo, loudness, and articulation and create models and predictions for each target separately.

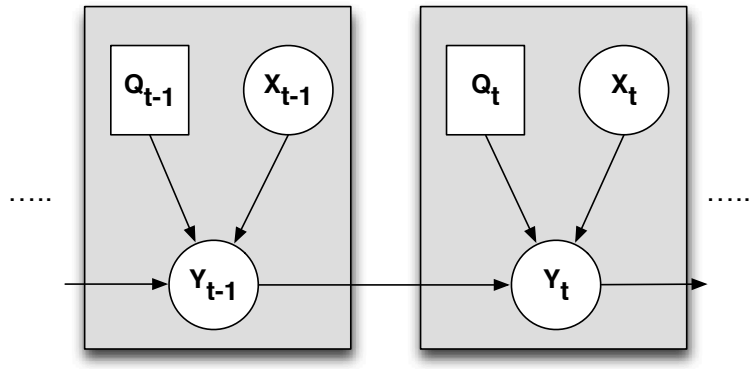


Figure 4.8: The network unfolded in time. $\mathbf{Q_{t-1}}$ are instantiations of $\{Q_1, \cdots, Q_{m_1}\}$ at time $t-1$; $\mathbf{X_{t-1}}$ likewise for $\{X_1, \cdots, X_{m_2}\}$.

4.5 Introducing performance context - Step Two

The predictions of the basic YQX system are note-wise; each prediction depends only on the score features at that particular score onset. In a real performance this is of course not the case: typically, dynamics or tempo evolve gradually. Clearly, this necessitates awareness of the surrounding expressive context.

In this section we present two extensions to the system that both introduce a dynamic component by incorporating the prediction made for the preceding score note into the prediction of the current score note. Graphically, this corresponds to first unfolding the network in time and then adding an arc from the target in time-step $t-1$ to the target in time-step t. Figure 4.8 shows the unfolded network. This should lead to smoother and more consistent performances with less abrupt changes and, ideally, to an increase in the overall prediction quality.

The context-aware prediction can be done in two different ways: (1) Using the previous target simply as an additional parenting probability distribution to the current target allows local optimisation with respect to one preceding prediction. Minimal adaptation has to be made to the algorithm (see 4.5.1). (2) Using an adaptation of the Viterbi decoding in Hidden Markov Models results in a predicted series that is optimal with respect to the complete piece (see 4.5.2).

4.5.1 YQX with local maximisation

The first method is rather straightforward: We use the linear Gaussian model and treat the additional parent (the target Y_{t-1}) to the target Y_t as an additional feature that we calculate from the performance data. In the training process, the joint distribution of the continuous features, the target Y_t, and the target in the previous time-step Y_{t-1} given the discrete score features – in mathematical terms $P(Y_{t-1}, Y_t, \overrightarrow{x}_t|\mathbf{q}_t)$ – is estimated. This alters the conditional distribution of the target Y_t to $P(Y_t|\mathbf{Q}, \mathbf{X}, Y_{t-1}) = \mathcal{N}(y_t; \mu, \sigma^2)$ with[4]

$$\mu = d_{\mathbf{q},y_{t-1}} + \overrightarrow{k}_{\mathbf{q},y_{t-1}} \cdot (\overrightarrow{x}, y_{t-1}).$$

The prediction phase is equally straightforward. The already predicted y_{t-1} is treated like additional evidence, and, as in the simple model, the mean of $P(Y_t|\mathbf{q}_t, \overrightarrow{x}_t, y_{t-1})$ is used as the prediction for the score note in time-step t. This is the value with the highest local probability.

4.5.2 Global Optimization

The second approach drops the concept of a linear Gaussian model. The goal is to construct a performance with maximum probability with respect to the complete history of predictions made up to that point.

In the training phase the joint gaussian distributions of the continuous features $\mathbf{X} = \{X_1, \cdots, X_{m_1}\}$, the target $\mathbf{Y_t}$, and the previous target $\mathbf{Y_{t-1}}$ conditioned on the set of discrete features $\mathbf{Q} = \{Q_1, \cdots, Q_{m_2}\}$ are estimated from the training data. In mathematical terms those are $P(Y_{t-1}, Y_t, \mathbf{X}|\mathbf{Q})$, and, as before, one joint distribution exists per possible instantiation $\mathbf{q} \in \mathcal{D}(\mathbf{Q})$ of the discrete features $\mathbf{Q}$. Mean and variance of the distributions can be estimated via relative frequencies (component-wise mean and covariance) or with a prior distribution. The latter is especially advisable if the number of discrete features and their respective domains are large: if the training data does not provide samples for each possible combination of discrete score features, prior distributions provide a controllable fallback in case the underrepresented combinations occur in the test data. We use a simple zero-mean gaussian with unity covariance as a default. Simply put, the training of the model only requires grouping together all samples sharing the same combination of values in the discrete feature, and calculating mean and

[4]The construct $(\overrightarrow{x}, y_{t-1})$ is a concatenation of the vector $\overrightarrow{x}$ and the value y_{t-1} leading to a new vector of dimension $dim(\overrightarrow{x}) + 1$.

covariance of the continuous features for all the samples in the separate groups.

In the prediction phase, given the score features of the test piece, we construct a sequence of predictions that maximizes the conditional probability with respect to the complete history of predictions made up to that point. This is calculated in analogy to the Viterbi-decoding in Hidden Markov Models, which tries to find the best explanation for the observed data [63]. Aside from the fact that the roles of evidence nodes and query nodes are switched, the main conceptual difference is that – unlike the HMM setup, which uses tabular distributions – our approach must deal with continuous distributions. This rules out the dynamic programming algorithm usually applied. Kalman filters also use continuous state-space variables, but use a different system of connections between the nodes [76]. We have a very specific network layout, and the algorithm does not have to account for arbitrary combinations of connections and variables. We can therefore search for a specialized, analytical rather than an algorithmic solution to our inference problem. As in the Viterbi algorithm, the calculation is done in two steps: a forward and a backward sweep. In the forward movement the most probable target is calculated relative to the previous time-step. In the backward movement, knowing the final point of the optimal path, the sequence of predictions is found via backtracking through all time-steps. The prediction phase, the forward and backward calculation is explained in detail in the following.

The forward calculation

Let $\overrightarrow{x}_t, \mathbf{q}_t$ be the sets of continuous and discrete features at time t, and N be the number of data points in a piece. Further, let $\alpha(Y_t)$ be a probability distribution over the values y_t of Y_t, indicating the probability that the optimal path from time-steps 1 to t ends in y_t. By means of a recursive formula, $\alpha(Y_t)$ can be calculated for all time-steps of the unfolded network:

$$\alpha(Y_1 = y_1) = P(Y_1 = y_1 | \mathbf{x}_1, \mathbf{q}_1) \tag{4.5}$$

$$\alpha(Y_t = y_t) = \max_{y_{t-1} \in \mathbb{R}} \left[P(Y_t = y_t, Y_{t-1} = y_{t-1} | \overrightarrow{x}_t, \mathbf{q}_t) \cdot \alpha(Y_{t-1} = y_{t-1}) \right] \tag{4.6}$$

This formula can be interpreted as follows: Assuming that we know for all the target values y_{t-1} in time step $t-1$ the probability of being part of the optimal path, we can calculate for each target value y_t in time step t the predecessor that yields the highest probability for each specific y_t of being on the optimal path. In the backward movement we start with the most probable final point of the path (the mean of the

last α) and then backtrack to the beginning by choosing the best predecessors. As we cannot calculate the maximum over all $y_{t-1} \in \mathbb{R}$ directly, we need an analytical way of calculating $\alpha(Y_t)$ from $\alpha(Y_{t-1})$, which we derive below. We will also show that $\alpha(Y_t)$ remains Gaussian through all time-steps. This is particularly important because we rely on the parametric representation using mean and variance. Anticipating our proof that the $\alpha(Y_t)$ are Gaussian, we refer to the mean and variance as $\mu_{\alpha,t}$ and $\sigma^2_{\alpha,t}$. For the sake of simplicity, we abbreviate the probabilities of specific values such as $\alpha(Y_t = y_t)$ and $P(Y_t = y_t, Y_{t-1} = y_{t-1})$ to the short forms $\alpha(y_t)$ and $P(y_t, y_{t-1})$ respectively. In cases in which not a specific value but the distribution as a whole is addressed, we write $\alpha(Y_t)$ and $P(Y_t, Y_{t-1})$.

Indexing the target variables by time steps (t and $t-1$) might be misleading, as actually only two and not N target variables exist. The variable called Y_t represents the current prediction, Y_{t-1} the prediction made in the previous time step. Accordingly, there is only two sets of feature variables, continuous ($\mathbf{X} = \{X_1, \cdots, X_{m_1}\}$) and discrete ($\mathbf{Q} = \{Q_1, \cdots, Q_{m_2}\}$), that take on different values in every time step t ($\overrightarrow{x}_t = (x_{1,t}, \cdots, x_{m_1,t})$ and $\mathbf{q}_t = \{q_{1,t}, \cdots, q_{m_2,t}\}$ respectively).

Given evidence $\mathbf{q}_t$ of the states of the discrete features at time t the joint probability of targets Y_t and Y_{t-1} and continuous features $\mathbf{X} = \{X_1, \cdots, X_{m_1}\}$ is the following multivariate gaussian:

$$P(Y_{t-1}, Y_t, \mathbf{X}|\mathbf{q}_t) \propto \mathcal{N}(\begin{pmatrix} y_{t-1} \\ y_t \\ x_1 \\ \vdots \\ x_{m_x} \end{pmatrix}; \begin{pmatrix} \mu_{y_{t-1}} \\ \mu_{y_t} \\ \mu_{x_1} \\ \vdots \\ \mu_{x_{m_x}} \end{pmatrix}; \Sigma_t), \tag{4.7}$$

with $\mu_{y_{t-1}}$, μ_{y_t}, and $\mu_{x_1}, \cdots, \mu_{x_{m_x}}$ and the following covariance matrix Σ_t as estimated from the data:

$$\Sigma_t = \begin{bmatrix} \sigma^2_{y_{t-1}} & \sigma^2_{y_{t-1},y_t} & \sigma^2_{x_1,y_{t-1}} & \cdots & \sigma^2_{x_{m_x},y_{t-1}} \\ \sigma^2_{y_{t-1},y_t} & \sigma^2_{y_t} & \cdots & \cdots & \sigma^2_{x_{m_x},y_t} \\ \vdots & \cdots & \sigma^2_{x_1} & \cdots & \vdots \\ \sigma^2_{x_{m_x},y_t} & \cdots & \cdots & \cdots & \sigma^2_{x_{m_x}} \end{bmatrix}$$

$$= \left[\begin{array}{cc|c} \sigma^2_{y_{t-1}} & \sigma^2_{y_{t-1},y_t} & C^T \\ \sigma^2_{y_{t-1},y_t} & \sigma^2_{y_t} & \\ \hline C & & B \end{array}\right]$$

The conditional joint distribution of targets Y_t and Y_{t-1} given both discrete and continuous features as evidence can be calculated from 4.7 in the following form (for details see e.g. [96], or appendix A):

$$P(Y_{t-1}, Y_t | \mathbf{q}_t, \overrightarrow{x}_t) \propto \mathcal{N}(\begin{pmatrix} y_{t-1} \\ y_t \end{pmatrix}; \begin{pmatrix} \hat{\mu}_{y_{t-1}} \\ \hat{\mu}_{y_t} \end{pmatrix}; \hat{\Sigma}_t) \tag{4.8}$$

with

$$\begin{pmatrix} \hat{\mu}_{y_{t-1}} \\ \hat{\mu}_{y_t} \end{pmatrix} = \begin{pmatrix} \mu_{y_{t-1}} \\ \mu_{y_t} \end{pmatrix} + CB^{-1} \begin{pmatrix} x_1 - \mu_{x_1} \\ \vdots \\ x_{m_x} - \mu_{x_{m_x}} \end{pmatrix}, \text{and}$$

$$\hat{\Sigma}_t = \begin{bmatrix} \sigma^2_{y_{t-1}} & \sigma^2_{y_{t-1}, y_t} \\ \sigma^2_{y_{t-1}, y_t} & \sigma^2_{y_t} \end{bmatrix}.$$

Using the same technique, the conditional probability of the previous target given the current target under the current evidence, $P(Y_{t-1}|Y_t, \mathbf{q}_t, \overrightarrow{x}_t)$, can then be formulated as follows:

$$P(Y_{t-1}|Y_t, \mathbf{q}_t, \overrightarrow{x}_t) \propto \mathcal{N}(y_{t-1}; \tilde{\mu}_{y_{t-1}}, \tilde{\sigma}^2_{y_{t-1}}) \tag{4.9}$$

$$\tilde{\mu}_{y_{t-1}} = \hat{\mu}_{y_{t-1}} + \frac{\sigma^2_{y_t, y_{t-1}}(y_t - \hat{\mu}_{y_t})}{\sigma^2_{y_t}}$$

$$\tilde{\sigma}^2_{y_{t-1}} = \sigma^2_{y_{t-1}} - \frac{\sigma^4_{y_t, y_{t-1}}}{\sigma^2_t}$$

From the definition of the conditional joint distribution of Y_t and Y_{t-1} (equation 4.8) the marginal distribution of Y_t (conditioned on $\mathbf{q_t}$ and $\overrightarrow{x}_t$) can easily be read:

$$P(Y_t|\mathbf{q}_t, \overrightarrow{x}_t) \propto \mathcal{N}(y_t; \tilde{\mu}_{y_t}, \tilde{\sigma}^2_{y_t}) \tag{4.10}$$

$$\tilde{\mu}_{y_t} = \hat{\mu}_{y_t}$$

$$\tilde{\sigma}^2_{y_t} = \sigma^2_{y_t}$$

Under the basic laws of conditional probability, the inductive definition of α (eq. 4.6) can be rewritten to replace the joint distribution with a conditional distribution

(the conditioning on $\mathbf{q}_t$, $\overrightarrow{x}_t$ is omitted for simplicity):

$$\alpha(Y_t) = \max_{y_{t-1}\in\mathbb{R}}\left[P(Y_{t-1} = y_{t-1}|Y_t)\cdot P(Y_t)\cdot\alpha(Y_{t-1} = y_{t-1})\right] \tag{4.11}$$

$$= \max_{y_{t-1}\in\mathbb{R}}\left[P(Y_{t-1} = y_{t-1}|y_t)\cdot\alpha(Y_{t-1} = y_{t-1})\right]\cdot P(Y_t) \tag{4.12}$$

$$= \max_{y_{t-1}\in\mathbb{R}}\left[\mathcal{N}(y_{t-1};\tilde{\mu}_{y_{t-1}},\tilde{\sigma}^2_{y_{t-1}})\cdot\alpha(Y_{t-1} = y_{t-1})\right]\cdot P(Y_t) \tag{4.13}$$

The distribution $P(Y_t)$ in equation 4.11 can be put outside of the scope of the maximum because it is does not depend on y_{t-1}. In 4.12 we can then plug in the distribution calculated in 4.9. Assuming that $\alpha(Y_{t-1})$ is Gaussian, multiplying the two distributions under the maximum in 4.13 results in the following Gaussian distribution:

$$\mathcal{N}(y_{t-1};\mu^*_t,\sigma^{*2}_t) \tag{4.14}$$

with

$$\mu^*_t = \sigma^{*2}_t\left(\frac{\tilde{\mu}_{y_{t-1}}}{\tilde{\sigma}^2_{y_{t-1}}} + \frac{\mu_{\alpha,t-1}}{\sigma^2_{\alpha,t-1}}\right)$$

$$\sigma^{*2}_t = \frac{\tilde{\sigma}^2_{y_{t-1}}\cdot\sigma^2_{\alpha,t-1}}{\tilde{\sigma}^2_{y_{t-1}} + \sigma_{\alpha,t-1}}$$

The normalising constant z of $\mathcal{N}(y_{t-1};\mu^*_t,\sigma^{*2}_t)$ itself is gaussian in the means of both factors of the multiplication, $\tilde{\mu}_{y_{t-1}}$ and $\mu_{\alpha,t-1}$:

$$z = \frac{1}{\sqrt{2\pi|\tilde{\sigma}^2_{y_{t-1}} + \sigma^2_{\alpha,t-1}|}} e^{\left(\frac{-(\tilde{\mu}_{y_{t-1}}-\mu_{\alpha,t-1})^2}{2(\tilde{\sigma}^2_{y_{t-1}}+\sigma^2_{\alpha,t-1})}\right)} \tag{4.15}$$

Later, z will be multiplied with a Gaussian distribution over y_t. Hence, z must be transformed into a distribution over the same variable. By finding a y_t such that the exponent in eq. 4.15 equals 0 we can construct the mean μ_z and variance σ^2_z of z as a distribution over y_t. Note that the variable $\tilde{\mu}_{y_{t-1}}$ is dependent on y_t due to the conditioning of $P(Y_{t-1}|Y_t)$ on y_t. With respect to y_t, the normalizing constant z follows the distribution in equation 4.16:

$$z \propto \mathcal{N}(y_t;\mu_z,\sigma^2_z) \tag{4.16}$$

$$\mu_z = -\frac{\sigma^2_{y_t}\cdot(\tilde{\mu}_{y_{t-1}} + \mu_{\alpha,t-1}) + \mu_{y_t}\cdot\sigma^2_{t,t-1}}{\sigma^2_{t,t-1}}$$

$$\sigma^2_z = \tilde{\sigma}^2_{y_{t-1}} + \sigma^2_{\alpha,t-1}$$

The calculation of $\alpha(Y_t)$ (equation 4.13) can now be simplified further:

$$\begin{aligned}\alpha(Y_t) &\propto \max_{y_{t-1}\in\mathbb{R}} \left[\mathcal{N}(y_{t-1}; \tilde{\mu}_{y_{t-1}}, \tilde{\sigma}^2_{y_{t-1}}) \cdot \alpha(Y_{t-1} = y_{t-1})\right] \cdot P(Y_t) \\ &= \max_{y_{t-1}\in\mathbb{R}} \left[z \cdot \hat{\mathcal{N}}(y_{t-1}; \mu^*_t, \sigma^{*2}_t)\right] \cdot P(Y_t) \qquad (4.17) \\ &\propto \max_{y_{t-1}\in\mathbb{R}} [\mathcal{N}(y_t; \mu_z, \sigma^2_z) \cdot \hat{\mathcal{N}}(y_{t-1}; \mu^*_t, \sigma^{*2}_t)] \cdot P(Y_t) \qquad (4.18)\end{aligned}$$

In equation 4.17 we separate $\mathcal{N}(y_{t-1}; \mu^*_t, \sigma^{*2}_t)$ from its normalizing constant z, leaving an unnormalized distribution $\hat{\mathcal{N}}(y_{t-1}; \mu^*_t, \sigma^{*2}_t)$. As z is independent of y_{t-1}, it is not affected by the calculation of the maximum in equation 4.18. Therefore we can put $\mathcal{N}(y_t; \mu_z, \sigma^2_z)$ outside of the scope of the maximum (equation 4.19). The y_{t-1} that maximizes the remaining $\hat{\mathcal{N}}(y_{t-1}; \mu^*_t, \sigma^{*2}_t)$ is the mean of the distribution, μ^*_t. As the distribution $\hat{\mathcal{N}}(y_{t-1}; \mu^*_t, \sigma^{*2}_t)$ is unnormalized, the result is 1 (equation 4.20):

$$\begin{aligned}\alpha(Y_t) &\propto \max_{y_{t-1}\in\mathbb{R}} [\hat{\mathcal{N}}(y_{t-1}; \mu^*_t, \sigma^{*2}_t)] \cdot \mathcal{N}(y_t; \mu_z, \sigma^2_z) \cdot P(Y_t) \qquad (4.19) \\ &= 1 \cdot \mathcal{N}(y_t; \mu_z, \sigma^2_z) \cdot P(Y_t). \qquad (4.20)\end{aligned}$$

The distribution $P(Y_t)$ (the conditional, marginal distribution $P(Y_t|\mathbf{q}_t, \overrightarrow{x}_t)$ defined in equation 4.10) is Gaussian by design, and hence the remaining product again results in a Gaussian and a normalising constant. As the means of both factors are fixed, the normalising constant in this case is a single factor. The mean $\mu_{\alpha,t}$ and variance $\sigma^2_{\alpha,t}$ of $\alpha(Y_t)$ follow:

$$\alpha(Y_t) \propto \mathcal{N}(y_t; \mu_{\alpha,t}, \sigma^2_{\alpha,t}) \qquad (4.21)$$

$$\sigma_{\alpha,t} = \frac{\sigma^2_t \cdot \sigma^2_z}{\sigma^2_t + \sigma^2_z} \qquad (4.22)$$

$$\mu_{\alpha,t} = \sigma_{\alpha,t}\left(\frac{\mu_z}{\sigma^2_z} + \frac{\tilde{\mu}_{y_t}}{\sigma^2_{y_t}}\right). \qquad (4.23)$$

Thus, $\alpha(Y_t)$ is Gaussian in y_t, assuming that $\alpha(Y_{t-1})$ is Gaussian. Since $\alpha(Y_1)$ is Gaussian, it follows that $\alpha(Y_t)$ is Gaussian for $1 \leq t \leq N$. This equation shows that the mean and variance of $\alpha(y_t)$ can be computed recursively using the mean $\mu_{\alpha,t-1}$ and variance $\sigma^2_{\alpha,t-1}$ of $\alpha(Y_{t-1})$. The parameters of $\alpha(Y_1)$ equal μ_{y_1} and $\sigma^2_{y_1}$, which are the mean and the variance of the distribution $P(Y_1|\overrightarrow{x}_1, \mathbf{q}_1)$, and are estimated from the data.

The backward calculation

Once the mean and variance μ_t, σ_t^2 of $\alpha(y_t)$ are known for $1 \leq t \leq N$, the optimal sequence $\dot{y}_1, ..., \dot{y}_N$ can be calculated. The final point of the sequence $\dot{y}_N$ is calculated first. By design, the value $\alpha(Y_t = y_t)$ indicates the probability that the optimal path from time-steps 1 to t ends in y_t. This means that the optimal path ends in the mean $\mu_{\alpha,N}$ as it is the most probable value of the final distribution $\alpha(Y_N)$. The predecessor to $\dot{y}_N$, $\dot{y}_{N-1}$, is the value of Y_{t-1} that maximized equation 4.19. The optimal sequence is calculated with the following two equations. The definition of μ_t^* (equation 4.14) is dependent on y_t, which, in this case is substituted with $\dot{y}_t$.

$$\dot{y}_N = \mu_{\alpha,N} \tag{4.24}$$

$$\dot{y}_{t-1} = \underset{y_{t-1}}{\operatorname{argmax}} \left[\mathcal{N}(y_{t-1}; \mu_t^*, \sigma_t^{*2}\right] \tag{4.25}$$

$$= \mu_t^* \tag{4.26}$$

4.6 Composite Performance Dimensions - Step Three

Up to now, expressive dimensions were considered as atomic entities, in the sense that they represent one aspect of music performance that manifests in a single number per note. Alternative definitions for tempo and loudness are introduced in the following section, which constitute the final extension to our rendering system. Regarding loudness, we take into account the expressive annotations that are given in the score; regarding tempo, we propose a way to split up the complete tempo curve into two components. Consequently, instead of expecting a single model to come up with an explanation for all facets of performance tempo or loudness, we can use several specialized models for the different aspects.

4.6.1 Loudness & Performance Directives

Section 4.1.4 discusses the importance of expressive annotations in the score. Musicians obey and realize those to a certain degree. This implies, and studies exist that support this experimentally [48, 49], that the loudness curve calculated from a performance can, at least partly, be explained by the annotations in the score. Given a loudness curve that only represents the expressive annotations, we can separate that from the performance curve, which reduces the amount of variation left for the model to explain. Obtaining

such a curve, of course, poses problems: (1) Which expressive annotations did the artist obey? This is to a certain degree related to the score edition they used, a question we cannot answer with certainty. However, when it comes to shaping a piece in terms of large gestures and shape, there is substantial common ground among musicians. This, in turn, makes it reasonable to assume that they all work on a basis that, if not identical, at least is somehow similar. (2) How expressive annotations are realized in detail is by no means generic. Shape and extent of each is dependent on context (and artist). Not only is there not one generic way to execute a *crescendo* which is then applied in all relevant situations, but also, even if two pianists obviously both play a *crescendo*, there can be substantial differences in the realization.

Conceptually, this also applies to tempo. However, annotations regarding tempo are much rarer in the Chopin scores than annotations regarding dynamics; in Mozart scores they are practically non-existent. While it is still important for a convincing rendition of a piece to obey the few annotations that are given in the score, the information is too sparse to be significant for the process of learning.

Grachten formulates the problem as a machine learning task [48, 49]: a least squares fitting of a set of basis functions is used to model the influence of notated loudness directives. The loudness trajectory of a piece is thought of as a linear sum of active loudness directives, each represented by a weighted basis function. Given suitable data, like the Magaloff Corpus, the weights can be estimated using least-squares optimization, minimizing the sum of the squared differences between the observed loudness values and the values predicted for the score the summed basis functions. This is the technique we use for our experiments in chapter 5.

We call the approximation of what we assume is the contribution of the score annotations *annotated loudness*. Based on this estimate we define *local loudness* as the residual of the complete loudness curve after removing the annotated loudness.

Local loudness: Let $annL_i$ be the annotated loudness estimate for the loudness for note s_i and vel_i the loudness calculated from the real performance. The local loudness $locL_i$ for s_i is calculated by:

$$locL_i = \frac{vel_i - annL_i}{annL_i}. \tag{4.27}$$

In order to produce a loudness curve vel for a piece, the two elements, rendered expressive annotations $annL$ and prediction of local loudness $locL$, can then be combined

by reversing the decomposition in equation 4.27:

$$vel_i = locL_i \cdot annL_i + annL_i \tag{4.28}$$

4.6.2 Tempo as a composite phenomenon

In music performances, tempo usually refers to a combination of three aspects: (1) *global tempo* refers to the initial tempo prescriptions at the beginning of a score; (2) *local tempo* describes localized tempo trends which, for example, outline larger musical units (e.g. phrases) and realize annotations in the score (like *ritardando* or *accelerando*); (3) *(local) note timing* refers to local (note-wise) deviations from the local tempo that emphasize single notes through delay or anticipation.

Viewing the previously defined complete tempo curve as a composite of local tempo and note timing, we associate *local tempo* with its low-frequency content, which we extract by applying a windowed, moving average. The residual, the curve that remains after subtracting the local tempo from the complete tempo curve, is associated with *note timing*. Formally, we define the two aspects as follows:

Local tempo: Let $ioiR_i$ be the IOI ratio of note s_i, and $n \in \mathbb{N}$ the window length in beats. Let further be $\mathbf{S_i^{\pm n}} = \{s_j : |on_i - on_j| < \frac{n-1}{2}\}$, where on_i is the onset of score note s_i, the set of melody notes that have an onset within the window of n beats surrounding s_i. The local tempo lt_i of the note s_i is calculated by:

$$lt_i = \frac{1}{|\mathbf{S_i^{\pm n}}|} \sum_{\mathbf{S_i^{\pm n}}} ioiR_j. \tag{4.29}$$

Note timing The residual high-frequency content can be considered as the local timing nt_i and, in relation to the local tempo, indicates that a note is either played faster or slower with respect to the local tempo:

$$nt_i = \frac{ioiR_j - lt_i}{lt_i}. \tag{4.30}$$

Figure 4.9 shows the result of applying the decomposition to the tempo curve of Magaloff's performance of the Chopin Mazurka Op. 56 No. 1. How n, the size of the window in the definition of local tempo, is set, influences the result of the decomposition. Informal experiments suggested 4 beats as a reasonable value for the Chopin pieces.

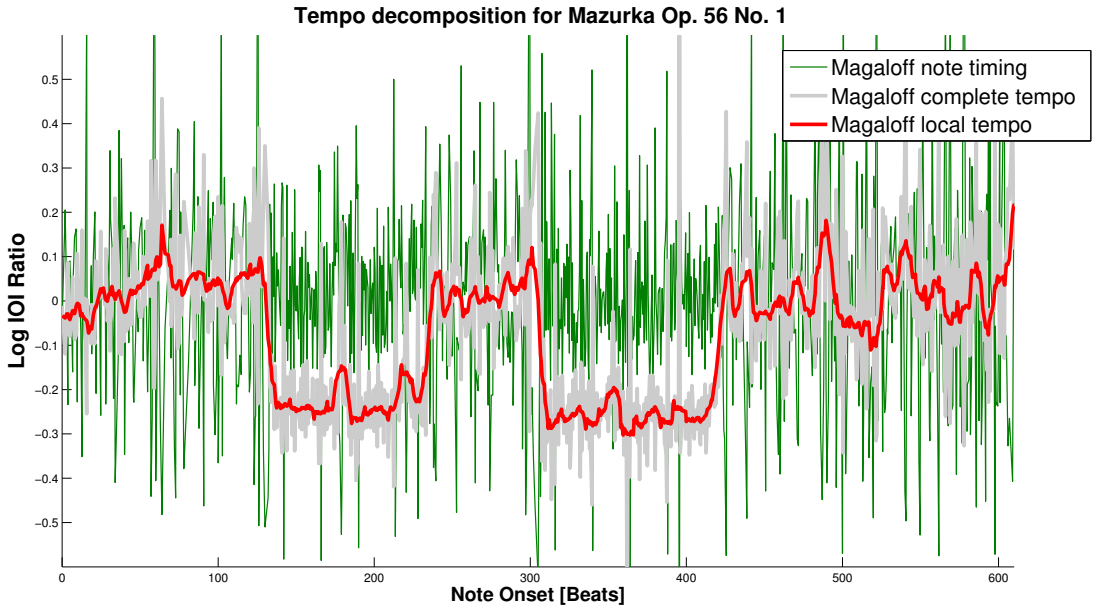


Figure 4.9: Tempo decomposition of Magaloff's performance of Mazurka Op. 56, No. 1: Magaloff's complete tempo (logarithmic IOI-ratios), the local tempo component, and the local timing residual.

Possibilities include making the window size dependent on the meter (bar length) of the piece, or, if a phrase segmentation of the piece is available, varying the window size in relation to the phrase length. However, as of now, we have not been able to establish a well-founded understanding of what constitutes a good split of the complete tempo, or what a local tempo curve should look like.

After predicting the two components separately, reversing the decomposition process in equation 4.30 provides a complete tempo curve for the piece:

$$ioiR_i = nt_i * lt_i + lt_i \tag{4.31}$$

A further notion that comes into play when combining the two predicted curves is *balance*: instead of just reversing the decomposition, an additional influence factor β can be introduced such that either one of the components can be made dominant. Given the local tempo lt_i and note timing nt_i of note s_i, and an influence factor $0 < \beta \leq 1, \beta \in \mathbb{R}$

the final tempo value $ioiR_i$ can be calculated using the following formula:

$$ioiR_i = \widehat{\beta} \cdot nt_i \cdot lt_i + lt_i \tag{4.32}$$

where $\widehat{\beta}$ scales the note timing to a fraction of the local tempo, according to the intended influence:

$$\widehat{\beta} = \frac{1-\beta}{\beta} \cdot \frac{\max_i lt_i}{\max_i nt_i} \tag{4.33}$$

Chapter 5

Evaluation and Experiments

It is important to be looking with eagle-ears.

R. Hagelauer, February 8, 2012

The goal of expressive performance rendering is to automatically generate performances that sound as "human" and "naturally expressive" as possible. The question then of course is, how can "humanness" or "naturalness" be measured? Human perception is exceptionally good at detecting irregularities in music performances. Minuscule differences in any expressive dimensions are perceived, and can draw the line between a note or sequence of notes sounding natural of completely out of place. It is not possible to emulate this ability algorithmically: A plethora of factors is at play – including aspects like musical background, personal taste, and the expectations that are formed accordingly – that are subjective and far too complex to be formalized and quantified. What makes human judgement impossible to implement, also makes it inherently subjective and unsuitable for representative large-scale evaluations across several datasets.

As measuring aesthetic quality in absolute terms is out of the question, a possible alternative is to attempt to judge similarity between computer-rendered and human performances. Although it has severe limitations with respect to musical considerations (further discussed in 5.2), correlation is an easy-to-compute similarity measure for two curves. As such it can serve as a similarity measure for automatic evaluation of expressive rendering systems.

At some point, however, human judgement has to come into play as the only adequate measure of aesthetic quality. The annual rendering contest RENCON offers a scientific platform on which performance rendering systems can be compared and rated by the audience. Although it only provides an evaluation of a few selected pieces, this is the closest an automatically rendered performance can come to an object evaluation.

The experiments described in the following chapter try to paint a picture of the qualities and limitations of our rendering system. The score model, the selection of features characterizing the score, is evaluated first. Its influence on the prediction is

immense. Key is the distribution of the targets across the values of a score feature, which essentially determines the predicted performance. Section 5.3 inspects the separation qualities of score features with respect to the different expressive dimensions and training datasets and describes an experiment searching for the most suitable feature sets for different prediction scenarios.

Using the so generated feature sets, the expressive dimensions are evaluated separately: First, experiments examining the similarity between predicted and original performances give a quantitative perspective of the generalization qualities of the different models. As the merits of correlation as a measure of aesthetic quality are debatable (see section 5.2), I then inspect qualitative aspects of the predicted performance trajectories. This will illustrate the advantages of the different approaches in the various scenarios, and show how the extensions described in sections 4.5 and 4.6 affect and, in some cases, improve the predictions.

The qualitative aspect we tried to emphasize with the different extensions, were based on the author's intuition and experience with the data. They find justification through the final evaluation in the Rendering Contest RENCON, where listener perception validated our models twice. Details of the competitions in 2008 and 2011, both of which we won, are given in section 5.7.

5.1 Data and Experiment Setup

All experiments are based on the following two datasets: (1) the complete works for solo piano by Chopin played by N. Magaloff, the corpus described in chapter 2, and (2) Mozart piano sonatas played by the Viennese pianist Roland Batik. The latter is a collection of 13 complete piano sonatas by W. A. Mozart (K.279, 280, 281, 282, 283, 284, 330, 331, 332, 333, 457, 475, and 533) performed on a Bösendorfer SE290 computer-controlled grand piano. G. Widmer, who prepared the data, gives a more detailed description [125]. As with the Magaloff data, all notes were matched to their counterparts in the score, resulting in a performance corpus consisting of roughly 106.000 played notes (about four hours of music). Melody notes were marked manually. As mentioned in section 4.2, our score and performance model, especially tempo and articulation, can only be calculated on homophonic music. Using only the matched melody notes reduces the number of usable data points from 307.900 to 100.236 in the Chopin corpus, and to 48.427 in the Mozart sonatas.

How the pieces are organized differs between the two datasets. In the Chopin data the smallest unit is a complete piece. This might be a complete opus (e.g. Scherzo Op. 31), part of an opus that contains several pieces (e.g. Nocturnes Op. 9 No. 2), or a movement of a sonata. In the Mozart data, most of the sonata movements are segmented into several parts, which are then treated separately and form the smallest unit in the dataset. The movements are split whenever a repetition, a change of key, or a change of meter is indicated in the score. This implies that the score part of passages that are repeated is stored twice in the dataset, once with each performance of the passage. For our approach, the effect of a repetition of the to-be-rendered test-piece being present in the training data is considerable, as can be observed in the respective quantitative evaluations. There are of course slight differences between the first and the second time the part is played, be they unintentional or on purpose, and one would expect the effect to be negligible due to the large number of pieces in the dataset. However, situations occur where score situations are unique under the chosen representation, and the only other occurrence being the repetition of the sequence. In absence of the repetition in the training data, the score situation has no performance information and the algorithm has to interpolate between known points preceding and following the gap. Thus, to evaluate a model on a set of test pieces in one fold of a cross validation run (see below), we have to exclude from the training data all pieces that are repetitions pieces in the test set.

The Mozart data were split into two different datasets – fast movements and slow movements – as they might reflect different interpretational concepts that would also be reproduced in the predictions. We also show the results for the Chopin data for different categories (ballades, nocturnes, etc.). The experiments consist of k-fold cross validations of the different datasets: fast Mozart movements (MOZ/F, $k = 10$), slow Mozart movements (MOZ/S, $k = 10$), the complete Chopin dataset (CHP, $k = 10$), and the separate categories ($k = 3$). In a k-fold crossvalidation each piece (i.e. a complete piece in the Chopin data, and a segment of a sonata movement in the Mozart data) is used once as a test piece, and $k - 1$ times as a training piece. The average performance quality over all test pieces in the k folds, which amounts to all pieces of a dataset, serves as a quality indicator for a model.

5.2 Problems of Automatic Evaluation

Every optimization process needs a measure of quality to compare different (partial) solutions of the problem and decide which is better. In the case of expressive performance rendering, the *musicality* of the rendered performance is the key factor. However, this is not only a highly subjective matter, but also a complex interaction of a plethora of factors, all of which are subject to the peculiarities of human perception and personal taste.

As a substitute to judging the absolute quality of a performance, assessing the similarity between two performances of the same piece, one played by a human, the other one computer generated, can give an estimate of how "human" and "natural" the generated performance might sound. This seems easier to quantify. Few examples of automatic evaluation of performance rendering systems can be found in the literature. Teramura [117] calculates the pointwise differences between prediction and target, and uses the standard deviation of the differences – normalized with respect to the standard deviation of the original – as a measure of quality (*normalized difference*). Widmer [130] and Grindlay [50] both employ correlation coefficients as a quality indicator, which is what we also use in the following experiments.

The correlation $r_{p,t}$ between the original target curve $t = (t_1, \cdots, t_n)$ and the predicted curve $p = (p_1, \cdots, p_n)$ is calculated as follows:

$$r_{p,t} = \frac{\Sigma_{p,t}}{\sigma_p \cdot \sigma_t} \tag{5.1}$$

$$= \frac{\sum_{i=1}^{n}(p_i - \overline{p})(t_i - \overline{t})}{\sqrt{\sum_{i=1}^{n}(p_i - \overline{p})^2}\sqrt{\sum_{i=1}^{n}(t_i - \overline{t})^2}} \tag{5.2}$$

Both curves are treated as random variables, with $\Sigma_{p,t}$ being the covariance of the two variables, σ_p and σ_t their respective standard deviations, and $\overline{t}$ and $\overline{p}$ the means of the curves. Values range from −1 to 1, with 1 being perfectly (linearly) correlated, −1 being in perfect negative (linear) correlation, 0 indicating no (linear) correlation.

Three major problems arise with the use of correlation as a similarity measure for music performances: (1) Correlation constitutes a point-wise comparison of the two curves by assessing the differences in deviation from the mean of corresponding elements. By design this is order invariant, and hence ignores the natural context dependency of time-series. However, perception of music is sensitive to context: the relative loudness of a note with respect to its immediate context is much more important than the relative

loudness with respect to the absolute mean of the complete piece. (2) Some points in a performance bear greater significance than others in transporting expressivity. For the perception of a phrase, for example, it is more important where a concluding ritardando reaches its slowest point than where it starts. In the correlation coefficient, each point carries the same importance. (3) Determining whether a prediction technique is suitable for a specific prediction scenario involves assessing certain qualitative aspects of the produced curves. The prediction of the local tempo of a piece, for example, is expected to be a smooth, slowly evolving curve as opposed to the fast fluctuations expected in the predictions for articulation, which does not carry long-term dependencies. Correlation does not provide the means for this kind of qualitative assessment.

Calculating the similarity between the prediction of a model and the desired outcome is a very common problem, encountered in such diverse flavours as predicting the stock market or the streamflow in water catchments [66]. However, the requirements are application specific: Given a suitable representation of the modeled system and its outcome, similar outcomes lead to representations that share certain characteristics. What those characteristics are depends on the chosen representation and the modeled system. Thoughts on how a music-specific similarity measure could be implemented can be found in 6.2.2.

For the time being we use the following compromise: Curves produced by different instances of the same concept (same algorithm, different score models), usually exhibit the same overall qualities. Correlation is used to decide which of the instances is more suitable. However, the decision which approach to use for specific scenarios is based mainly on a separate, manual assessment of the desired general qualities of the produced curves. Ultimately, of course, only listener perception can verify if those qualities lead to musically sensible performances.

5.3 Score model evaluation

5.3.1 Separation Qualities of Different Score Features

The main characteristic for discrete score features is how differently target values are distributed across the values of the feature. For the simple YQX prediction, the means of the target distributions for the individual values of a feature can be seen as building blocks with which the systems tries to reconstruct the performances. If those building blocks are very similar, the capacities for learning and reproducing highly variant perfor-

mances are small. If the means are different from each other, the base for reconstruction is more expressive; the feature discriminates situations that require different realizations. We refer to that characteristic as *separation*.

Generally, across all features, the Chopin data are much harder to separate than the Mozart sonatas. The differences in size (the Chopin dataset is three times the size of the fast Mozart movements and eight times the size of the slow movements) play a major role in this, as the means have to account for more instances and hence even out more easily. Also, in the Chopin data the features have to account for much more extensive expressive variations.

Figure 5.1 displays how the discrete feature *Rhythm Context* separates the target IOI Ratio in the different datasets. Shared across all three datasets, and the most pronounced of all relations between rhythmic context and tempo, is the tendency to delay a note that is preceded by a shorter note and a rest (*-nl*). This seems to hold true also for other situations where the last note of a triplet is longer than the other two: *ssl*, *nsl* and *snl*. In the fast Mozart movements the pattern *snl* is more pronounced than *ssl*, and vice versa in the slow movements. The Chopin data only share this trend for the *-nl* pattern. However, a tendency to speed up the middle note of a triplet of equal durations *nnn* can be discerned, which is unremarkable in the Mozart data. The tendency to delay the third note in short-short-long patterns was also discovered in rule extraction experiments by Widmer [127].

For articulation an effect of pitch interval to the next note can be seen. In both the fast Mozart movements and the Chopin data there is a strong trend to join notes together if they are close in pitch, but to separate notes on the same pitch, and notes with a large interval. This is less pronounced in the slow Mozart movements, where generally the melody is played more legato.

5.3.2 Feature Selection

Playing Mozart requires fundamentally different interpretational approaches than playing Chopin. Fast and slow movements of Mozart sonatas also follow (slightly) different rules. The (supposedly) systematic differences in performance result in largely different feature/target distributions. Establishing a set of features for each target and each of the three datasets, Chopin, fast and slow Mozart movements, seems a reasonable middle ground between searching a general (and unspecific) representation and overfitting to the data.

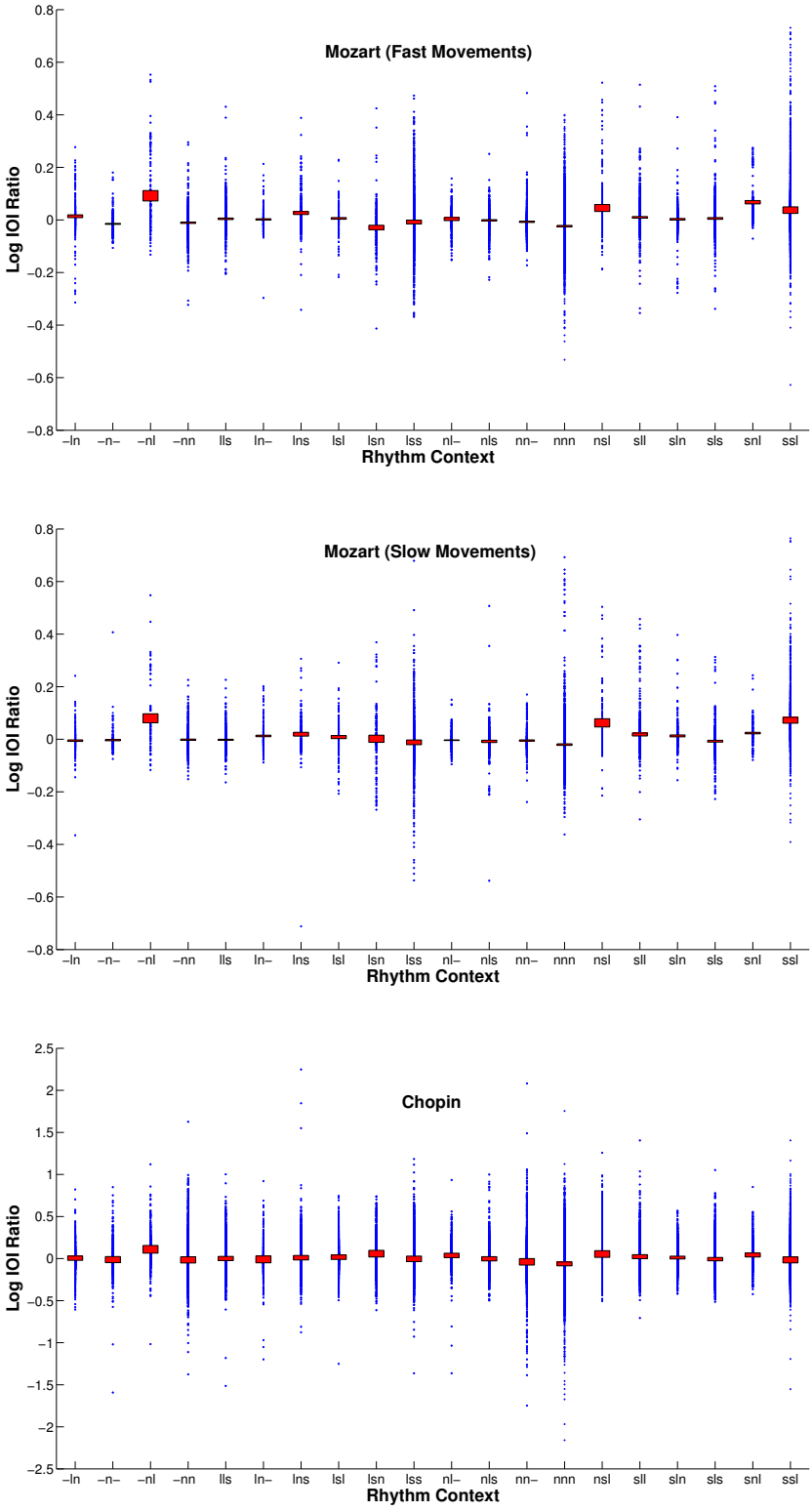


Figure 5.1: Separation of ioiRatio in the three different datasets by the feature Rhythm Context. Red Boxes indicate mean and variance of the target distribution for individual values of the feature. Top to bottom: Mozart Fast, Mozart Slow, Chopin complete.

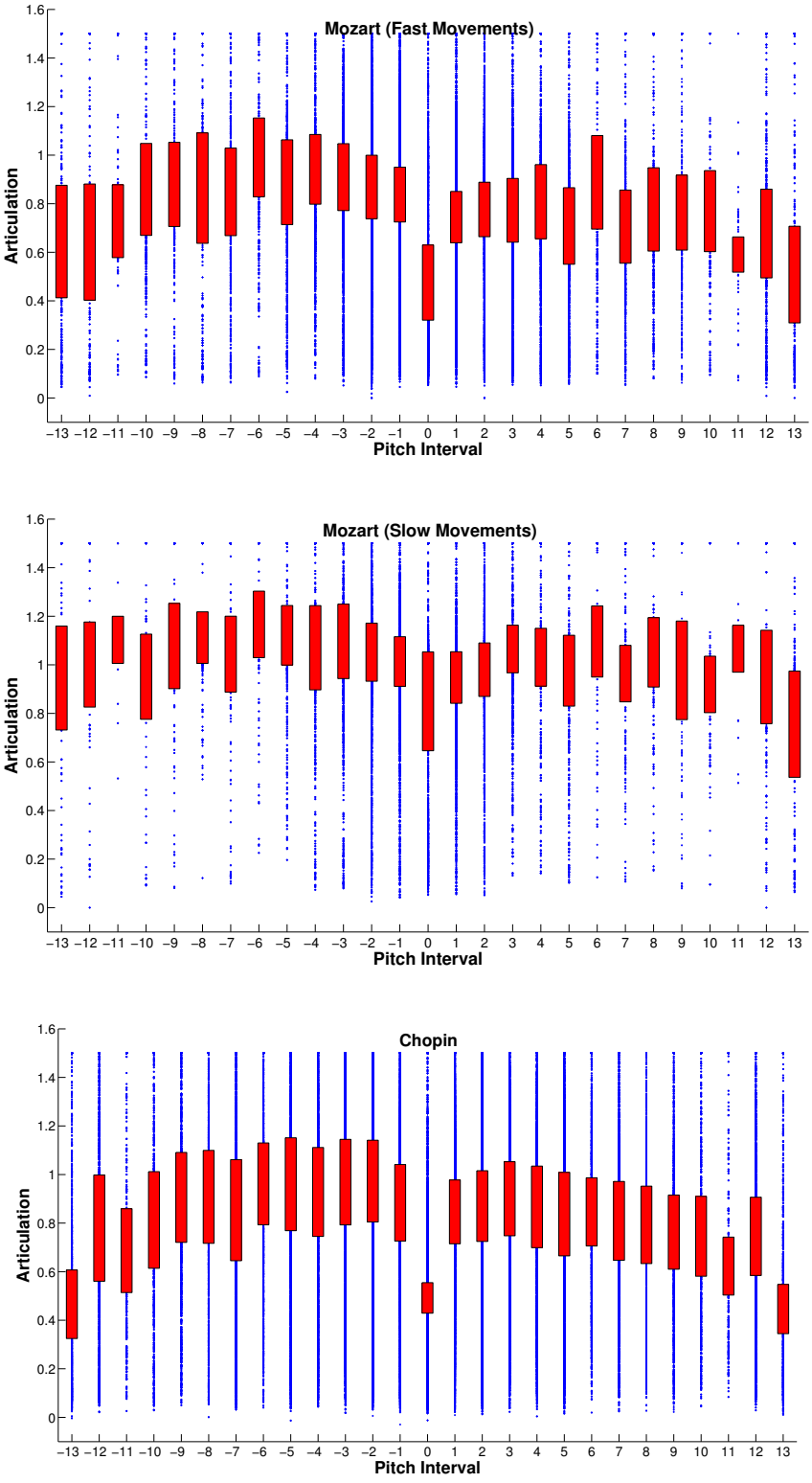


Figure 5.2: Separation of articulation in the three different datasets by the feature Pitch Interval. Red Boxes indicate mean and variance of the target distribution for individual values of the feature. Top to bottom: Mozart Fast, Mozart Slow, Chopin complete.

Searching for the overall best set of features for the respective datasets and targets is computationally very expensive, due to the exponential growth of the number of possible combinations of features. Instead, we conducted a *greedy hill-climbing* search in the space of possible feature combinations: starting with an empty feature set, in each iteration the one feature is added that increases the average correlation of a ten fold cross validation of the dataset most. A decrease in correlation by 0.02 is allowed once per search to avoid one local minimum and favor more features over a slightly higher correlation. Table D.5 shows the results for the different datasets and targets, based on the following set of features: IR Arch (IR-A), IR Label (IR-L), Pitch Interval (PI), Grouped Pitch Interval (PI-G), Consonance Difference (CD), Local Consonance (LC), Melodic Max Peaks (MaxP), Melodic Min Peaks (MinP), Average Max Peaks (AvMax), Average Min Peaks (AvMin), Metrical Strength (MS), Rhythmic Context (RC), and Duration Ratio (DR). The experiment was done separately for all three models (YQX, YQX/L, and YQX/G) described in chapter 4.

In the experiments that follow, I will use the features sets determined for YQX also for YQX/L, although feature selection indicates differently. The algorithms are conceptually similar and this way a direct comparison is more meaningful. For the Mozart data it seems justified, for the sake of simplicity and comparability, to use the complete set of features for all targets and both fast and slow movements (although we will continue to train separate models).

In situations where the new data is limited, as it is the case in the RENCON, it can be beneficial to tailor the set of features and the trained model specifically to the piece(s) in question. Given a good intuition about the desired stylistic elements and profound knowledge of the training datasets, one can select the set of features that performed best on a subset of the training data that is stylistically similar to the test piece.

5.4 Articulation Prediction

Articulation describes how closely two successive notes are joined together. As defined in section 4.3.1 we measure articulation by relating the size of the gap between two notes in the performance to the gap notated in the score. Values theoretically start at 0 and can take any positive real number. However, in practice we limit the values to a range from 0.15 to 1.5, as anything below 0.15 sounds unnaturally short, and above 1.5 just blurs the acoustic image.

5.4.1 Quantitative Evaluation

For the experiments we use the following sets of features, as determined by the feature selection algorithm in section 5.3.2. *YQX+/L* stands for both simple YQX prediction (see section 4.4) and the locally optimized version described in 4.5.1, *YQX/G* refers to the globally optimized system described in 4.5.2.

Chopin (YQX+/L) IR-Arch, Pitch Interval, Grouped Pitch Interval, Rhythm Context, and Duration Ratio.

Chopin (YQX/G) Pitch Interval, Grouped Pitch Interval, Consonance Difference, Metrical Strength

Mozart (YQX+/L) IR-Arch, IR-Label, Pitch Interval, Grouped Pitch Interval, Consonance Difference, Local Consonance, Average Max Peaks, Average Min Peaks, Metrical Strength, Rhythm Context, and Duration Ratio

Mozart (YQX/G) Pitch Interval, Grouped Pitch Interval, Metrical Strength, and Rhythm Context.

Table 5.1 shows the correlation achieved on average for the datasets. Values in brackets show the results on the Mozart data with repetitions not excluded from the training set (see section 5.1). All algorithm work best on the fast Mozart movements. MOZ/S seems to follow a different articulation regime than MOZ/F that is more difficult to model. Across the different categories of Chopin pieces results differ considerably. The Études and Scherzos score lowest. In the case of the Etudes this could be due to the fact that, while within a single Étude there is usually very little stylistic variation (with the exceptions of Op. 25 No. 5 and No. 10), the differences between Études can be substantial (for example Op. 10 No. 4, with its fast and sharply articulated runs in the right hand and Op. 25 No. 1, where the right hand is to be played highly legato and only a selection of notes needs to be heard very clearly and above all others).

Prediction quality decreases with the introduction of performance context (YQX/L and YQX/G). This is not surprising, as articulation is a local phenomenon that does not benefit from long-term modeling. It is noteworthy that the average correlation is higher for the complete Chopin dataset than for all subsets of the data. While this is clearly because we chose the feature set optimizing efficiency on the complete Chopin dataset, the effect is still surprising given the sheer differences in size and the expected diminished

	Ballades	Études	Mazurkas	Nocturnes	Pieces
YQX	0.29	0.17	0.20	0.21	0.26
YQX/L	0.27	0.18	0.19	0.22	0.26
YQX/G	0.29	0.09	0.20	0.19	0.19
	Polonaises	Préludes	Scherzos	Sonatas	Waltzes
YQX	0.30	0.16	0.12	0.23	0.29
YQX/L	0.29	0.19	0.11	0.22	0.26
YQX/G	0.19	0.14	0.00	0.14	0.18
	Chopin	Mozart fast	Mozart slow		
YQX	0.31	0.41 (0.61)	0.22 (0.54)		
YQX/L	0.30	0.41 (0.61)	0.26 (0.56)		
YQX/G	0.32	0.31 (0.36)	0.17 (0.21)		

Table 5.1: *Articulation* prediction: Average correlations achieved by the different prediction methods over different datasets. Values in parentheses are the results on the Mozart data with repetitions not excluded from the training set. The generic category *Pieces* comprises: Rondos (Opp. 1, 5 & 16), Variations Op. 12, Bolero Op. 19, Impromptus (Opp. 36 & 51), Tarantelle Op. 43, Allegro de Concert Op. 46, Fantaisie Op. 49, Berceuse Op. 57, and Barcarolle Op. 61.

explanatory power of the features. The set of features seems to be more suitable to represent the overall Chopin perspective than the more specific characteristics of single categories.

5.4.2 Qualitative Evaluation

Figure 5.3 shows a prototypical constellation for the prediction of articulation. The plot shows the articulation values extracted from Magaloff's performance of the Chopin Marzuka Op. 6 No. 2, and three predictions made by YQX, YQX/L, and YQX/G. Magaloff's curve presents with fast fluctuations, spanning a range from around 0.05 to the cut-off value 1.5. Of the three predictions, both simple inference and local optimization seem to reproduce the general shape to a certain degree. At 0.38, local optimization ranks slightly higher in terms of correlation than simple inference at 0.36. Global optimization (correlation of 0.31 to the original) dampens the curve suppressing fast fluctua-

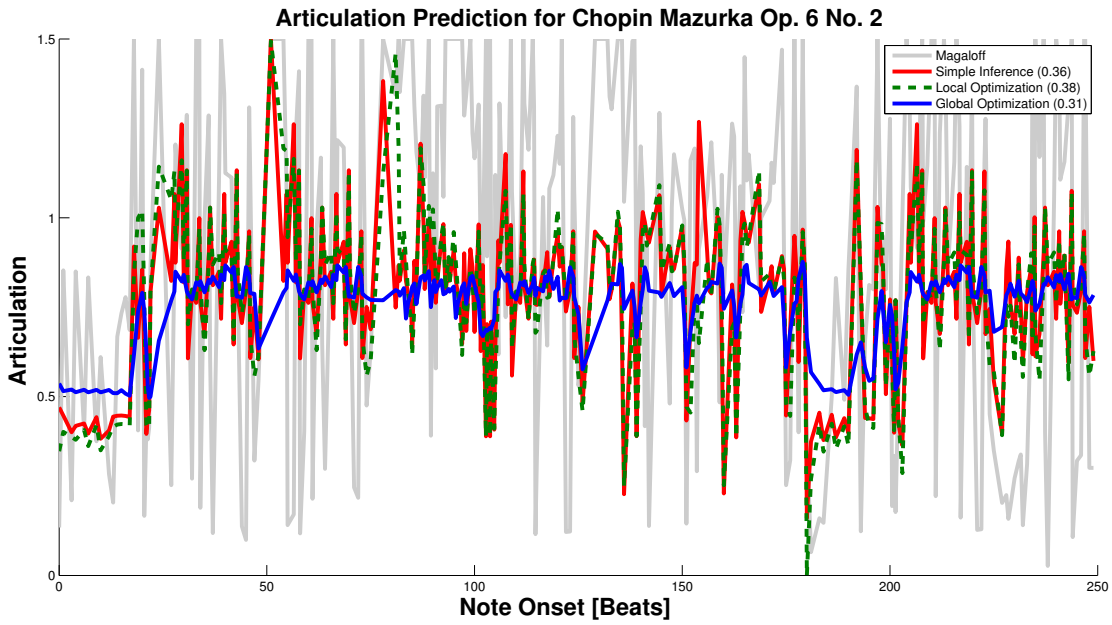


Figure 5.3: Articulation curve predicted for Chopin Mazurka Op. 6 No. 2 with Simple Inference, Local Optimization, and Global Optimization. Also shown is the original articulation curve measured from Magaloff's performance.

tions and introduces a counter-intuitive context-sensitivity. Therefore, regardless of the fact that global optimization numerically increases correlation in some cases, qualitative considerations argue against applying that approach to the prediction of articulation.

5.5 Loudness Prediction

Two definitions of loudness are suggested in chapter 4: (1) loudness as measured from a performance, which we call *complete loudness curve*, introduced in 4.3.2 and (2) *local loudness*, the residual after subtracting the loudness curve rendered based on the expressive annotations in the score, introduced in 4.6. I evaluate the two definitions separately.

5.5.1 Complete Loudness Curve

Loudness is modeled not in absolute MIDI loudness values, but as a ratio of current loudness and average loudness of the complete piece. Hence, the predictions also represent relative loudness values, and are independent of the general loudness chosen for a piece. As defined in 4.3.2 we use the logarithm of the loudness ratio, which results in values usually ranging between −1 and +1. In practice, we restrict the final MIDI loudness values to lie between 15 and 105.

Quantitative Evaluation

Experiments were conducted using the following sets of features, as determined by the feature selection algorithm in section 5.3.2. *YQX+/L* stands for both simple YQX prediction (see section 4.4) and the locally optimized version described in 4.5.1, *YQX/G* refers to the globally optimized system described in 4.5.2.

Chopin (YQX/+L) Average Max Peaks, Melodic Min Peaks, and Duration Ratio.

Chopin (YQX/G) Average Max Peaks, Melodic Max Peaks, Melodic Min Peaks, and Metrical Strength.

Mozart (YQX/+L) IR-Arch, IR-Label, Pitch Interval, Grouped Pitch Interval, Consonance Difference, Local Consonance, Average Max Peaks, Average Min Peaks, Melodic Max Peaks, Melodic Min Peaks, Metrical Strength, Rhythm Context, and Duration Ratio.

Mozart (YQX/G) IR-Label, Pitch Interval, Grouped Pitch Interval, Average Max Peaks, Melodic Min Peaks, Metrical Strength, and Rhythm Context.

Table 5.2 shows the correlation achieved on average for the datasets. Values in brackets show the results on the Mozart data with repetitions not excluded from the training set. The prediction quality for the Mozart sonatas is on par with the articulation predictions, and, as before, the fast movements are considerably easier to model than the slow movements. For both, the prediction quality is noticeably higher than for the Chopin data. The difference between the complete Chopin dataset and the individual categories is even more prominent for loudness than for articulation. For the Mozart sonatas and the majority of Chopin datasets the average correlation decreases with the introduction of performance context, although not as pronouncedly as with the articulation predictions.

	Ballades	Etudes	Mazurkas	Nocturnes	Pieces
YQX	0.08	0.14	0.11	0.11	0.09
YQX/L	0.11	0.13	0.09	0.09	0.07
YQX/G	0.10	0.14	0.10	0.12	0.13
	Polonaises	Preludes	Scherzos	Sonatas	Waltzes
YQX	0.03	0.07	0.07	0.07	0.08
YQX/L	0.02	0.07	0.04	0.05	0.03
YQX/G	0.05	0.02	0.11	0.11	0.05
	Chopin	Mozart fast	Mozart slow		
YQX	0.17	0.39 (0.66)	0.27 (0.65)		
YQX/L	0.15	0.37 (0.64)	0.23 (0.64)		
YQX/G	0.17	0.31 (0.47)	0.24 (0.39)		

Table 5.2: *Complete Loudness* prediction: Average correlations achieved by the different prediction methods over different datasets. Values in parentheses are the results on the Mozart data with repetitions not excluded from the training set.

5.5.2 Local Loudness

As discussed in sections 4.1.4 and 4.6.1, expressive annotations in the score already establish a coarse picture of the loudness evolution throughout the piece. By eliminating this (presumably already explained) part from the loudness curves, the training focuses the model on the part that is not accounted for by score annotations. Loudness can therefore be understood as a composite of two elements: (1) *annotated loudness*, the part prescribed by score annotations, and (2) *local loudness*, the residual local variations. As a prerequisite we need to construct a loudness curve from the annotations in the score, in a way that fits the realizations of the annotations by the pianist. The following experiments were done based on loudness curves generated by Grachten's Basis Mixer [48, 49].

The Basis Mixer assigns functions to all loudness annotations in the score, which are active at the same time as the annotations. The weighted sum of all functions forms the loudness curve accounted for by the annotations. The weights associated with individual annotations can be learned from the Magaloff data. In a scenario where one tries to predict a loudness curve for a new piece, the weights learned for a specific type

of annotation would be averaged and applied to render all instances of this annotation. Here however, as we want to eliminate as much "annotated loudness" as possible and examine if our approach works better on the residual, we need to fit the annotated loudness curve as closely as possible to Magaloff's real performance. Therefore, the weights for each individual annotation are chosen optimally, which makes the resulting curves fittings rather than actual predictions. As described in section 4.6.1, the local loudness is than calculated by subtracting the annotated loudness from the Magaloff's loudness curve and viewing the remaining values relative to the annotated loudness.

Expressive Annotations are relatively sparse in Mozart's Urtext compared to Chopin's extensively annotated scores. The effect of disentangling the variations assumed to be caused by expressive annotations from the measured loudness is therefore marginal in the Mozart sonatas. Moreover, the Mozart corpus not being available as musicXML but only as MIDI files makes it much more difficult and laborious to add the annotations to the corpus. Hence, the following experiments only include the Chopin Data. Except for the Nocturnes, where every expressive direction given in the printed score was consistently transferred to the musicXML files, mainly the *crescendo* and *decrescendo* wedges have be transcribed from the printed score. Consistently including the remaining annotations, which I intend to do for future versions of the corpus (see section sec:CON-Future-CORP), should further improve the situation.

Quantitative Evaluation

The following sets of features were used, as determined by the feature selection algorithm in section 5.3.2:

Chopin (YQX/+L) IR-Arch, Melodic Min Peaks, Local Consonance, Metrical Strength, and Rhythm Context

Chopin (YQX/G) Average Max Peaks, Average Min Peaks, Metrical Strength, and Rhythm Context

Table 5.3 shows the correlations achieved on average over three-fold crossvalidations of the datasets. The results are similarly low as for the complete loudness predictions. For the individual categories, no clear advantage of one algorithm over the others can be detected. Averaged over the complete Chopin, the globally optimized model has a clear lead.

	Ballades	Etudes	Mazurkas	Nocturnes	Pieces
YQX	0.13	0.06	0.04	0.06	0.06
YQX/L	0.16	0.05	0.04	0.07	0.12
YQX/G	0.06	0.08	0.08	0.05	0.05
	Polonaises	Preludes	Scherzos	Sonatas	Waltzes
YQX	0.05	0.10	0.15	0.07	0.05
YQX/L	0.08	0.08	0.17	0.07	0.07
YQX/G	0.06	0.06	0.17	0.10	0.00
	Chopin				
YQX	0.12				
YQX/L	0.10				
YQX/G	0.18				

Table 5.3: *Local Loudness* prediction: Average correlations achieved by the different prediction methods over different datasets.

5.5.3 Qualitative Evaluation

Complete Loudness Curve

Figure 5.4 shows the loudness predictions of YQX and YQX/G together with the original loudness curve extracted from R. Batiks performance of an excerpt of Mozarts Sontata KV 280 in F Major (3^{rd} Movement, *Presto*, Bars 1 – 77). In the upper panel the training set contained the repetition of the passage, in the lower panel, the repetition was excluded. The locally optimized prediction is virtually identical with the simple inference and is not shown in the plots.

Both predictions in the upper panel score an exceptionally high correlation (0.73 and 0.63 for YQX and YQX/G, respectively), copy almost all major trends of the original, agreeing in many peaks and fluctuations. The global optimization seems more conservative than the simple inference, smoothing over some of the faster changes and generally reacting more slowly and less lively. The original curve in this case justifies many of the peaks and fluctuations in the YQX curve. In practice however, a more conservative rendering might be more advisable, as an undue peak at a prominent position can very easily disrupt the rendering.

Correlations for the predictions in the lower panel (trained without the repetition)

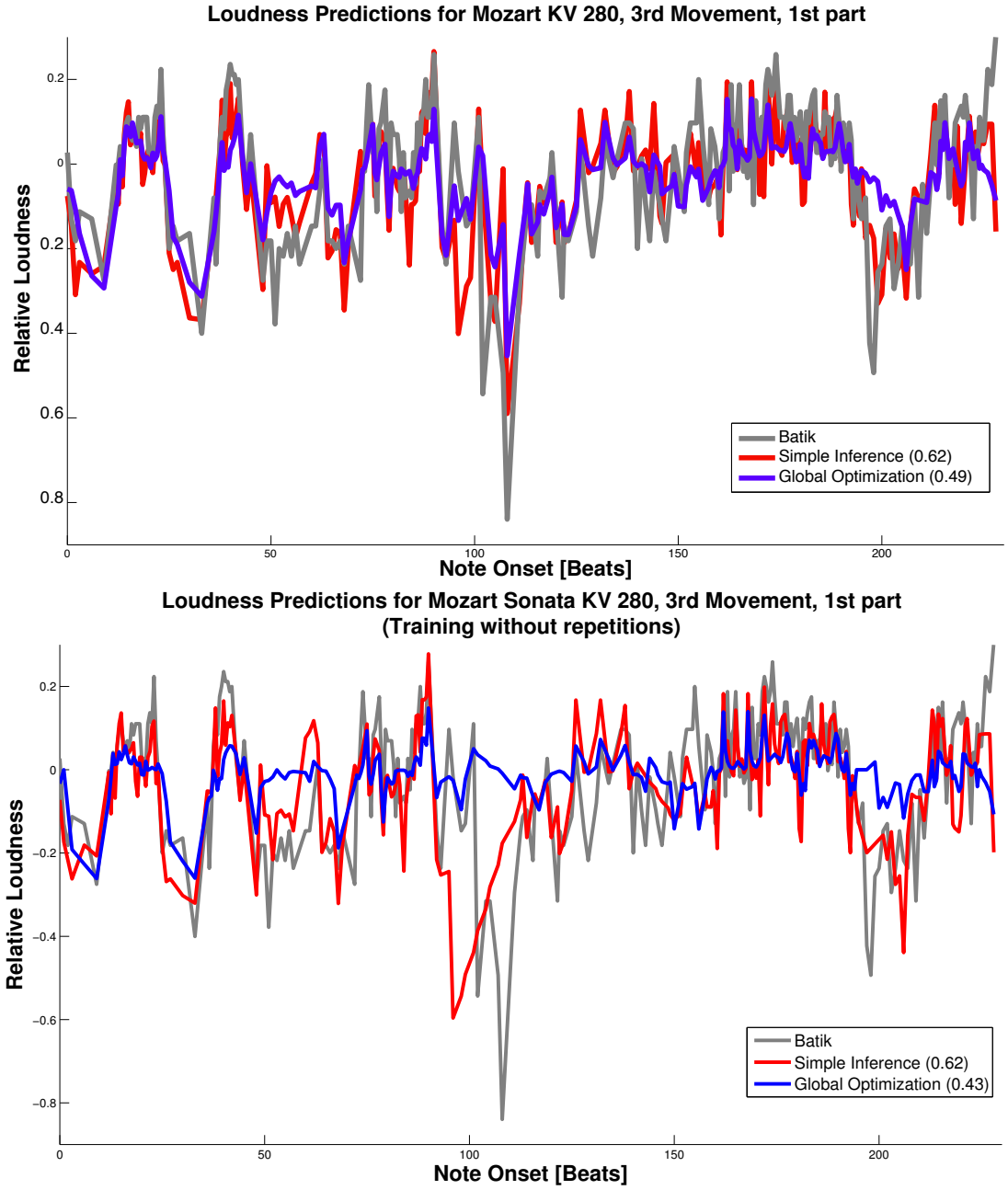


Figure 5.4: Loudness curve predicted for Mozart Sonata KV 280, 3rd Movement, Bars 1-77 with Simple Inference, and Global Optimization. Also shown is the original loudness curve measured from Batik's performance of the piece. *Upper panel:* Training data did include Batik's performance of the repetition of the passage. *Lower panel*: Batik's performance of the repetition of the passage was not included in the training data.

are still very high (0.62 and 0.49 respectively). YQX still follows the major trends and most of the local peaks and fluctuations. Some situations, however, are clearly worse than in the "informed" prediction. Most prominently around beat 100 where YQX misses the overall softest point in the piece, and, probably even worse, places its softest point where Batik is at a loudness peak. YQX/G skips over several of the larger trends (also around beat 100 and again near beat 200), but, while being less expressive, the prediction is less prone to mishaps.

Loudness predictions for the Chopin data paint a less favorable picture. Figure 5.5 shows the loudness predictions for the Chopin Etude Op. 25 No. 2 in F minor. The predictions in the upper panel were made by a model trained on the complete Chopin dataset, while in the lower panel, a model trained only on the Etudes was used. The most obvious difference between the two are the large, negative peaks in the latter. The situations occur at beats 36, 58, 93, and 95, completely unwarranted by the musical content of the score. The first two occur with the first notes in bars 19 and 30 (Pitches $C5$ and $E\flat 5$ respectively) and in the middle of (identical) bars 46 and 47 (Pitch $B\flat 4$). Where the peaks come from is hard to trace, but in all likelihood, only very few instances populate the specific score situation, and by chance represent very soft notes. The peaks are missing in the model trained on the complete Chopin dataset, hence, the score situation is now represented by more instances and the loudness evened out to a less extreme value. This, of course, also affects all other score situations, and generally leads to interpretations with a less expressive variation. The much less pronounced *crescendo-descrescendo* combination near beat 20 serves as an example. As before, the difference between simple inference and local optimization was marginal, which is why the curve is not displayed here. In both cases (upper and lower panel), the curves predicted by YQX and YQX/G leave a lot to be desired. The globally optimized prediction exhibits some coherent trends, consistent with *crescendo-decrescendo* combinations, some of which have corresponding movements in the original. Some of the trends are also present in the YQX curve, but are blurred by the fast fluctuations, which will also impede perception of the trends in the corresponding audio.

Generally, the amplitude of the predictions is significantly smaller. In order to create a "more expressive" performance, it is possible to scale the curve, such that it either covers a specified range or that its absolute maximum has a specified value. This, however, should be done with extreme care, as too excessive variation can very easily sound unnatural.

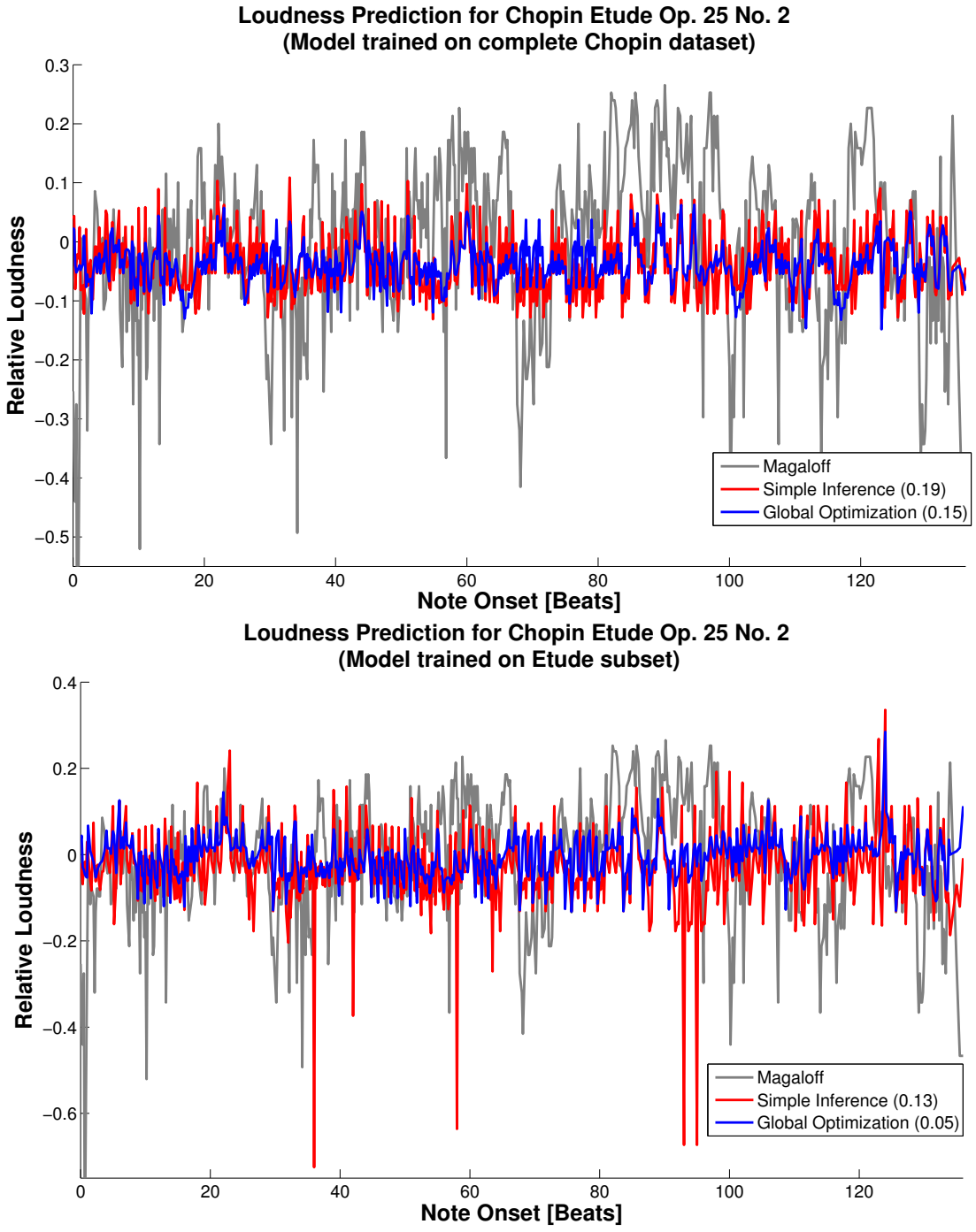


Figure 5.5: Loudness curve predicted for Chopin Etude Op. 25 No. 2 with Simple Inference, and Global Optimization. Also shown is the original loudness curve measured from Magaloff's performance of the piece. *Upper panel*: The models were trained on the complete Chopin dataset. *Lower panel:* The training data contained only the Etude subset of the data.

Local and annotated loudness

Figure 5.6 shows the effect of eliminating the annotated loudness from Magaloff's performed loudness curve (again for the Etude Op. 25 No. 2 in F minor). At places with no annotated dynamic changes, the two performance targets *complete loudness* and *local loudness* are congruent. The effect of removing annotated dynamic changes from the loudness curve can be seen, for instance, from onsets 110-128: the annotated loudness describes a *crescendo-decrescendo* combination which is also contained in the complete loudness curve. The local loudness remains more stationary and displays only a reduced version of this. Consequently, the prediction model does not have to account for the *crescendo-decrescendo* combination in full.

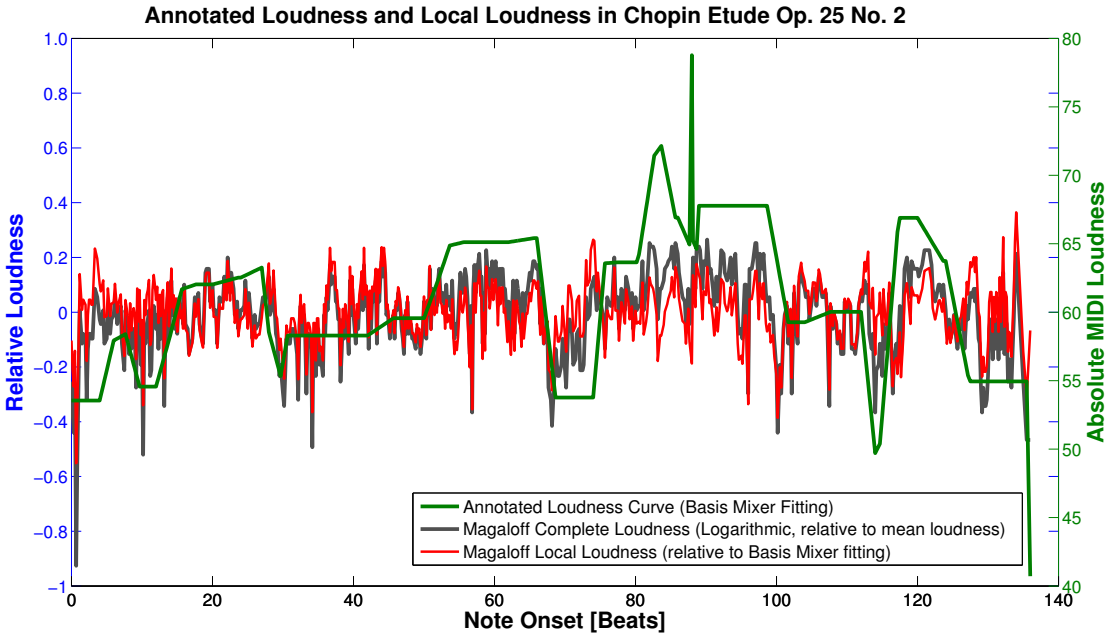


Figure 5.6: Local loudness curves for Chopin Etude Op. 25 No. 2, measured from Magaloff's performance of the piece. *Left Axis*: Complete loudness curve (grey, logarithmic and relative to the mean loudness of the piece) and local loudness (red, relative to the Basis Mixer fitting). *Right Axis*: the Basis Mixer fitting for the same piece (green).

Figure 5.7, finally, shows the predictions of the global model for the local loudness of the first 15 bars of the Etude. The prediction has been scaled to match the overall mean

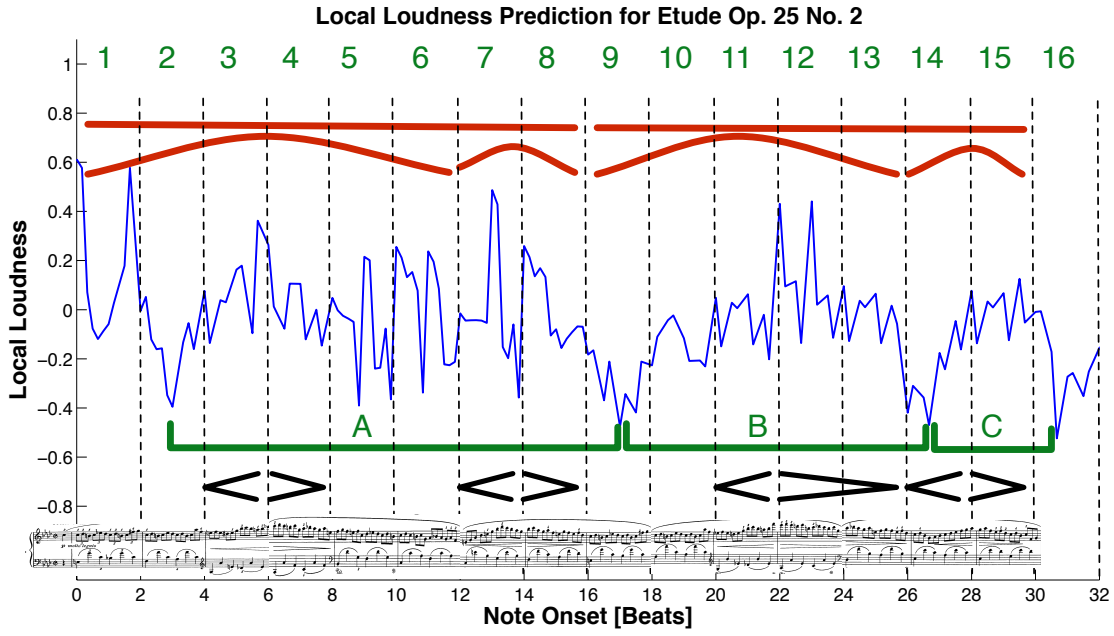


Figure 5.7: Local loudness predicted for the first 16 bars of Chopin Etude Op. 25 No. 2 with Global Optimization. Dotted vertical lines indicate bar lines (bar numbers in green). Red lines roughly represent phrases (straight) and subphrases (curved). Black wedges correspond to the crescendos and decrescendos annotated in the score.

(−0.0241) and variance (0.0536) seen over the complete corpus. As indicated by the red lines on top of the plot, this excerpt roughly corresponds to the first two major phrases of the piece, each of which consists of a longer subphrase (6 and 5 bars, respectively) followed by a shorter 2 bar phrase. Figure 5.8 shows the score of the excerpt together with the suggested phrasing. A common hypothesis (e.g. [120]) about expressive variation of loudness postulates that phrases are framed by local loudness minima. With respect to local minima, the YQX/G prediction of the local loudness curve can be considered to consist of three main parts: (1) Onsets 3 to 17 (part A) , (2) Onsets 17 to 26.6 (part B) , and (3) Onsets 27 to 31 (part C). The end of part A in the predicted curve coincides almost precisely with end of the first main phrase. The first subsegment in part A, an arc spanning onsets 3 to 9 (the first local minimum in A), surrounds the crescendo-decrescendo combination annotated in the score. The decrescendo wedge at the end of

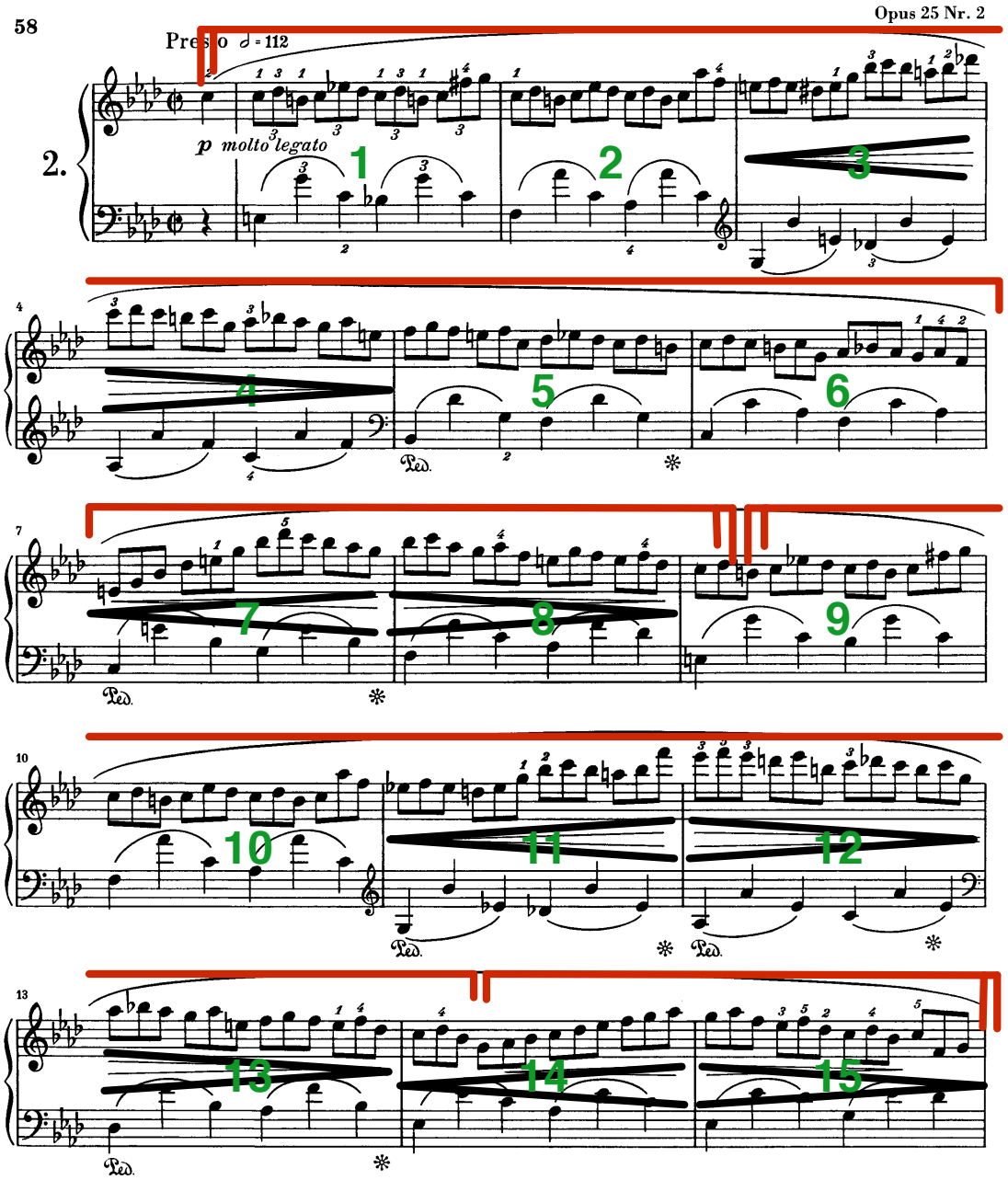


Figure 5.8: The first 15 bars of Chopin Etude Op. 25 No. 2 (Henle Urtext Edition [137]). Double vertical red lines indicate phrase boundaries, single vertical red lines indicate sub phrase boundaries. Black wedges correspond to the crescendos and decrescendos annotated in the score. Green brackets indicate a suggested partition of the predicted curve in three parts (A, B, and C)

part A (bar 8) is also reflected in the predicted curve. Parts B and C coincide with the two sub phrases of the second main phrase, both demarcating beginning and end of the respective phrases with local minima and describing an arc in-between. Although the overall correlation between prediction and Magaloff's original curve is not very high (0.17 for the complete piece), the curve predicted for the shown excerpt contains all expected major loudness trends.

The reason why this example works, is that here the intended loudness variations (as indicated e.g. by the phrase structure and the supporting dynamic annotations) are closely related to mainly one aspect in the score, namely pitch: the suggested peaks in loudness roughly correspond to the locally highest pitch (and the middle of the phrase), while local pitch minima often coincide with phrase boundaries. In the score model used for this prediction pitch is represented by the distance to the next turning point of the pitch sequence (*Average Max Peaks* and *Average Min Peaks*). Most of the time the links between score and performance are much more complex and not as obvious as here. However, sometimes a simple score model like the one used here is able to capture such connections. To some extent, this might be seen as proof of concept for the complete approach. Regarding the separation of local loudness and annotated loudness, what can be seen in this example is that both are indispensable parts of the whole complex. The annotated loudness can only account for the directives given the score. Hence, in 6 of the 15 bars of the example above the annotated loudness curve is completely flat. This does never happen in a real performance, which means that the annotated directives alone are insufficient. This is even more evident in pieces, where annotations are scarcer than in this example. The same holds true for the local loudness variation: in pieces, where the connection between score and performance is less obvious than here, or the interpretation intended by the composer (indicated by according dynamic annotations) contradicts that connection the predicted loudness curve will often lead to an unacceptable performance. Obeying the annotations in the score ensures that the performance stays in acceptable limits.

5.6 Tempo Prediction

Two conceptually different possibilities of defining and modeling performance tempo are proposed in section 4: (1) directly modeling the *complete tempo curve* of a performance (section 4.3.3); (2) viewing tempo as a composite phenomenon of several components

(section 4.6).

5.6.1 Complete Tempo Curve

As defined in 4.3.3, we call the series of logarithmic ratios of score and performance inter-onset-intervals (IOIs), the *complete tempo curve*. It contains all aspects of tempo note-by-note, relative to the beat tempo chosen for the complete piece. A value below 0 describes an IOI that was, with respect to the current tempo, played shorter than indicated in the score, which anticipates the succeeding note. A positive value describes an IOI that was lengthened, delaying the succeeding note, and therefore slowing down.

The predicted tempo curves usually have a much smaller amplitude than the curves measured in performances. In practice it sometimes can be beneficial to linearly scale the prediction to make the expressive variations more pronounced and audible. Care has to be exercised, because a misplaced, large tempo change can easily destroy all favorable impressions of the generated performance. This issue is much more prominent with the Chopin data, because the overall range of expressive variations is larger. The most obvious solution to this problem is to scale the predictions to match the training data in overall mean (−0.02) and variance (0.026).

Quantitative Evaluation

Automatic feature selection (section 5.3.2) suggests the following sets of features for the different algorithms and datasets:

Chopin (YQX/+L) IR-Arch, Consonance Difference, Metrical Strength, Rhythm Context, and Duration Ratio

Chopin (YQX/G) Pitch Interval, Grouped Pitch Interval, Metrical Strength, and Rhythm Context.

Mozart (YQX/+L) IR-Arch, IR-Label, Pitch Interval, Grouped Pitch Interval, Consonance Difference, Local Consonance, Average Max Peaks, Average Min Peaks, Metrical Strength, Rhythm Context, and Duration Ratio

Mozart (YQX/G) IR-Label, Pitch Interval, Grouped Pitch Interval, Average Max Peaks, Metrical Strength, and Rhythm Context.

	Ballades	Études	Mazurkas	Nocturnes	Pieces
YQX	0.23	0.14	0.20	0.13	0.14
YQX/L	0.20	0.13	0.18	0.11	0.13
YQX/G	0.31	0.09	0.17	0.12	0.21
	Polonaises	Préludes	Scherzos	Sonatas	Waltzes
YQX	0.18	0.16	0.27	0.11	0.30
YQX/L	0.18	0.13	0.24	0.08	0.32
YQX/G	0.15	0.07	0.18	0.11	0.33
	Chopin	Mozart fast	Mozart slow		
YQX	0.22	0.39 (0.60)	0.37 (0.67)		
YQX/L	0.20	0.39 (0.60)	0.34 (0.67)		
YQX/G	0.21	0.34 (0.41)	0.36 (0.47)		

Table 5.4: *Complete tempo curve* prediction: Average correlations achieved by the different prediction methods over different datasets. Values in parentheses are the results on the Mozart data with repetitions not excluded from the training set.

Table 5.4 shows the correlation achieved on average for the datasets. Values in brackets show the results on the Mozart data with repetitions not excluded from the training set. Overall, the prediction quality is higher than for the loudness curves, and, as before, the Mozart data lends itself more easily to modeling efforts than the Chopin pieces. Considering the fact, that the chosen feature set represents the complete Chopin data best, the Ballades, Mazurkas, Scherzos, and Waltzes seem to be the most characteristic of the categories for Chopin's particular style.

In same cases, most notably the Chopin Ballades, Pieces, and Waltzes performance context increases the numerical prediction quality.

5.6.2 Tempo as a composite phenomenon

Tempo can be considered to be an aggregation of several components: (1) the basic tempo of the piece, often indicated by an initial tempo marking, which we call *global tempo*, (2) a slowly evolving tempo trend, which we call *local tempo*, used for instance to carry ritardandi and accelerandi, and shape phrases, and (3) deviations of the individual notes from the local tempo, which we call *note timing*, shaping each note through antic-

ipation and delay, generating tension and relief. As defined in section 4.6.2, we extract aspects (2) and (3) from the complete tempo curve (which is already independent of global tempo) in two steps: First, we extract the low-frequency components from the curve by applying a moving average filter. This we associate with local tempo. Second, we calculate the residual of the complete tempo curve relative to the local tempo. This constitutes note timing.

Instead of training the model with complete tempo curves, and then interpreting the prediction in the same way, models can be trained on local tempo and note timing separately. The separation process can afterwards be reversed, and the two predictions reassembled to represent the complete tempo curve. This way, the (different) characteristics of the two components can be handled, and attended to, separately by different models.

By construction the extracted local tempo curves are smoother than the complete tempo curves, a quality we deem desirable for the slowly evolving tempo trend. Accordingly, we search for a model that reproduces that particular quality in its predictions. The principle also holds for note timing, which is much more specific to an individual note and less context-dependent. The curves, by construction, exhibit fast fluctuations, and sharp peaks, qualities we like to see retained in predictions of the target. All experiments are based on a window size of 4 beats for calculating the local tempo component, see section 6.2.2 for a further discussion.

Quantitative Evaluation

The feature sets established by the selection algorithm in section 5.3.2 for the Mozart data are the same for local tempo, note timing, and complete tempo (with the exception of the addition feature *local consonance* for the globally optimized version). For the Chopin data the suggested feature sets are the same for local tempo and note timing for the simple YQX and the locally optimized version (YQX/+L), but differ in one feature for the globally optimized version (melodic min peaks for tempo, grouped pitch interval for timing).

Local tempo prediction is based on the following feature sets:

Chopin (YQX/+L) IR-Arch, Consonance Difference, Metrical Strength, Rhythm Context, Duration Ratio

Chopin (YQX/G) IR-Label, Metrical Strength, Rhythm Context

Mozart (YQX/+L) IR-Arch, IR-Label, Pitch Interval, Grouped Pitch Interval, Local Consonance, Melodic Max Peaks, Melodic Min Peaks, Metrical Strength, Rhythm Context, Duration Ratio

Mozart (YQX/G) IR-Label, Grouped Pitch Interval, Melodic Min Peaks, Metrical Strength, Rhythm Context

	Ballades	Etudes	Mazurkas	Nocturnes	Pieces
YQX	0.25	0.13	0.06	0.05	0.13
YQX/L	0.13	0.16	0.11	0.07	0.07
YQX/G	0.34	0.17	0.17	0.07	0.23
	Polonaises	Preludes	Scherzos	Sonatas	Waltzes
YQX	0.11	0.05	0.15	0.12	0.31
YQX/L	0.10	0.12	0.03	0.08	0.33
YQX/G	0.12	0.08	0.36	0.20	0.44
	Chopin	Mozart fast	Mozart slow		
YQX	0.14	0.37 (0.69)	0.21 (0.61)		
YQX/L	0.15	0.38 (0.68)	0.21 (0.62)		
YQX/G	0.19	0.39 (0.48)	0.18 (0.29)		

Table 5.5: *Local tempo* prediction: Average correlations achieved by the different prediction methods over different datasets. Values in parentheses are the results on the Mozart data with repetitions not excluded from the training set.

The following features are suggested for the prediction of note timing:

Chopin (YQX/+L) IR-Arch, Consonance Difference, Metrical Strength, Rhythm Context, Duration Ratio

Chopin (YQX/G) Grouped Pitch Interval, Metrical Strength, Rhythm Context

Mozart (YQX/+L) IR-Arch, IR-Label, Pitch Interval, Grouped Pitch Interval, Consonance Difference, Local Consonance, Average Max Peaks, Average Min Peaks, Metrical Strength, Rhythm Context, Duration Ratio

Mozart (YQX/G) Pitch Interval, Grouped Pitch Interval, Average Max Peaks, Metrical Strength, Rhythm Context

Tables 5.5 and 5.6 show the correlation achieved on average for the datasets. Values in brackets refer to the results on the Mozart data with repetitions not excluded from the training set. Local tempo prediction of the Chopin data, individual categories as well as the complete dataset, benefits greatly from using the performance context: in all instances the average correlation is higher for the globally optimized model (YQX/G) than for the context-free version (YQX). Gains range from 0.02 (Nocturnes and Mazurkas) to 0.21 (Scherzos) in correlation.

Note timing presents with considerably lower correlations, suggesting that the target is much harder to learn, and generalize. Contrary to expectation, performance context and global optimization seems to have a positive effect in some cases. A noteworthy detail is that for the slow Mozart movements timing prediction is much more successful than local tempo, as opposed to the fast Mozart movements where the local tempo seems to be easier to generalize and learn than note timing.

	Ballades	Etudes	Mazurkas	Nocturnes	Pieces
YQX	0.07	0.05	0.18	0.08	0.03
YQX/L	0.06	0.05	0.19	0.08	0.05
YQX/G	0.05	0.05	0.16	0.08	0.09
	Polonaises	Preludes	Scherzos	Sonatas	Waltzes
YQX	0.07	0.09	0.06	0.02	0.14
YQX/L	0.06	0.09	0.06	0.02	0.14
YQX/G	0.11	0.05	0.10	0.07	0.16
	Chopin	Mozart fast	Mozart slow		
YQX	0.13	0.36 (0.58)	0.37 (0.66)		
YQX/L	0.13	0.35 (0.57)	0.33 (0.65)		
YQX/G	0.14	0.30 (0.37)	0.38 (0.45)		

Table 5.6: *Note timing* prediction: Average correlations achieved by the different prediction methods over different datasets. Values in parentheses are the results on the Mozart data with repetitions not excluded from the training set.

As described above, the predicted local tempo and note timing curves can be reassembled to form a complete tempo curve. The correlations presented in table 5.7 are calculated between the reassembled curves and the original complete tempo curves of the pieces in question. The results vary with the prediction paradigms used for the two com-

ponents. For the Mozart data, combined tempo prediction, while numerically slightly inferior to prediction of the complete tempo curve, still presents with a very high average correlation for both slow and fast movements. In both cases, the correlations only take a severe plunge when the globally optimized prediction is used for the note timing component. From a numerical point of view, the other combinations are equivalent. The correlations for combined and complete tempo prediction do not feature big differences in case of the Chopin data. With the exception of the Ballades and the Waltzes, only minor differences can be noticed. No clear tendency can be stated suggesting that one combination of models is superior for all subsets of the data. A numerical comparison with the results of the complete tempo curve prediction (table 5.4) does not show a significant improvement in terms of average correlation. With the exception of the Preludes and the Scherzos, the highest average correlation achieved by combining local tempo and note timing is equal or lower than the highest average correlation with a single model approach. This might suggest that the two methods, complete tempo prediction and combined tempo prediction, are equivalent and hence recommend the former as it is the simpler of the two. This, however, ignores all qualitative considerations of the different prediction strategies, which will be examined in the next section.

5.6.3 Qualitative Evaluation

A comparison between the average correlation coefficients for the *complete tempo* predictions (5.6.1) and the *combined tempo* (5.6.2) predictions does not reveal any advantage of one method over the other. However, the curves predicted by the different algorithms in the different prediction scenarios feature certain qualities, that make them more appropriate in some contexts than others. Starting with the complete tempo curve prediction, the examples in the following section examine the different scenarios and show how a combined prediction can be superior.

Complete Tempo Curve

Figure 5.9 shows complete tempo curve predictions for the Mozart Sonata KV280 in F Major, Bars 1-40, generated with the simple YQX model and the YQX/G. Both algorithms score reasonably high correlations for this example. The paradigmatic differences of the two algorithms are very obvious in the two displayed predictions: the context-free YQX presents with fast fluctuations, which, in this case, very often mirror

Timing	Tempo	Ballades	Etudes	Mazurkas	Nocturnes	Pieces
YQX	YQX	0.22	0.10	0.19	0.11	0.12
YQX	YQX/L	0.18	0.09	0.19	0.10	0.08
YQX	YQX/G	0.22	0.10	0.19	0.11	0.11
YQX/L	YQX/L	0.15	0.11	0.19	0.11	0.08
YQX/L	YQX/G	0.18	0.11	0.20	0.11	0.11
YQX/G	YQX/G	0.25	0.10	0.19	0.13	0.19
		Polonaises	Preludes	Scherzos	Sonatas	Waltzes
YQX	YQX	0.13	0.14	0.27	0.07	0.24
YQX	YQX/L	0.12	0.12	0.21	0.07	0.23
YQX	YQX/G	0.12	0.11	0.28	0.07	0.24
YQX/L	YQX/L	0.11	0.12	0.22	0.07	0.22
YQX/L	YQX/G	0.11	0.11	0.29	0.07	0.23
YQX/G	YQX/G	0.13	0.06	0.25	0.12	0.31
		Chopin	Mozart fast	Mozart slow		
YQX	YQX	0.20	0.37 (0.58)	0.36 (0.65)		
YQX	YQX/L	0.19	0.37 (0.58)	0.36 (0.65)		
YQX	YQX/G	0.19	0.37 (0.58)	0.36 (0.64)		
YQX/L	YQX/L	0.19	0.36 (0.57)	0.32 (0.64)		
YQX/L	YQX/G	0.19	0.36 (0.58)	0.32 (0.64)		
YQX/G	YQX/G	0.22	0.34 (0.40)	0.38 (0.44)		

Table 5.7: *Combined tempo* prediction: Average correlations achieved by the different prediction methods over different datasets. Values in parentheses are the results on the Mozart data with repetitions not excluded from the training set.

the original curve. Taking the previously predicted values into account, the context-sensitive YQX/G leads to a much smoother curve, rather outlining the general trends. The latter approach does not seem to fit the Mozart data very well. The reason for that lies in the stylistic characteristics of Mozart: Gradual tempo changes, like ritardandi or accelerandi, are extremely rare; tempo variations are much more constrained than for instance in the Chopin pieces; small, local variations are much more important and dominant. As confirmed by the quantitative comparison of the different algorithms (section 5.6.1), the Mozart data cannot profit from long term optimization over the of

performance context. Simple inference and local optimization, which, in this example, again, are virtually identical, both reproduce the necessary qualities for Mozart better than global optimization.

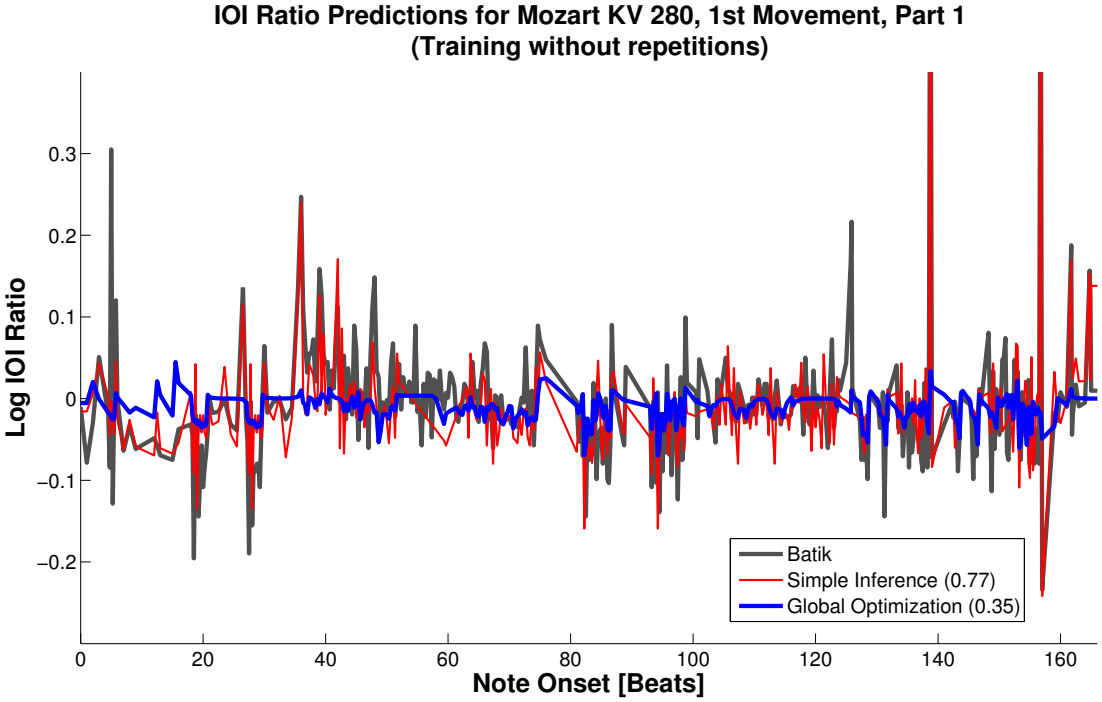


Figure 5.9: Complete tempo curve predicted for Mozart Sonata KV280, 1^{st} Movement, Bars 1-40 with Simple Inference, and Global Optimization. Also shown is the original loudness curve measured from Batik's performance of the piece. The training data did not contain the repetition of passage.

The Chopin pieces pose a much harder nut to crack. Figure 5.10 (upper panel) shows predictions of complete tempo curves for the Mazurka, Op. 56 No. 1, generated by YQX/L and YQX/G. Both algorithms achieve a correlation that is exceptionally high for the Chopin corpus (0.61 (YQX/L) and 0.58 (YQX/G), respectively). However, inspecting the curves behind those numbers, reveals very poor resemblance to the original tempo curve extracted from Magaloff's performance. Scaling and shifting the curves, such that mean and variance are congruent with the average mean and variance encountered in the Chopin data, leads to the curves displayed in the lower panel of figure

5.10. The two faster passages (beats 132-240, and 303-426), can be clearly distinguished in both predictions. Like before in the prediction of the local loudness curve (section 5.5.3), this is due to a coupling between one particular aspect, in the case the rhythmic patterns, and the tempo: predominant in the first, slower 44 bars (beats 1-132, *Allegro non tanto*) is the syncopated rhythm typical for Mazurkas, while the melody voice in the following faster section (*Poco piu mosso, leggiero*) only contains eighth notes.

Composite Tempo

Employing the idea of a composite tempo and exploiting the characteristics of the different algorithms can lead to a tempo prediction that appears more sensible and related to the expected results. Figure 5.11 shows the decomposition of Magaloff's complete tempo curve into the local tempo and note timing components. The local tempo is a smooth curve, setting the general trend of the tempo: the transitions to the fast passages and back are clearly visible, as are small local peaks, consistent with *ritardando-accelerando* combinations, that could coincide with phrase boundaries. The note timing curve is centered around 0 and with rapidly changing values.

Training with local tempo curves instead of complete tempo curves affects the models differently. The upper panel of figure 5.12 shows the respective results. The YQX/L prediction, although heading in the same general direction as Magaloff's local tempo curve, adapts the tempo very slowly. Consequently, the highest tempo is reached at the very end of the fast passage, after an unnaturally long accelerando. Global optimization and simple inference depict the same general tempo, the former a dampened version of the latter, and are in that more similar to the local tempo curve extracted from Magaloff's performance.

The opposite can be observed in the lower panel of figure 5.12. Here, the timing predictions generated by YQX and YQX/G are displayed (differences between local optimization and simple inference are marginal). Again, the curve predicted by the context-sensitive global optimization algorithm, although not as smooth as before, displays discernible trends, and passages coherent with *ritardando-accelerando* combinations. However, a comparison with the targeted note timing curve suggests that this is not appropriate. The curve left after removing the local tempo trends contains the temporal placement of individual notes with respect to the current tempo: Its mean is around 0 (a note being on time) and large positive values (late placement) and large negative values (early placement) may alternate free. The prediction generated without

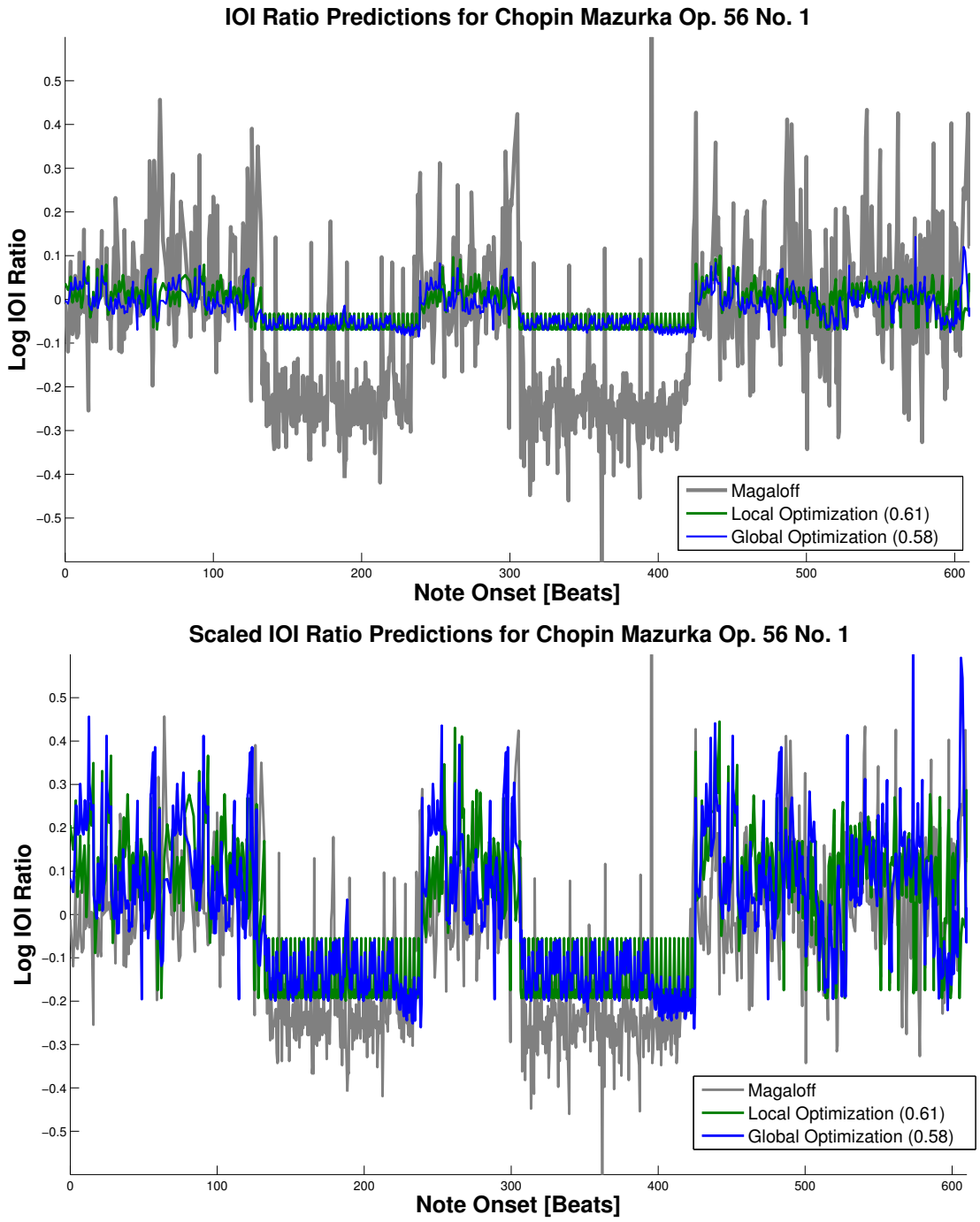


Figure 5.10: *Upper Panel:* Complete tempo curve predictions for Chopin Mazurka Op. 56 No. 1 with local and Global Optimization. *Lower Panel:* Same curves, scaled to match the overall mean and variance in the corpus.

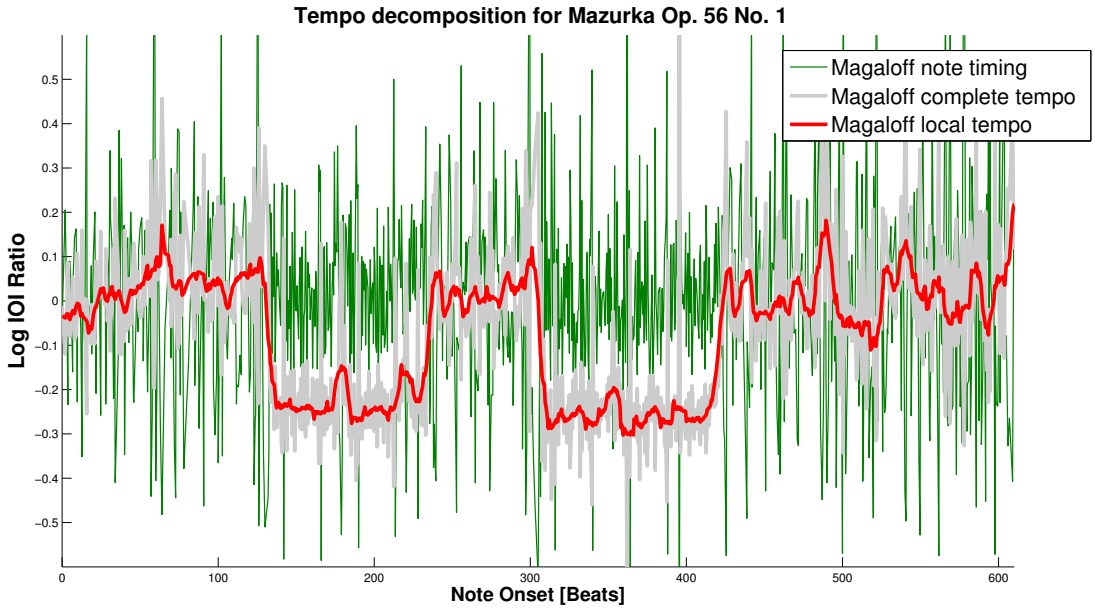


Figure 5.11: Effect of the tempo decomposition applied to Magaloff's performance of Mazurka Op. 56 No. 1.

consideration of the performance context, captures this general idea more adequately.

The two separately predicted curves can be recombined to form the complete tempo curve of the piece. Figure 5.13 shows the result of combining the local tempo curve predicted by YQX/G and the note timing curve predicted by YQX. This seems a reasonable choice considering the characteristics of the individual models. The single model prediction for the complete tempo curve of piece yielded a correlation of 0.61 and 0.58 for Local Optimization and Global Optimization, respectively. The correlations of the combined tempo predictions with the original curve are similar (0.54, 0.62, and 0.63 for influence parameters of 0.3, 0.5, and 0.7, respectively). However, generating the tempo curve this way has two major advantages. (1) Separate models can be based on different score models, taking different aspects of the score into account. The different aspects tempo and timing may prescribe opposite behavior for the same note: For example, a note can at the same time be part of an accelerando and still be placed late with respect to the current tempo. In a setup with two models, the two models base their decisions on different aspects of the score and the local tempo model may decide to

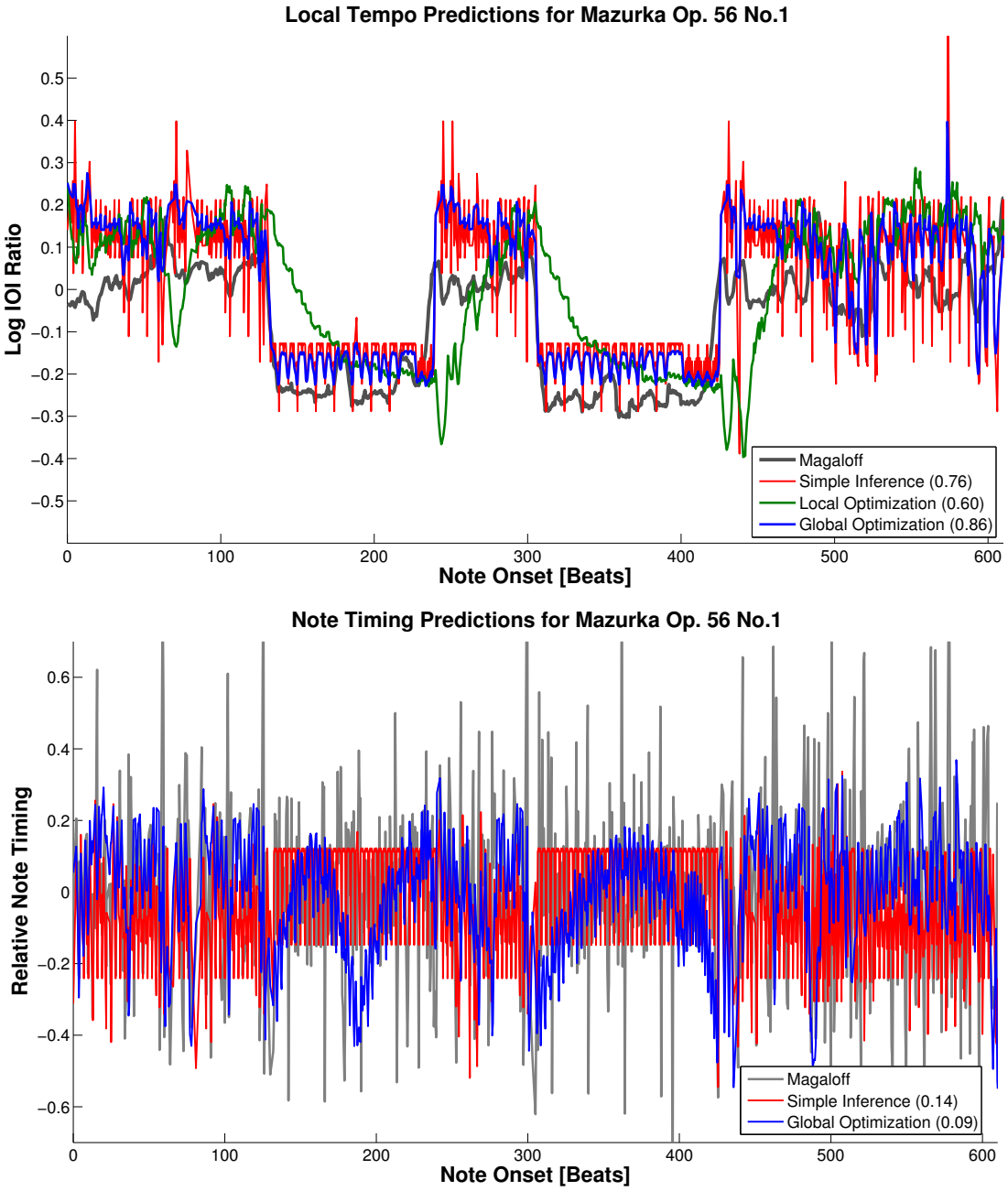


Figure 5.12: *Upper Panel:* Local tempo curve predicted for Chopin Mazurka Op. 56 No. 1 with Simple Inference, Local Optimization, and Global Optimization. *Lower Panel:* Note timing predicted for Chopin Mazurka Op. 56 No. 1 with Simple Inference, and Global Optimization.

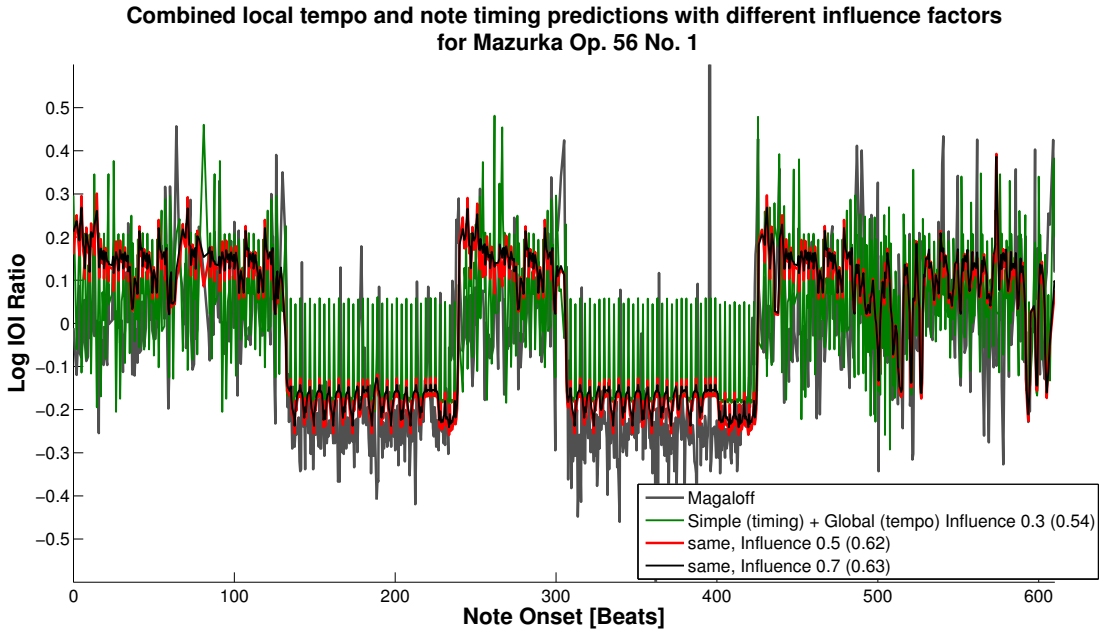


Figure 5.13: Tempo curve for Chopin Mazurka Op. 56 No. 1 as a combination of a note timing prediction with Simple Inference, and local tempo prediction with Global Optimization (three different influence factors). Also shown is the original tempo curve measured from Magaloff's performance of the piece.

predict an accelerando, while the timing model predicts a late placement independently. In a single model setup, the model in a way has to draw two different conclusions from one characterization of the score, and merge them into one prediction. (2) Equation 4.33 in section 4.6.2 introduces a way to assign different levels of influence to the separate curves. Figure 5.13 shows three different combinations of the same curves, with a varying balance between the two components. As expected, the higher the influence, the closer the resulting curve follows the local tempo curve. Provided that a reasonable local tempo curve is established, this can prevent the note timing component disrupting the perception of the tempo trends. The underlying assumption is that local tempo is more important to a expressive rendering than note timing. While this might be unwarranted as a general rule, what can be said is that a discernible local trend – regardless, how musically (in)appropriate it might be in that particular situation – is easier to process

for the human perception than a series of randomly placed notes. Moreover, larger differences between successive values lead to more obvious effects. While obvious effects are desirable in the right places, in the wrong places they disrupt the overall effect much more than applying no effect at all. However, a passage without any fluctuations in the placement of notes sounds equally unnatural. A "safe" approach is therefore to first establish a local tempo curve to introduce discernible tempo trends and then introducing local variation into that curve, with an influence parameter representing the confidence in the local prediction.

5.7 Listener Evaluation – the Rendering Contest RENCON

The Rendering Contest RENCON [53] is an annual international competition, where computer rendered performances are judged and evaluated by a jury and an audience. Ultimately, the event serves as a turing test for performance rendering systems. The project started in 2002 as a satellite workshop of the International Conference on Auditory Display (ICAD2002), and "aims at winning the Chopin Piano Competition, one of the most prestigious international piano competitions, in 2050" [58][1]. In the following I am going to describe the two contests we entered our system into, the RENCON 2008 and 2011, and how the systems were set up.

5.7.1 Putting it all together

Prediction mechanisms, like YQX, form the heart of expressive rendering systems. Several more elements are needed for a complete system. Generally, the following steps have to be executed to render an expressive performance, assuming that the prediction model is already trained. We assume the piece is represented in musicXML format. While not

[1] This is the goal proposed by the originators of the competition, which is, in my opinion, misleading. This contest is the only international scientific forum committed to the quasi-representative evaluation of research in this area. While this in itself is essential and of the utmost importance for any computerized approach to aesthetics and art, this should not be done with the purpose of trying to best human beings in what is a highly creative process. This is, in my opinion neither possible, nor even desirable. It also draws the focus away from trying to understand the elusive art of expressive music performance and find the role computers an artificial intelligence can play in this. Instead it ties up many resources in producing dedicated software, with the sole purpose of trying to win a piano contest. I refer the reader to [128] for some thoughts on creativity in expressive rendering systems.

necessary for the prediction part of the process, rendering the score annotations requires a format able to store this information:

1. Parse the musicXML file of the new piece and convert the musical content to an "expressionless" MIDI
2. Extract the expressive annotations, and generate the corresponding basic tempo and loudness curves.
3. Extract the score model (Feature Extraction)
4. Predict performance trajectories and combine the predicted performance parameters with the result of 2
5. Post processing (Apply additional rules, limit loudness to sensible range, etc.)

A description of the rendering systems we used in the rendering contests RENCON 2008 and RENCON 2011 follows.

5.7.2 RENCON 2008

The RENCON 2008 [55] was hosted alongside the 10^{th} *International Conference on Music Perception and Cognition (ICMPC10)* in Sapporo, Japan. The 2008 competition introduced *on-site expression generation*: two previously unknown pieces, composed specifically for the competition by Prof. Tadahiro Murao, were to be rendered within the time frame of one hour. The two pieces, "My Nocturne" in a Chopin-like style and "My Mozart in Sentiment" in a Mozart-like style (scores of the pieces can be found in appendix C), were intended to capture very prototypical stylistic elements of piano music: the classic, constrained, and delicate Mozart and the romantic Chopin, which requires more pronounced tempo and velocity fluctuations. In contests prior to 2008, participants had been handed the evaluation piece beforehand, and had time to prepare and fine-tune their systems to it.

Two sections were available for the contestants: the *interactive section*, and the *autonomous section*. Time frame and pieces were the same for both sections. Entrants to the autonomous section were not allowed any audio feedback from the system during the rendering process. Hence, the expressive content of the rendered piece was generated without any human adjustments or tailoring parameters to the pieces at hand. Competitors in the interactive section were allowed audio output from their systems which made

it possible to fine-tune the performance. Performances from the different categories were judged separately.

Four contestants entered the autonomous section and competed for three awards: The Rencon award was to be given to a winner selected by audience vote (both through web and on-site voting), the Rencon technical award was to be given to the entrant judged most interesting from a technical point of view, and finally the Rencon Murao Award was to be given to the entrant that most impressed the composer Prof. T. Murao. Our system YQX, as described in section 4.4 won all three prizes (see appendix C). While this is no proof of the absolute quality of the model, it does give some evidence that the model is able to capture and reproduce certain aesthetic qualities of music performance. Videos of YQX performing "My Nocturne" and "My Mozart in Sentiment" at RENCON08 can be seen at `http://www.cp.jku.at/projects/yqx`

5.7.3 YQX 0.1 - the RENCON 2008 model

The system presented in the RENCON 2008 was the first attempt at expressive rendering. Figure 5.14 shows an overview of the system. Prediction of all expressive dimensions was done by the context-free, local YQX model described in 4.4. Based on our experience with the different stylistic attributes of two specific composers and types of pieces, we used different sets of score features for the two test pieces and the different performance targets. For the Chopin-like piece "My Nocturne" we used Duration Ratio and Pitch Interval for articulation, Rhythm Context and IR-arch for tempo, and Duration Ratio and IR-arch for loudness. For the Mozart-like piece "My Mozart in Sentiment" we used Rhythm Context, IR-label, IR-arch and Pitch Interval for all three targets (see 4.2 for a description of the features).

Rendering of expressive annotations in the score was done in a very straight forward way: The global tempo was set as suggested by the metronome markings provided in the score. A basic tempo curve was then set up in the following way: for each *accelerando* (*ritardando*) in the score, the final tempo was determined by multiplying the tempo value with 1.2 (0.8) and interpolating linearly between the values at the beginning and the end. The same was done for loudness, where we used a combination of simple mapping of absolute dynamic annotations, like *p, pp, and f*, to fixed midi velocities, and the same linear interpolation as for tempo.

In 2003, Widmer developed a rule extraction algorithm for musical expression [127, 126]. In Batik's performances of the Mozart sonatas, the algorithm discovered a small

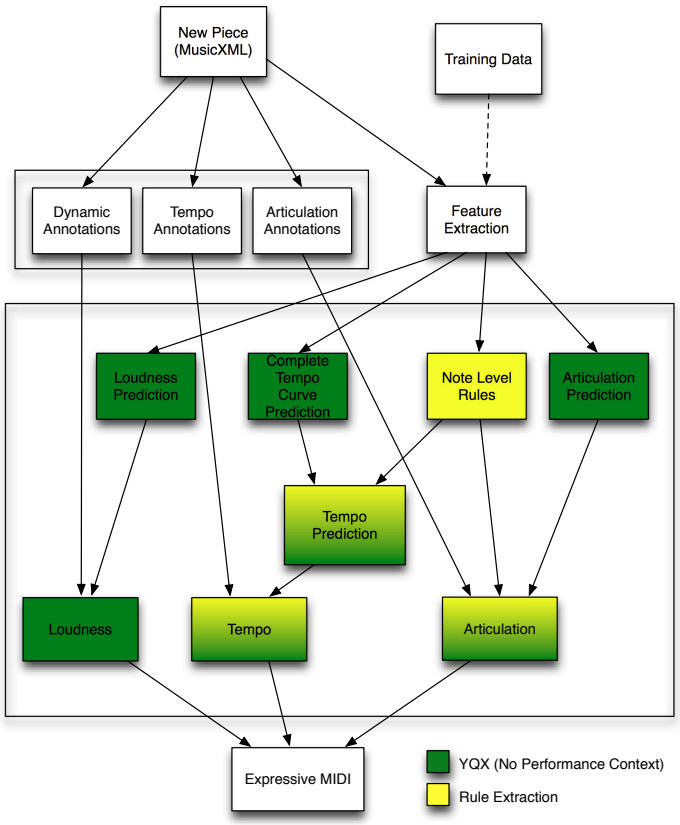


Figure 5.14: Schematic overview of the YQX system used in the RENCON 2008 in Sapporo, Japan. Green boxes only involve the probabilistic graphical model, yellow boxes concern Widmer's performance rules, green and yellow boxes are a result of combining both.

number of simple rules suggesting expressive change under certain melodic or rhythmic circumstances. We use two of the rules to further enhance the aesthetic qualities of the rendered performances:

Staccato Rule: If two successive notes (not exceeding a certain duration) have the same pitch, and the second of the two is longer, then the first note is played staccato. In our implementation the predicted articulation is substituted with a fixed small value, usually around 0.15, which amounts to 15% of the duration in

the score in terms of the current performance tempo.

Delay Next Rule: If two notes of the same length are followed by a longer note, the last note is played with a slight delay. The IOI ratio of the middle note of a triple satisfying the condition is calculated by taking the average of the two preceding notes and adding a fixed amount.

5.7.4 RENCON 2011

The RENCON 2011 [54] was hosted alongside the 8^{th} *Conference on Sound and Music Computing (SMC2011)* in Padua, Italy. The contest was held in two stages. In the first stage, which took place three months before the conference, contestants were given 2 days time to render "A Little Consolation" – a piece specifically composed by Prof. T. Murao for the competition (the score can be found in appendix C). The musical quality of the submissions was rated blindly by 7 reviewers with a strong musical background. The contestants were also required to submit a two-page extended abstract which was reviewed and rated blindly by a technical jury with regard to the technical relevance and interestingness of the submission. The latter established the "Rencon 2011 Technical Award".

The second part of the competition was held in Padua, Italy, as part of the SMC Conference. The set piece, an excerpt from the third movement of Beethoven's Piano Sonata No. 8, op. 13 ("Pathetique"), (the score is displayed in appendix C) was selected randomly from the list of 20 pieces shown in table 5.8, that was announced 2 weeks before the competition. As before, the rendering had to be finished within one hour. Two different performances of the set piece were asked for, displaying different performance styles. The musical quality of the performances was then rated by the audience (both through web and on-site voting) based on the following criterion: 'How much applause would you give the performance?'.

Although systems were marked either autonomous or interactive, they were not judged separately: in both stages the musical evaluation did not distinguish between autonomously and interactively generated performances. The scores from both stages were combined to establish the winner of the "Rencon 2011 Award" for the system with the best musical qualities. Our system, 'YQX featuring the BasisMixer' (which was completely autonomous), won both the technical and the musical award.

J. S. Bach	Wohltemperiertes Klavier Book I, Prelude C Major
J. S. Bach	Two-part Invention No. 15 in B minor
J. S. Bach	Menuette from "Little Notebook for A. M. Bach"
T. Badarzewska	A Maiden's Prayer, Op. 4
L. v. Beethoven	Piano Sonata Op. 13, 3^{rd} Mv.
L. v. Beethoven	Bagatelle No. 25 in A minor 'For Elise', WoO 59
J. Brahms	Hungarian Dance No. 5 in F♯ minor
F. Chopin	Nocturne Op. 9 No. 2 in E♭ Major
F. Chopin	Etude Op. 10 No. 3 in E Major
F. Chopin	Waltz Op. 69 No. 1 in A♭ Major
E. Elgar	Salut d'amour Op. 12
G. Faur	Sicilienne Op. 78 in G minor
G.F. Händel	Aria from 'The Harmonious Blacksmith'
F. Liszt	Etude S. 145, No. 1, 'Waldesrauschen'
F. Mendelssohn	Songs Without Words Op. 30 No. 6 in F♯ minor
W. A. Mozart	Piano Sonata K. 545, 1^{st} Mv.
W. A. Mozart	Piano Sonata K. 331, 3^{rd} Mv.
D. Scarlatti	Sonata in C Major, K. 159
R. Schumann	Abegg Variations Op.1, 'Theme'
P.I. Tchaikovsky	The Seasons Op. 37a No. 7, 'July'

Table 5.8: List of potential set pieces for Stage II of the RENCON 2011.

5.7.5 YQX 0.2 Featuring the BasisMixer - the RENCON 2011 model

The system entered into the RENCON 2011 introduced a multi-level tempo prediction – tempo considered as composed of local tempo and note timing – and an alternative concept of loudness modeling. As proposed in section 4.6, tempo was modeled as a composite dimension. Local tempo and note timing were predicted separately by algorithms with different levels of context awareness: the globally optimized version of YQX (section 4.5.2) was used for local tempo, and the locally optimized version (section 4.5.1) for note timing. The two predictions were then combined to form the prediction for the complete tempo curve, retaining the slow evolving tempo trends of the former for the local tempo component and the fast fluctuations of the latter for the note timing. For articulation, we used the non context-aware predictions of the original YQX. The idea

of combining a rendition of the expressive annotations with a prediction of the residual, local loudness variations (as proposed in section 4.6),was established after the competition. Instead, loudness was entirely modeled by Grachten's Basis Mixer. The two note-level rules described in 5.7.3 were, as before, used to post-process the performance.

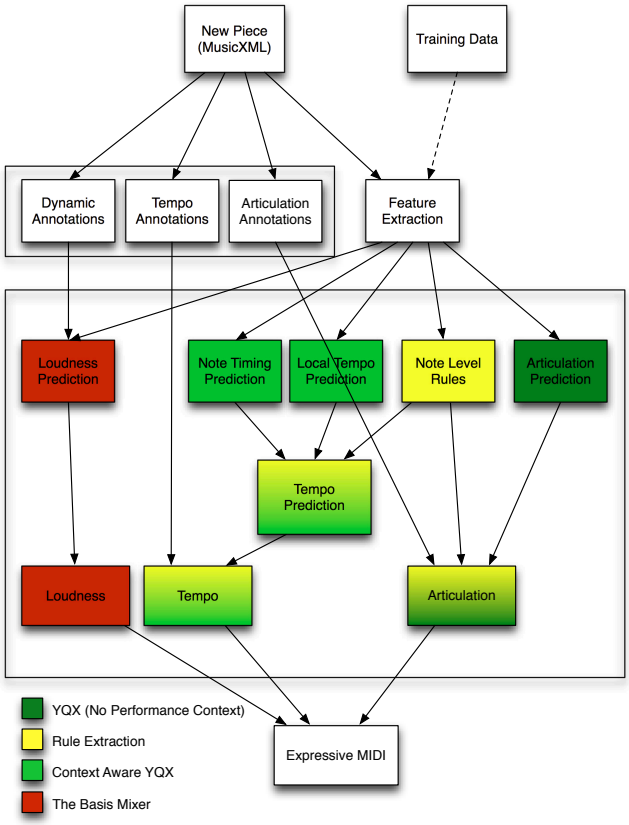


Figure 5.15: Schematic overview of YQX 0.2 Featuring the BasisMixer, as used in the RENCON 2011 in Padua, Italy. Light green boxes indicate a performance context aware probabilistic model, dark green boxes a context-free probabilistic model, yellow boxes concern Widmer's performance rules. Red boxes involve the Basis Mixer.

5.8 Summary

This chapter shows experiments done with the models presented in chapter 4. The different performance dimensions are evaluated separately. First, for each performance dimension the numerical prediction quality of the different algorithms, measured through the similarity between the predicted performance dimension and Magaloff's real performance, is assessed. Then, the results of the different algorithms are evaluated from a qualitative point of view to determine if the produced performance curves exhibit characteristics similar to the real performance curves. A summary of the results follows:

Articulation: The results reflect the fact that articulation is a very local phenomenon without long term dependencies: Sensitivity to the immediate, local context can be beneficial. Integrating a large performance context leads to a prediction with qualities uncharacteristic of the fast fluctuations of the articulation measured in Magaloff's performances. Consequently, although the quantitative evaluation shows a slight lead for the global optimization, it is advisable to use local methods.

Loudness: Loudness seems to be the toughest of the three performance targets to model with our current statistical model. In the original formulation none of the three algorithms produces good results consistently on the Chopin data. Loudness is evidently much more related to the dynamic annotations in the scores. The proposed split of the loudness curve into "annotated loudness" and residual, and assigning focused approaches to each of the problems seems promising. With the current score model the three proposed algorithms still struggle with predicting the residual. A score characterization with more information on phrase boundaries could probably ameliorate the performance. On the Mozart sonatas the results are on the whole acceptable with a clear lead of the context insensitive simple inference algorithm.

Tempo: The benefits of the two proposed extensions are most obvious in the tempo predictions of the Chopin data: Splitting the tempo in local tempo trends and timing of individual notes, and predicting the two components separately with different algorithms achieves two goals: (1) The different algorithms create curves with very different characteristics. The context-aware curves share characteristics of the local tempo curves, the context-insensitive simple inference shares characteristics of the note timing component. This way we can preserve the qualities of both compo-

nents in the result. (2) We can use different score models for the two components. This is congruent with the author's intuition about the matter: decisions how to shape a larger unit, e.g. a phrase, are made based on different information from the score than decisions related to anticipation or delay of individual notes.

Chapter 6

Conclusions and Future Work

This thesis is centered around the Magaloff corpus, a unique collection of performances comprising the complete works for solo piano by Chopin. It describes how we prepared the data, and presents studies of some of the peculiarities of Magaloff's style of playing. The main focus is the application of the data in the field of expressive performance rendering. This chapter summarizes the different parts presented in this thesis, describes further possible applications, and proposes ideas for supplementing and enhancing the results.

6.1 Main Contributions and Results

What sets the Magaloff corpus apart from other collections of recordings is (1) its size – over 10 hours of music, over 150 pieces, around 330.000 played notes, (2) its content – a single pianist playing all pieces of one composer in front of an audience, and (3) its precision – recorded as precisely measured list of note and pedal events on a Bösendorfer computer controlled grand piano instead of audio picked up by a microphone. Making the corpus accessible and preparing the data for research was the firs dt major goal of this thesis. Apart from resulting in a unique resource, this very time-consuming task (digitizing over 900 pages of music scores, converting them into computer readable symbolic scores, correcting the errors of the conversion process, and aligning the digital scores to the performances) yielded a software tool for handling (assembling, inspecting, and correcting) alignments between score and performance, and annotating musical scores.

Several application scenarios exist for a corpus of this precision and dimensions. From an expressive performance research perspective it is both testament of Magaloff's personal style and idiosyncrasies, and representative of piano performance in general. Regarding the former, we examined to what extent Magaloff conformed to a general model of successful aging. As it turned out, the two criteria of the model that were testable – reduction of repertoire, and reduction of performance tempo – are not fulfilled.

This suggests that instead of compromising his musical ideals, Magaloff rather chose to risk an increase in performance errors. Two further exploratory studies were conducted assessing musicological questions. Although, strictly speaking, the results of the studies only apply to Magaloff's performances, the corpus is large enough for the results to still have a certain validity in general. The first study investigated performance errors. More specifically, we examined if, and to what extent, erroneous notes are played in a way that make them more subtle and less noticeable by the audience. The findings mostly corroborate an earlier study by Repp [100], who examined the same question under different circumstances (laboratory conditions rather than a concert situation, graduate piano students instead of a world-class pianist). Further investigations of the phenomenon of errors lead to a catalogue of error patterns: sequences of errors that share commonalities are grouped together and viewed in context. A considerable number of errors are not just isolated mishaps but reoccur and form patterns. Finally, we presented the results of an analysis done in cooperation with W. Goebl that explored the matter of between-hand asynchronies. Similar to ensembles, where often the player currently in charge of the melody voice precedes the others by a perceptible amount of time, Magaloff uses asynchronies between his left and right hand as an expressive device. Two specific types of asynchrony – bass anticipation and tempo rubato – were automatically identified in the data. Both were used to a considerable extent.

Apart from analyzing the artistic and musical content, the corpus plays an essential role as ground truth or training data for data-driven applications. One example for such an endeavor is expressive performance rendering, which is the main focus of this thesis. Chapter 4 proposes a rendering system based on a graphical probabilistic model. Based on a simple model, we develop two important extensions of the system. The first extension integrates awareness of performance context into the design. From a technical point of view, the algorithm developed for this purpose is a closed-form solution to calculate exact inference in a special type of probabilistic network. With regard to the predicted performances this reduces fast fluctuations in the predictions, and emphasizes long term trends. As frequent large changes without discernible trend lead to an unsatisfactory overall shape of a piece, this is a desirable effect for the overall development of tempo or loudness. However, it impedes the equally important local tempo variation. Combining both aspects, long term development and local variation, is the goal of the second extension, which re-defines tempo and loudness: Tempo is regarded as composed of a slowly evolving tempo component and local timing variation. Loudness is also split in

two: one component associated with the performance directives in the score setting the overall evolution and the other one containing local deviations from that. An evaluation of the different components leads to the following conclusions:

- Articulation is best modeled with simple inference or local optimization. Global optimization destroys the characteristic properties of this expressive dimension.

- Decomposing tempo into local tempo trends and note timing improves the prediction quality considerably. This is especially the case when the long term components are predicted using the context-aware model with global optimization, and the local variations are predicted with either short-term performance context or no performance context at all.

- Loudness prediction is the most difficult of the three. As with tempo, it is necessary to regard loudness not as an atomic component, but to deal with different aspects of dynamic change separately. The performance directives in the score are an integral part of loudness evolution and need special treatment. Using the here proposed model to predict the residual *local loudness* does not yet work to a satisfying degree.

All of the above point in the following direction: While probabilistic models certainly are a reasonable approach to performance rendering, there is not one single model that is capable of handling all different aspects of expressive performance. The system we used in the 2011 Rendering contest, as described in 5.7.5, is a combination of several specialized subsystems, using different approaches for different aspects of music performance. The concluding accounts of the two Rendering Contests we participated in and which we won, have to be taken *cum grano salis*: It is warranted to say that there are certain statistical relationships between score and performance which, given a suitable score characterization, can be learned and reproduced, to some degree, by a graphical model. However, there is still a long way to go for computers to be able to a produce something profoundly musical. One major obstacle here is that it is not possible to analyze a music score automatically at the level of understanding and complexity achieved by a musician.

6.2 Future Directions...

6.2.1 ... concerning the state of the corpus

Several aspects can further improve the Magaloff corpus and open up more application scenarios. Audio recordings of all pieces in a studio environment will make the corpus an ideal resource for research on automatic transcription [6]. A collection comprising several hours of audio material with precise timing, pitch, and loudness information of all played notes is both invaluable training data for machine learning algorithms, and ground truth for evaluation. Another application that could greatly benefit from having audio recordings of the complete corpus, is score-following [2]. Efforts are currently made, to record the data on the newer Bösendorfer CEUS system, which replaced the Bösendorfer SE in 2005. This is proving difficult, as the way Magaloff played was tailored to the instrument he played on and the concert hall with its acoustic properties. As no two instruments are identical, some of the elements of Magaloff's performances do not translate well to a different grand piano. Adjustments have to be made to the loudness and pedaling measurements, which introduces a subjective bias into the recordings.

As of now, the score information contained in the corpus mainly concerns the actual note content. Performance directives are only included consistently in selected parts of the data, mainly the Nocturnes. Including these annotations in the complete corpus facilitates detailed studies of how they were realized in the performances.

The interactive score display developed in the jGraphMatch interface (see section 2.5.1) can be extended in a way that broadens the field of application considerably. Replacing the piano roll display with a general graph display the tool can also be used to inspect performance curves interactively and replay performances note-by-note.

6.2.2 ... concerning performance rendering

Hierarchical Score Model

Music is an essentially hierarchical construct of phrases and subphrases, which are reflected in the performance trajectories [119, 130]. By viewing tempo as a composite phenomenon we tried to account for this in the performance model; choosing different sets of score descriptors for the different components is a first approach to extend this to the score model. Eventually, our score model should account for the hierarchy in the compositions. This, however, requires much more elaborate automatic music analysis

of scores than possible at the moment – most important, reliable detection of phrase boundaries, which is, as described in section 4.1.1, only possible with restrictions.

One aspect that might ameliorate the results of phrase analysis in case of the Magaloff corpus is that it contains *slurs*, symbols indicating that the embraced notes are to be played without separation. Consequently, phrase boundaries are more likely to occur on end- or starting points of slurs, than on notes were no slurs begin or end, and very unlikely to occur in the midst of a slur. Hence, the slurs printed in the sheet music give a good indication of structural entities, and can serve as additional cue for phrase detection algorithms.

Performance Target Dependencies

At the moment we predict the three performance targets separately. The underlying assumption is that the performance dimensions are independent of each other. This, of course, is not true. An obvious counterexample are accentuated notes which may combine an increased loudness with a staccato articulation and a slightly delayed onset. Todd proposes that loudness and tempo are coupled together in a "faster equals louder" relation [120]. While this, in general, oversimplifies the issue, the two performance dimensions are certainly linked together. With the proposed model it is possible to simulate this effect by not only considering the performance context of one but of all performance dimensions.

Smoothed Score Features

In [29], we investigated the influence of the *scope* of features, the size of the context they describe, on the prediction quality for tempo and timing. Experiments, in which we predicted note timing and local tempo for the Mozart Sonatas, suggested that the prediction quality for local tempo increased when certain features were averaged over a window of up to four beats, while the quality for note timing prediction decreased with the smoothing. This idea has not been applied in the present scenario, and might lead to additional improvement for the local tempo predictions.

Local Tempo Calculation - Smoothing Window Size

Crucial in the decomposition of the complete tempo into local tempo and note timing it the window size of the moving average. For the presented evaluation we chose the

parameter based on informal experiments and intuition. Especially at phrase boundaries, the wrong choice can have an adverse effect on the result: If we assume that the tempo is reduced rapidly towards a phrase boundary, reaches a minimum at or around a phrase boundary, and rises again, then the average will be considerably higher then the tempo curve at the minimum. So, instead of having the expected pronounced minimum in the local tempo, we actually get a moderate curve for the local tempo while all activity is captured in the residual. This suggests that a sensible smoothing window should not cross phrase boundaries. This, however, requires reliable automatic phrase analysis. A first step in this direction could be to use slurs in the scores as restrictions for the smoothing window (see above).

Alternative efficiency criteria

All quantitative evaluations of our model are based on correlation coefficients between original and predicted curve as a quality measurement. As discussed in section 5.2, this seems inadequate, as musical considerations have no part in the measurement. The objective would be to develop a measurement that works on different hierarchical levels, and on different levels emphasizes different criteria – similar to the different qualitative evaluation criteria applied for the different performance dimensions in chapter 5. Accordingly, for the evaluation of tempo predictions, such a quality measurement could incorporate the following ideas.

Tempo Evaluation: In analogy to the proposed tempo decomposition, the evaluation should be done for local tempo and note timing separately. The local tempo evaluation should focus on overall shape, as well as piecewise trends. One possibility is, to use a segmentation of the piece in musically sensible units, preferably phrases, fit parameterized functions to the local tempo curve in the segments, and calculate the similarity between the fitted functions. For note timing emphasis lies not on the overall trends, but in the realization of individual notes. Furthermore, a suitable quality criterion should take into account that some notes are much more exposed and/or important for the overall impression of the interpretation, and consequently penalize large differences to the original more for important notes. Comprehensively ranking the relative importance of individual notes is of course impossible. Very basic heuristics could be applied, considering, for example, rhythmic position (notes on full beats over notes in-between beats), rhythmic context (notes that disrupt an otherwise homogenous rhythmic pattern more important than notes that

blend into the pattern), and harmonic context (obvious dissonances and harmonic turning points over harmonical "stuffing"). Parncutt's theory of accents [87, 88] is an even more sophisticated alternative to determine salient and important notes in the score.

The local loudness component, the residual after subtracting the part explained by performance directives, could be evaluated in analogy to note timing.

1/f Fractal Noise

Studies of the random fluctuations arising naturally in psychological experiments discovered that they follow the power distribution of $1/f$ noise (pink noise) [40]. It is assumed that the hitherto "unexplained variance" is an intrinsic aspect of tasks involving cognition. Repp [101] also suggests that pianistic expression is not controlled by a deterministic motor system but subject to random variation. Listening and tapping experiments by Rankin et al. [92] lend further evidence to the "determinism of randomness" involved in music performance and cognition. As this seems to be a factor inseparable from music production, and although very subtle, present even in expert performances, it seems appropriate to include this in a system modeling music performance and expression. The GERM model by Juslin et. al. [64] includes a random component based on this observation. This, of course, also applies to the Magaloff Corpus. Hence, all models trying to learn from the data also have to account for the implicit random component. The variance in the data, and with it the complexity of the learning task, could be reduced by eliminating the noise from the data.

Bibliography

[1] Arcos, J., and de Mántaras, R. An interactive CBR approach for generating expressive music. *Journal of Applied Intelligence 27*, 1 (2001), 115–129.

[2] Arzt, A., Widmer, G., and Dixon, S. Automatic page turning for musicians via real-time machine listening. In *Proceedings of the 18th European Conference on Artificial Intelligence (ECAI '08)* (Patras, Greece, 2008).

[3] Baltes, P. B., and Baltes, M. M. Psychological perspectives on successful aging: The model of selective optimization with compensation. In *Successful Aging*, P. B. Baltes and M. M. Baltes, Eds. Cambridge University Press, Cambridge, UK, 1990, pp. 1–34.

[4] Berry, D., Iversen, E., Gudbartsson, D., Hiller, E., Garber, J., Peshkin, B., Lerman, C., Watson, P., Lynch, H., Hilsenbeck, S., Rubinstein, W., Hughes, K., and Parmigiani, G. BRCAPRO validation, sensitivity of genetic testing of BRCA1/BRCA2, and prevalence of other breast cancer susceptibility genes. *Journal of Clinical Oncology 20*, 11 (2002), 2701–2712.

[5] Bisesi, E., Parncutt, R., and Friberg, A. An accent-based approach to performance rendering: Music theory meets music psychology. In *Proceedings of the International Symposium on Performance Science 2011 (ISPS '11)* (Toronto, Canada, 2011).

[6] Böck, S., and Schedl, M. Enhanced Beat Tracking with Context-Aware Neural Networks. In *Proceedings of the 14th International Conference on Digital Audio Effects 2011 (DAFx-11)* (Paris, France, 2011).

[7] Bonkowski, W. Nikita Magaloff. `http://en.chopin.nifc.pl/chopin/persons/detail/name/magaloff/id/6636`.

[8] Buntine, W. A guide to the literature on learning probabilistic networks from data. *IEEE Transactions on Knowledge and Data Engineering 8* (1996), 195–210.

[9] Cambouropoulos, E. Musical rhythm: A formal model for determining local boundaries, accents and metre in a melodic surface. In *Music, Gestalt, and Computing*, M. Leman, Ed., vol. 1317 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 1997, pp. 277–293.

[10] Cambouropoulos, E. The local boundary detection model (LBDM) and its application in the study of expressive timing. In *Proceedings of the International Computer Music Conference 2001 (ICMC '01)* (Havana, Cuba, 2001).

[11] Canazza, S., dePoli, G., Drioli, C., Rodà, A., and Vidolin, A. Modeling and control of expressiveness in music performance. *Proceedings of the IEEE 92*, 4 (2004), 686–701.

[12] Canazza, S., dePoli, G., Rodà, A., and a Vidolin. An abstract control space for communication of sensory expressive intentions in music performance. *Journal of New Music Research 32*, 3 (2003), 281–294.

[13] Cella, F., and Magaloff, I. *Nikita Magaloff*. Nuove Edizione, Milano, 1995.

[14] Clarke, E. F. Some aspects of rhythm and expression in performances of Erik Satie's "Gnossienne No. 5". *Music Perception 2* (1985), 299–328.

[15] Dannenberg, R. An on-line algorithm for real-time accompaniment. In *Proceedings of the International Computer Music Conference 1984 (ICMC '84)* (Paris, France, 1984).

[16] Dannenberg, R., and Hu, N. Polyphonic audio matching for score following and intelligent audio editors. In *Proceedings of the International Computer Music Conference 2003 (ICMC '03)* (San Francisco, CA, USA, 2003).

[17] Davis, G. A. Bayesian reconstruction of traffic accidents. *Law, Probability and Risk 2*, 2 (2003), 69–89.

[18] de Campos, L. M., Fernandez-Luna, J. M., and Huete, J. F. Bayesian networks and information retrieval: an introduction to the special issue. *Information Processing & Management 40*, 5 (2004), 727 – 733.

[19] Dixon, S. Automatic extraction of tempo and beat from expressive performances. *Journal of New Music Research 30*, 1 (2001), 39–58.

[20] Dixon, S. Evaluation of the audio beat tracking system beatroot. *Journal of New Music Research 36* (2007), 39–50.

[21] Dixon, S., Goebl, W., and Widmer, G. The performance worm: Real time visualisation of expression based on Langner's tempo-loudness animation. In *Proceedings of the International Computer Music Conference 2002 (ICMC '02)* (Goteborg, Sweden, 2002).

[22] Dixon, S., Goebl, W., and Widmer, G. The "Air Worm": An interface for real-time manipulation of expressive music performance. In *Proceedings of the International Computer Music Conference 2005 (ICMC '05)* (Barcelona, Spain, 2005).

[23] Dixon, S., and Widmer, G. Match: A musical alignment tool chest. In *Proceedings of the 6th International Conference on Music Information Retrieval 2005 (ISMIR '05)* (London, UK, 2005).

[24] Dorard, L., Hardoon, D., and Shawe-Taylor, J. Can style be learned? A machine learning approach towards 'performing' as famous pianists. In *Proceedings of Music, Brain & Cognition Workshop - The Neural Information Processing Systems 2007 (NIPS 2007)* (Whistler, Canada, 2007).

[25] Duda, R. O., Hart, P. E., and Stork, D. G. *Pattern Classification*, 2. ed. Wiley, New York, 2001.

[26] Flossmann, S., Goebl, W., Grachten, M., Niedermayer, B., and Widmer, G. The Magaloff Project: An interim report. *Journal of New Music Research 39*, 4 (2010), 363–377.

[27] Flossmann, S., Goebl, W., and Widmer, G. Maintaining skill across the life span: Magaloff's entire chopin at age 77. In *Proceedings of the International*

Symposium on Performance Science 2009 (ISPS '09) (Auckland, New Zealand, 2009).

[28] Flossmann, S., Goebl, W., and Widmer, G. The magaloff corpus: An empirical error study. In *Proceedings of the 10th International Conference on Music Perception and Cognition 2010 (ICMPC '10)* (Seattle, WA, USA, 2010).

[29] Flossmann, S., Grachten, M., and Widmer, G. Experimentally investigating the use of score features for computational models of expressive timing. In *Proceedings of the 10th International Conference on Music Perception and Cognition 2008 (ICMPC '08)* (Sapporo, Japan, 2008).

[30] Flossmann, S., Grachten, M., and Widmer, G. Expressive performance rendering: Introducing performance context. In *Proceedings of the 6th Sound and Music Computing Conference 2009 (SMC '09)* (Porto, Portugal, 2009).

[31] Flossmann, S., Grachten, M., and Widmer, G. Expressive performance rendering with probabilistic models. In *Guide to Computing for Expressive Music Performance (in press)*, E. Miranda and A. Kirke, Eds. Springer, 2012.

[32] Flossmann, S., and Widmer, G. Toward a model of performance errors: A qualitative review of Magaloff's Chopin. In *Proceedings of the International Symposium on Performance Science 2011 (ISPS '11)* (Toronto, Canada, 2011).

[33] Flossmann, S., and Widmer, G. Toward a multilevel model of expressive piano performance. In *Proceedings of the International Symposium on Performance Science 2011 (ISPS '11)* (Toronto, Canada, 2011).

[34] Forte, A. Schenker's conception of musical structure. *Journal of Music Theory 3*, 1 (1959), 1–30.

[35] Friberg, A. *A Quantitative Rule System for Musical Performance*. PhD thesis, Royal Institute of Technology, Stockholm, 1995.

[36] Friberg, A., Bresin, R., and Sundberg, J. Overview of the KTH rule system for musical performance. *Advances in Cognitive Psychology 2*, 2-3 (2006), 145–161.

[37] Friberg, A., and Sundberg, J. Does music performance allude to locomotion? A model of final ritardandi derived from measurements of stopping runners. *Journal of the Acoustical Society of America 105*, 3 (1999), 1469–1484.

[38] Friberg, A., and SundströM, A. Swing ratios and ensemble timing in jazz performance: Evidence for a common rhythmic pattern. *Music Perception 19*, 3 (2002), 333–349.

[39] Fry, S., and Lihoreau, T. *Stephen Fry's incomplete and utter history of classical music*. Pan Books, London, 2005.

[40] Gilden, D. L. Cognitive emissions of 1/f noise. *Psychological Review 108* (2001), 33–56.

[41] Gingras, B., and McAdams, S. Improved score-performance matching using both structural and temporal information from midi recordings. *Journal of New Music Research 40*, 1 (2011), 43–57.

[42] Goebl, W. Numerisch-klassifikatorische Interpretationsanalysis mit dem "Bösendorfer Computerflügel" [Numerical-classificatorical performance analysis with the bösendorfer computer controlled grand piano]. Master's thesis, University of Vienna, Vienna, 1999.

[43] Goebl, W., and Bresin, R. Measurement and reproduction accuracy of computer-controlled grand pianos. *Journal of the Acoustical Society of America 114*, 4 (2003), 2273–2283.

[44] Goebl, W., Flossmann, S., and Widmer, G. Computational investigations into between-hand synchronization in piano playing: Magaloff's complete chopin. In *Proceedings of the 6th Sound and Music Computing Conference 2009 (SMC '09)* (Porto, Portugal, 2009).

[45] Goebl, W., Flossmann, S., and Widmer, G. Investigations into between-hand synchronisation in magaloff's chopin. *Computer Music Journal 34*, 3 (2010), 35–44.

[46] Grachten, M. *Expressivity-Aware Tempo Transformations of Music Performances Using Case Based Reasoning*. PhD thesis, Pompeu Fabra University, Barcelona, 2006.

[47] Grachten, M., and Widmer, G. Who is who in the end? recognizing pianists by their final ritardandi. In *Proceedings of the 10th International Society for Music Information Retrieval Conference 2009 (ISMIR '09)* (Kobe, Japan, 2009).

[48] Grachten, M., and Widmer, G. Explaining musical expression as a mixture of basis functions. In *Proceedings of the 8th International Conference of Sound and Music Computing 2011 (SMC '11)* (Padova, Italy, 2011).

[49] Grachten, M., and Widmer, G. A method to determine the contribution of annotated performance directives in music performances. In *Proceedings of the International Symposium on Performance Science 2011 (ISPS '11)* (Toronto, Canada, 2011).

[50] Grindlay, G., and Helmbold, D. Modeling, analyzing, and synthesizing expressive piano performance with graphical models. *Machine Learning 65*, 2-3 (2006), 361–387.

[51] Grindlay, G. C. Modeling expressive musical performance with Hidden Markov Models. Master's thesis, University of California, Santa Cruz, 2005.

[52] Hamanaka, M., Hirata, K., and Tojo, S. Implementing "A Generative Theory of Tonal Music". *Journal of New Music Research 35*, 4 (2006), 249–277.

[53] Hashida, M. RENCON - Performance Rendering Contest for computer systems. http://www.renconmusic.org/, September 2008.

[54] Hashida, M., Hirata, K., and Katayose, H. Rencon workshop 2011: Performance rendering contest for computer systems. In *Proceedings of the 8th Sound and Music Computing Conference 2011 (SMC '11)* (Padova, Italy, 2011).

[55] Hashida, M., Nakra, T., Katayose, H., Murao, T., Hirata, K., Suzuki, K., and Kitahara, T. Rencon: Performance rendering contest for automated music systems. In *Proceedings of the 10th International Conference on Music Perception and Cognition 2008 (ICMPC '08)* (Sapporo, Japan, 2008).

[56] Heijink, H., Desain, P., Honing, H., and Windsor, L. Make me a match: An evaluation of different approaches to score-performance matching. *Computer Music Journal 24*, 1 (2000), 43–56.

[57] Herttrich, E. *Chopin Klavierstücke, Urtext*. G. Henle Verlag, Munich, Germany, 1978.

[58] Hiraga, R., Hashida, M., Katayose, H., Hirata, K., and Noike, K. Rencon: toward a new evaluation method for performance rendering systems. In *Proceedings of the International Computer Music Conference 2002 (ICMC '02)* (Gothenburg, Sweden, 2002).

[59] Honing, H. Poco: An environment for analyzing, modifying and generating expression in music. In *Proceedings of the International Computer Music Conference 1990 (ICMC '90)* (San Francisco, CA, USA, 1990).

[60] Honing, H. Is there a perception-based alternative to kinematic models of tempo rubato? *Music Perception 23*, 1 (2005), 79–85.

[61] Hudson, R. *Stolen Time: The History of Tempo Rubato.* Clarendon Press, Oxford, 1994.

[62] Juang, B., and Rabiner, L. An introduction to Hidden Markov Models. *IEEE ASSP magazine* (1986), 4–16.

[63] Juang, B. H., and Rabiner, L. R. Hidden Markov Models for speech recognition. *Technometrics 33*, 3 (1991), 251–272.

[64] Juslin, P. N., Friberg, A., and Bresin, R. Toward a computational model of expression in music performance: The GERM model. *Musicae Scientiae (special Issue 2001/2)* (2002), 63–122.

[65] Kim, T. H., Fukayama, S., Nishimoto, T., and Sagayama, S. Performance rendering for polyphonic piano music with a combination of probabilistic models for melody and harmony. In *Proceedings of the 7th Sound and Music Computing Conference 2010 (SMC '10)* (Barcelona, Spain, 2010).

[66] Krause, P., Boyle, D. P., and Bäse, F. Comparison of different efficiency criteria for hydrological model assessment. *Advances in Geosciences 5* (2005), 89–97.

[67] Krumhansl, C. L., and Kessler, E. J. Tracing the dynamic changes in perceived tonal organization in a spatioal representation of musical keys. *Psychological Review 89* (1982), 334–368.

[68] Large, E. W. Dynamic programming for the analysis of serial behaviors. *Behaviors Research Methods Instruments & Computers 25*, 2 (1993), 238–241.

[69] Lerdahl, F., and Jackendoff, R. *A Generative Theory of Tonal Music*. The MIT Press, Cambridge, 1983.

[70] Mazzola, G. *The Topos of Music - Geometric Logic of Concepts, Theory, and Performance*. Birkhäuser Verlag, Basel, 2002.

[71] Mazzola, G. Rubato software. http://www.rubato.org, 2006.

[72] Meyer, L. *Emotion and meaning in Music*. University of Chicago Press, Chicago, 1956.

[73] Milmeister, G. *The Rubato Composer Music Software: Component-Based Implementation of a Functorial Concept Architecture*. PhD thesis, Universität Zürich, Zürich, 2006.

[74] Moog, R. A., and Rhea, T. L. Evolution of the keyboard interface: The Boesendorfer 290 SE recording piano and the moog multiply-touch-sensitive keyboards. *Computer Music Journal 14*, 2 (1990), 52–60.

[75] Mueller, M., Kurth, F., and Clausen, M. Audio matching via chroma-based statistical features. In *Proceedings of the 5th International Conference on Music Information Retrieval 2005 (ISMIR '05)* (London, UK, 2005).

[76] Murphy, K. *Dynamic Bayesian Networks: Presentation, Inference and Learning*. PhD thesis, University of California, Berkeley, 2002.

[77] Murphy, K. P. How to use the bayes net toolbox. http://bnt.googlecode.com/svn/trunk/docs/usage.html, 2007.

[78] Narmour, E. *The Analysis and Cognition of Basic Melodic Structures: The Implication-Realization Model*. University of Chicago Press, Chicago, 1990.

[79] Narmour, E. *The Analysis and Cognition of Melodic Complexity: The Implication-Realization Model*. University of Chicago Press, Chicago, 1992.

[80] Niedermayer, B. Non-negative matrix division for the automatic transcription of polyphonic music. In *Proceedings of the 9th International Conference on Music Information Retrieval 2008 (ISMIR '08)* (Philadelphia, PA, USA, 2008).

[81] Paderewski, I. J. *Chopin Complete Works, XII Rondos*, 14 ed. The Fryderyk Chopin Institute, Warsaw, Poland, 1999.

[82] Paderewski, I. J. *Chopin Complete Works, VI Sonatas*, 17 ed. The Fryderyk Chopin Institute, Warsaw, Poland, 2006.

[83] Palmer, C., and Holleran, S. Harmonic, melodic, and frequency height influences in the perception of multivoiced music. *Perception & Psychophysics 56*, 3 (1994), 301–312.

[84] Palmer, C., and van de Sande, C. Units of knowledge in music performance. *Journal of Experimental Psychology 19*, 2 (1993), 457–470.

[85] Palmer, C., and van de Sande, C. Range of planning in music performance. *Journal of Experimental Psychology: Human Perception and Performance 21*, 5 (1995), 947–962.

[86] Pardo, B., and Birmingham, W. Improved score following for acoustic performances. In *Proceedings of the International Computer Music Conference 2002 (ICMC '02)* (Gothenborg, Sweden, 2002).

[87] Parncutt, R. Accents and expression in piano performance. In *Perspektiven und Methoden einer Systematischen Musikwissenschaft (Festschrift Fricke)*, K. W. Niemöller and B. Gätjen, Eds. Peter Lang, Germany, 2003, pp. 163–185.

[88] Parncutt, R. Modeling immanend durational accent in musical rhythm. In *Proceedings of the 5th Triennial Conference of the European Society for the Cognitive Sciences of Music (ESCOM '03)* (Hannover, Germany, 2003).

[89] Perez, A., Maestre, E., Ramirez, R., and Kersten, S. Expressive irish fiddle performance model informed with bowing. In *Proceedings of the International Computer Music Conference 2008 (ICMC '08)* (Belfast, Northern Ireland, 2008).

[90] Pickens, J. Feature selection for polyphonic music retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New Orleans, LA, USA, 2001).

[91] Ramirez, R., Hazan, A., Gòmez, E., and Maestre, E. Understanding expressive transformations in saxophone Jazz performances using inductive machine

learning in saxophone jazz performances using inductive machine learning. In *Proceedings of the Sound and Music Computing International Conference 2004 (SMC '04)* (Paris, France, 2004).

[92] Rankin, S. K., Large, E. W., and Fink, P. W. Fractal tempo fluctuation and pulse prediction. *Music Perception 26*, 5 (2009), 401–413.

[93] Raphael, C. Aligning music audio with symbolic scores using a hybrid graphical model. *Machine Learning 65*, 2-3 (2006), 389–409.

[94] Raphael, C. Music plus one and machine learning. In *Proceedings of the 27th International Conference on Machine Learning (ICML '10)* (Haifa, Israel, 2010).

[95] Rasch, R. A. Synchronization in performed ensemble music. *Acustica 43* (1979), 121–131.

[96] Rasmussen, C. E., and Williams, C. K. I. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[97] Recordare. MusicXML definition. http://www.recordare.com/xml.html, 2003.

[98] Repp, B. H. Diversity and commonality in music performance - an analysis of timing microstructure in Schumann's "Träumerei". *Journal of the Acoustical Society of America 92*, 5 (1992), 2546–2568.

[99] Repp, B. H. Expressive timing in Schumann's "Träumerei": An analysis of performances by graduate student pianists. *Journal of the Acoustical Society of America 98* (1995), 2413–2427.

[100] Repp, B. H. The art of inaccuracy: Why pianists' errors are difficult to hear. *Music Perception 14*, 2 (1996), 161–184.

[101] Repp, B. H. Variability of timing in expressive piano performance increases with interval duration. *Psychonomic Bulletin and Review 4* (1997), 530–534.

[102] Rosenblum, S. P. *Performance Practices in Classic Piano Music*. Indiana University Press, Bloomington, IN, USA, 1988.

[103] Ruiz, M., Strübing, F., Jabusch, H., and Altenmüller, E. EEG oscillatory patterns are associated with error prediction during music performance and are altered in musician's dystonia. *Neuroimage 55*, 4 (2011), 1791–1803.

[104] Russell, S., and Norvig, P. *Artificial Intelligence: A Modern Approach*, 2nd ed. Prentice-Hall, Englewood Cliffs, NJ, 2003.

[105] Schaffrath, H. *The Essen Folksong Collection in Kern Format.* Center for Computer Assisted Research in the Humanities, Menlo Park, CA, 1995.

[106] Schellenberg, G. E. Expectancy in melody: Tests of the Implication-Realization model. *Cognition 58*, 1 (1996), 75–125.

[107] Shaffer, L. H., Clarke, E. F., and Todd, N. P. M. Metre and rhythm in pianoplaying. *Cognition 20* (1985), 61–77.

[108] Sinkovicz, W. Chopin für Magaloff Fans [Chopin for Magaloff fans]. *Die Presse Mittwoch, 18. Jänner [Wednesday, January 18]* (1989).

[109] Sloboda, J. A. *The musical mind: The cognitive psychology of music.* Oxford: Clarendon Press, 1985.

[110] Stadler, P. Chopin-Marathon [Chopin-marathon]. *Der Standard 18. Jänner [January 18]* (1989).

[111] Sulzer, J., Blankenburg, C., Schulz, J., Kirnberger, J., and Tonelli, G. *Allgemeine Theorie der schönen Künste: in einzelnen, nach alphabetischer Ordnung der Kunstwörter aufeinanderfolgenden, Artikeln abgehandelt [General Theory of the Beaux Arts]*, 2 ed. G. Olms, 1799.

[112] Sundberg, J., Askenfelt, A., and Frydén, L. Musical performance. A synthesis-by-rule approach. *Computer Music Journal 7* (1983), 37–43.

[113] Suzuki, T. The second phase development of case based performance rendering system "Kagurame". In *Working Notes of the IJCAI-03 Rencon Workshop* (Acapulco, Mexico, 2003).

[114] Taruskin, R. *Text and act: Essays on music and performance.* Oxford University Press, New York, 1995.

[115] Temperley, D. *The cognition of basic musical structures.* The MIT Press, Cambridge, MA, USA, 2001.

[116] Temperley, D. *Music And Probability.* MIT Press, Cambridge, MA, USA, 2007.

[117] Teramura, K., Okuma, H., and et al. Gaussian process regression for rendering music performance. In *Proceedings of the 10th International Conference on Music Perception and Cognition 2008 (ICMPC '08)* (Sapporo, Japan, 2008).

[118] Tobudic, A., and Widmer, G. Learning to play like the great pianists. In *Proceedings of the 19th International Joint Conference on Artificial intelligence (IJCAI '05)* (2005).

[119] Tobudic, A., and Widmer, G. Relational IBL in classical music. *Mach. Learn. 64*, 1-3 (2006), 5–24.

[120] Todd, N. P. M. The dynamics of dynamics: A model of musical expression. *Journal of the Acoustical Society of America 91* (1992), 3450–3550.

[121] Vitouch, O. Erwerb musikalischer expertise [Acquisition of musical expertise]. In *Allgemeine Musikpsychologie (Enzyklopädie der Psychologie)*, T. H. Stoffer and R. Oerter, Eds., vol. D/VII/1. Hogrefe, Göttingen, Germany, 2005, pp. 657–715.

[122] Wagner, R. A., and Fischer, M. J. The string-to-string correction problem. *Journal of the Association of Computing Machinery 21*, 1 (1974), 168–173.

[123] Watson, A. H. D. What can studying musicians tell us about motor control of the hand? *Journal of Anatomy 208* (2006), 527–542.

[124] Widmer, G. Learning expressive performance: The structure-level approach. *Journal of New Music Research 25*, 2 (1996), 179–205.

[125] Widmer, G. Large-scale induction of expressive performance rules: first quantitative results. In *Proceedings of the International Computer Music Conference 2000 (ICMC '00)* (Berlin, Germany, 2000).

[126] Widmer, G. Machine discoveries: A few simple, robust local expression principles. *Journal of New Music Research 31*, 1 (2002), 37–50.

[127] Widmer, G. Discovering simple rules in complex data: A meta-learning algorithm and some surprising musical discoveries. *Artificial Intelligence 146*, 2 (2003), 129–148.

[128] Widmer, G., Flossmann, S., and Grachten, M. YQX plays Chopin. *AI Magazine 30*, 3 (2009), 35–48.

[129] Widmer, G., and Goebl, W. Computational models of expressive music performance: The state of the art. *Journal of New Music Research 33*, 3 (2004), 203–216.

[130] Widmer, G., and Tobudic, A. Playing Mozart by analogy: Learning multi-level timing and dynamics strategies. *Journal of New Music Research 32*, 3 (2003), 259–268.

[131] Widmer, G., and Zanon, P. Automatic recognition of famous artists by machine. In *Proceedings of the 16th European Conference on Artificial Intelligence 2004 (ECAI '04)* (Amsterdam, The Netherlands, 2004).

[132] Williamon, A. *Musical Excellence*. Oxford University Press., Oxford, 2004.

[133] Zimmermann, E. *Chopin Balladen, Urtext*. G. Henle Verlag, München, Germany, 1976.

[134] Zimmermann, E. *Chopin Klaviersonate B-moll, Opus 35, Urtext*. G. Henle Verlag, München, Germany, 1976.

[135] Zimmermann, E. *Chopin Walzer, Urtext*. G. Henle Verlag, München, Germany, 1978.

[136] Zimmermann, E. *Chopin Nocturnes, Urtext*. G. Henle Verlag, München, Germany, 1980.

[137] Zimmermann, E. *Chopin Etüden, Urtext*. G. Henle Verlag, München, Germany, 1983.

[138] Zimmermann, E. *Chopin Polonaisen, Urtext*. G. Henle Verlag, München, Germany, 1990.

[139] Zimmermann, E. *Chopin Préludes, Urtext*. G. Henle Verlag, München, Germany, 1996.

[140] Zimmermann, E. *Chopin Andante spianato und Grande Polonaise brillante, Op. 22, Urtext*. G. Henle Verlag, München, Germany, 1998.

[141] Zimmermann, E. *Chopin Impromptus, Urtext*. G. Henle Verlag, München, Germany, 1999.

[142] Zimmermann, E. *Chopin Scherzi, Urtext.* G. Henle Verlag, München, Germany, 2001.

[143] Zimmermann, E. *Chopin Mazurken, Urtext.* G. Henle Verlag, München, Germany, 2003.

[144] Zimmermann, E. *Chopin Klaviersonate H-moll, Opus 58, Urtext.* G. Henle Verlag, München, Germany, 2004.

Appendix A

Graphical Probabilistic Networks

The following describes a very generic setup, encountered frequently in all types of different situations: several interdependent factors constitute a system; some of the factors can be measured (e.g. through any kind of sensor) and some can be determined indirectly or only guessed at from the state of the rest of the system; changing one factor influences the most likely state the other factors are in. Together the factors form a (possibly very large) joint probability distribution that covers all possible states the complete system can be in. As the number of possible states of the complete system grows exponentially with the number of variables, calculating the joint probability distribution is unfeasible in most cases.

Probabilistic graphical models use graphs to encode such distributions over multi-dimensional spaces. The graph is a representation of the set of statistical (in)dependences between the variables in the system. If the edges in the graph are directed, the model is called a *Bayesian Network* (BN) (or Belief Network, causal model, etc), undirected graphs are the characteristic of *Markov Random Fields* (MRF) (or Markov Networks). Due to the independence statements that can be encoded by Markov Networks, undirected models are mainly used in image processing, computer vision, and physics. The family will not be covered here. Bayesian Networks are used to model expert knowledge in a variety of fields ranging from risk calculation in cancer research [4] and law [17] to Information Retrieval in Process management [18]. *Dynamic Bayesian Networks* (DBNs) are an extension of BNs that can be used to monitor systems over time, with variables depending not only on the current but also on previous states of the system. Algorithms exist that (1) estimate the model parameters from data, which is called *training*, and (2) calculate the changes in the probability distributions caused by parts of the system becoming known, a process called *inference*.

A short, informal overview to Bayesian networks (A.2) and Dynamic Bayesian Networks (A.3) is given in this chapter. This only serves the purpose of introducing notation and general concepts and is by no means meant to be exhaustive. A very thorough and

excellent introduction can be found in [76].

A.1 Basic statistical concepts

In this section the statistical elements used in this introduction and in chapter 4 of the thesis are listed and defined. A thorough introduction to random processes and variables, and probability theory can for instance be found in [25].

A.1.1 Discrete Random Variables

A random variable, denoted by a capital letter, either has continuous or discrete states. Continuous variables are here usually marked X or Y, discrete variables are marked Q. The different states $\{q_1, \cdots, q_n\}$ that a discrete variable Q can be in, are finite and countable, and form the *domain* $\mathbb{Q}$ of Q. The distribution $P(Q)$ of a discrete random variable Q assigns a probability $P(Q = q_i)$ to all $q_i \in \mathbb{Q}$. This is often abbreviated to $P(q_i)$. The probability distribution of a single discrete variable is represented by a vector containing the probabilities of the different states. A set of several random variables is denoted by a bold letter: $\mathbf{Q} = \{Q_1, \cdots, Q_n\}$.

Joint and Conditional Probability Distributions

Joint probability distributions (JPDs) set probabilities for two or more variables simultaneously being in designated states and are represented by $P(Q_1, \cdots, Q_n)$, $P(Q_i)$, or $P(\mathbf{Q})$. $P(Q_1 = q_1, \cdots, Q_n = q_n)$, or $P(q_1, \cdots, q_n)$ for simplicity, represents the probability of variables Q_1 to Q_n simultaneously being in states q_1 to q_n respectively. The joint domain $\mathcal{D}(\mathbf{Q})$ of variables $\mathbf{Q} = \{Q_1, \cdots, Q_n\}$ is the product of all domains:

$$\mathcal{D}(\mathbf{Q}) = \prod_{i=1}^{n} \mathbb{Q}_i \tag{A.1}$$

$$= \mathbb{Q}_1 \times \mathbb{Q}_2 \times \cdots \times \mathbb{Q}_n \tag{A.2}$$

$$= \{(q_1, \cdots, q_n) | q_1 \in \mathbb{Q}_1, \cdots, q_n \in \mathbb{Q}_n\} \tag{A.3}$$

The elements in $\mathcal{D}(\mathbf{Q})$ are n-dimensional tuples $\mathbf{q} = (q_1, \cdots, q_n)$. The size of a JPD is the product of the sizes of all involved variables. For $P(\mathbf{Q})$ to be a well-defined probability distribution, the probabilities for all $\mathbf{q} \in \mathcal{D}(\mathbf{Q})$ have sum to 1. A JPD of 2 variables can be represented by a table (Joint probability Table, JPT) with one variable along the

vertical and one along the horizontal axis of the table. Each additional variable adds one dimension. Usually, JPDs of more than 2 variables are represented by a 2 dimensional table with one variable along the top, and all possible combinations of the remaining variables along the left side.

The conditional probability distribution (CPD) of a variable Q_1 given that the state of another variable Q_2 is known to be q_2, is indicated by $P(Q_1|Q_2)$. The CPD of Q_1 given $Q_2 = q_2$ is calculated by normalizing the joint probability of the two variables with the total probability of Q_2 being in state q_2:

$$P(Q_1|Q_2 = q_2) = \frac{P(Q_1, Q_2 = q_2)}{P(Q_2 = q_2)}. \tag{A.4}$$

If the JPD of Q_1 and Q_2 is known, the probability $P(Q_2 = q_2)$ can be calculated using the law of *total probability*:

$$P(Q_2 = q_2) = \sum_{q_1 \in \mathbb{Q}_1} P(Q_1 = q_1, Q_2 = q_2) = \sum_{q_1 \in \mathbb{Q}_1} P(q_2|q_1)P(q_1). \tag{A.5}$$

For n random variables $\mathbf{Q} = \{Q_1, \cdots, Q_n\}$ this takes the following form:

$$\begin{aligned} P(Q_1) &= \sum_{\mathbf{q} \in \mathcal{D}(\mathbf{Q} \setminus Q_1)} P(Q_1, \mathbf{q}) && \text{(A.6)} \\ &= \sum_{\mathbf{q} \in \mathcal{D}(\mathbf{Q} \setminus Q_1)} P(Q_1|\mathbf{q})P(\mathbf{q}) && \text{(A.7)} \\ &= \sum_{(q_2, \cdots, q_n) \in \mathcal{D}(\mathbf{Q} \setminus Q_1)} P(Q_1|Q_2 = q_2, \cdots, Q_n = q_n)P(Q_2 = q_2, \cdots, Q_n = q_n). && \text{(A.8)} \end{aligned}$$

Instead of calculating the probabilities of one variable (Q_1 in the formula above) this can be formulated to cover the joint probability of a subset of $\mathbf{Q}$. Let $\mathbf{R} = \{Q_1, \cdots, Q_m\}, m < n \in \mathbb{N}$ be a subset of $\mathbf{Q} = \{Q_1, \cdots, Q_n\}$, then

$$\begin{aligned} P(\mathbf{R}) &= \sum_{\mathbf{q} \in \mathcal{D}(\mathbf{Q} \setminus \mathbf{R})} P(R, \mathbf{q}) && \text{(A.9)} \\ &= \sum_{\mathbf{q} \in \mathcal{D}(\mathbf{Q} \setminus \mathbf{R})} P(R|\mathbf{q})P(\mathbf{q}). && \text{(A.10)} \end{aligned}$$

Formula A.10 expresses a process called *Marginalization* of the variables in $\mathbf{Q} \setminus \mathbf{R}$.

Bayes' Rule, finally, makes it possible to switch the variable with the condition in a CPD, which is essential for all inference calculations in Bayesian Networks:

$$P(Q_1|Q_2)P(Q_2) = P(Q_2|Q_1)P(Q_1). \tag{A.11}$$

A.1.2 Continuous Random Variables - Gaussian Distributions

The domain $\mathbb{X} \subseteq \mathbb{R}$ of a continuous random variable X is an interval (or collection of intervals). A probability density function (pdf) $P(X)$ assigns a nonnegative value to events in $\mathbb{X}$. The integral of $P(X)$ over the complete domain is 1. The probability of X taking on a particular $x \in \mathbb{X}$ is always zero (or infinitesimally small). Instead, the probability of X to fall within a particular region or interval equals the integral of $P(X)$ over the interval[1].

There are several well-known probability density functions or continuous probability distributions. For the purpose of this introduction I only consider the *normal*, or Gaussian distribution $\mathcal{N} : \mathbb{R} \to \mathbb{R}^+$:

$$\mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{A.12}$$

where $\mu \in \mathbb{R}$ is the *mean* of the distribution, and $\sigma^2 > 0$ the *variance*[2]. Mean and variance are called *sufficient statistics* of $\mathcal{N}$: the mean is the expected value of the distribution, the value that gets assigned the highest probability, and the variance is a measure of how concentrated the distribution is around the mean. The function $\mathcal{N}(x)$ is unimodal and symmetric around $x = \mu$, the inflection points of the curve occur 1 σ away from μ (at $x = \mu \pm \sigma$)

Joint and Conditional Gaussians

The joint (gaussian) distribution of variables $X_1, \cdots, X_n$ (X_i distributed normally with $\mathcal{N}(x_i; \mu_i, \sigma_i^2)$ for all $i \in \{1, \cdots, n\}$) is the multivariate gaussian distribution

$$\mathcal{N}(\vec{x}; \vec{\mu}, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2}(\vec{x} - \vec{\mu})^T)\Sigma^{-1}(\vec{x} - \vec{\mu})\right), \tag{A.13}$$

[1] Formally, X has density f (f being non-negative and Lebesgue-integrable) if the probability of X to fall within [a,b] is the integral of f over [a,b].

[2] σ is known as *standard deviation*.

where

$$\vec{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \prod_{i=1}^{n} \mathbb{X}_i = \mathbb{R}^n,$$

$$\vec{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} \in \mathbb{R}^n,$$

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} & \cdots & \sigma_{1,n} \\ \sigma_{1,2} & \sigma_2^2 & \cdots & \sigma_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1,n} & \sigma_{2,n} & \cdots & \sigma_n^2 \end{pmatrix} \in \mathbb{R}^{n \times n},$$

and $|\Sigma|$ the determinant of Σ. The matrix Σ is called *covariance matrix* of the distribution, and consists of the variances σ_i^2 of the distributions $\mathcal{N}_i$ and the covariances $\sigma_{i,j}$ of all pairs X_i and X_j of random variables. The covariance $\sigma_{i,j}$ measures how the variables X_i and X_j change together: A positive covariance indicates that they show similar behavior, negative covariance indicates opposite behavior. If the two variables are independent, their covariance is zero. The covariance is calculated as follows[3]:

$$\sigma_{i,j} = \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])], \tag{A.14}$$

where $\mathbb{E}[X]$ is the expected value, or mean, of the random variable X. The sample covariance of a set of observed value pairs $\{(x_i^{(1)}, x_j^{(1)}), \cdots, (x_i^{(m)}, x_j^{(m)})\}$ can be calculated via:

$$\sigma_{i,j}^2 = \frac{1}{m} \sum_{k=1}^{m} (x_i^{(k)} - \overline{\mu}(X_i))(x_j^{(k)} - \overline{\mu}(X_j)), \tag{A.15}$$

with

$$\overline{\mu}(X_i) = \frac{1}{m} \sum_{k=1}^{m} x_i^{(k)} \text{the sample mean.}$$

The normalized version of the covariance is the *correlation coefficient*, which is often used to measure the strength of linear dependence between two variables:

$$r_{i,j} = \frac{\sigma_{i,j}^2}{\sigma_i \sigma_j}. \tag{A.16}$$

[3]The calculation is valid for arbitrary random variables, which is why the general concept of expected value, or *first moment* $\mathbb{E}$ is used here.

From the joint gaussian distribution of several continuous variables, arbitrary conditional distributions, as well as the joint distributions of subsets of the variables can be calculated. Let $P(\mathbf{X}) = \mathcal{N}(\vec{x}; \vec{\mu}, \Sigma)$ be a joint gaussian distribution over the variables $\mathbf{X} = \{X_1, \cdots, X_n\}$, with $\vec{\mu}$ and Σ as defined above. Let further $\mathbf{Y_1}$ and $\mathbf{Y_2}$ be a partition of $\mathbf{X}$, such that $\mathbf{Y_1} = \{X_1, \cdots, X_k\}$, and $\mathbf{Y_2} = \{X_{k+1}, \cdots, X_n\}$, $1 \leq k < n$. Accordingly,

$$\vec{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} = \begin{pmatrix} \vec{\mu}_{\mathbf{Y}_1} \\ \vec{\mu}_{\mathbf{Y}_n} \end{pmatrix}, \text{ and}$$

$$\Sigma = \begin{pmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}^T & \mathbf{B} \end{pmatrix},$$

where $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are the following submatrices of Σ:

$$\mathbf{A} = \begin{pmatrix} \sigma_1^2 & \cdots & \sigma_{1,k} \\ \vdots & \ddots & \vdots \\ \sigma_{1,k} & \cdots & \sigma_k^2 \end{pmatrix} \in \mathbb{R}^{k \times k}$$

$$\mathbf{B} = \begin{pmatrix} \sigma_{k+1}^2 & \cdots & \sigma_{k+1,n} \\ \vdots & \ddots & \vdots \\ \sigma_{k+1,n} & \cdots & \sigma_n^2 \end{pmatrix} \in \mathbb{R}^{(n-k) \times (n-k)}$$

$$\mathbf{C} = \begin{pmatrix} \sigma_{1,k+1} & \cdots & \sigma_{1,n} \\ \vdots & \ddots & \vdots \\ \sigma_{k,k+1} & \cdots & \sigma_{k,n} \end{pmatrix} \in \mathbb{R}^{k \times (n-k)}$$

The joint distributions of the subsets $\mathbf{Y_1}$ and $\mathbf{Y_2}$ are then simply:

$$P(\mathbf{Y_1}) \propto \mathcal{N}(\vec{y_1}; \vec{\mu}_{\mathbf{Y_1}}, \mathbf{A}), \text{ and} \tag{A.17}$$

$$P(\mathbf{Y_2}) \propto \mathcal{N}(\vec{y_2}; \vec{\mu}_{\mathbf{Y_2}}, \mathbf{B}). \tag{A.18}$$

The conditional distributions have the following form:

$$P(\mathbf{Y_1} = \vec{y_1} | \mathbf{Y_2} = \vec{y_2}) \propto \mathcal{N}(\vec{y_1}; \vec{\mu}_{\mathbf{Y_1}} + \mathbf{C}\mathbf{B}^{-1}(\vec{y_2} - \vec{\mu}_{\mathbf{Y_2}}), \mathbf{A} - \mathbf{C}\mathbf{B}^{-1}\mathbf{C}^T), \tag{A.19}$$

$$P(\mathbf{Y_2} = \vec{y_2} | \mathbf{Y_1} = \vec{y_1}) \propto \mathcal{N}(\vec{y_2}; \vec{\mu}_{\mathbf{Y_2}} + \mathbf{C}^T\mathbf{A}^{-1}(\vec{y_1} - \vec{\mu}_{\mathbf{Y_1}}), \mathbf{B} - \mathbf{C}^T\mathbf{A}^{-1}\mathbf{C}). \tag{A.20}$$

Multiplying two gaussian functions (over the same set of variables $\mathbf{X} = \{X_1, \cdots, X_n\}$) results in another gaussian function:

$$\mathcal{N}(\vec{x}; \vec{\mu}_1, \Sigma_1) \cdot \mathcal{N}(\vec{x}; \vec{\mu}_2, \Sigma_2) \propto \mathcal{N}(\vec{x}; \vec{\mu}_3, \Sigma_3), \tag{A.21}$$

where

$$\Sigma_3 = (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1}$$
$$\vec{\mu}_3 = \Sigma_3\Sigma_1^{-1}\vec{\mu}_1 + \Sigma_3\Sigma_2^{-1}\vec{\mu}_2.$$

The factor z, that normalizes the resulting distribution $\mathcal{N}(\vec{\mu}_3, \Sigma_3)$ is a Gaussian itself, either in $\vec{\mu}_1$ or $\vec{\mu}_2$:

$$z = \frac{\sqrt{|\Sigma_3|}}{\sqrt{(2\pi)^n|\Sigma_1||\Sigma_2|}} \exp\left(-\frac{1}{2}(\vec{\mu}_1^T\Sigma_1^{-1}\vec{\mu}_1 + \vec{\mu}_2^T\Sigma_2^{-1}\vec{\mu}_2 + \vec{\mu}_3^T\Sigma_3^{-1}\vec{\mu}_3)\right) \tag{A.22}$$
$$z(\vec{\mu}_1) \propto \mathcal{N}(\vec{\mu}_1; (\Sigma_1^{-1}\Sigma_3\Sigma_1^{-1})^{-1}(\Sigma_1^{-1}\Sigma_3\Sigma_2^{-1})\vec{\mu}_2, \Sigma_1^{-1}\Sigma_3\Sigma_1^{-1})^{-1}) \tag{A.23}$$
$$z(\vec{\mu}_2) \propto \mathcal{N}(\vec{\mu}_2; (\Sigma_2^{-1}\Sigma_3\Sigma_2^{-1})^{-1}(\Sigma_2^{-1}\Sigma_3\Sigma_1^{-1})\vec{\mu}_1, \Sigma_2^{-1}\Sigma_3\Sigma_2^{-1})^{-1}) \tag{A.24}$$

A.2 Bayesian Networks

Let $\mathbf{Q} = \{Q_1, \cdots, Q_n\}$ be a set of discrete random variables with distributions $P(Q_i)$. Let further $G = (V, E)$ be a directed acyclic graph, $V = \{v_{Q_1}, \cdots, v_{Q_n}\}, n \in \mathbb{N}$ the set of nodes in G, $E = \{(v_{Q_i}, v_{Q_j}) | v_{Q_i} \in V, v_{Q_j} \in V, i \neq j\}$ the set of edges between nodes in G. Each node v_{Q_i} in the graph represents one of the random variables Q_i, and each edge (v_{Q_i}, v_{Q_j}) in G indicates a direct statistical dependence between the variables Q_i and Q_j represented by the two nodes. Associated with each node is the probability distribution of the corresponding variable conditioned on all parent nodes in the graph. Given, for example, $\mathbf{V} = \{v_{Q_1}, v_{Q_2}, v_{Q_3}\}$ and $\mathbf{E} = \{(v_{Q_1}, v_{Q_3}), (v_{Q_2}, v_{Q_3})\}$, associated with the parentless nodes v_{Q_1} and v_{Q_2} are the marginal distributions of Q_1 and Q_2, and associated with v_{Q_3} is the conditional distribution $P(Q_3|Q_1, Q_2)$. From here on I often use Q_i instead of v_{Q_i} to address the node associated with Q_i to simplify the notation. Accordingly, $\mathbf{Q} = \{Q_1, \cdots, Q_n\}$ can either be a set of nodes in the graph or the random variables associated with the set of nodes.

The following popular example (figure A.1, [77]) demonstrates basic principles of Bayesian Networks. The system models the state of a lawn in the summer: the grass being wet (W=True) has two possible causes: the water sprinkler (S) or a bout of rain (R). The probability of the sprinkler being turned on or a bout of rain happening depends on the sky (C): a cloudy sky raises the probability of rain, a clear sky the probability of the sprinkler being turned on.

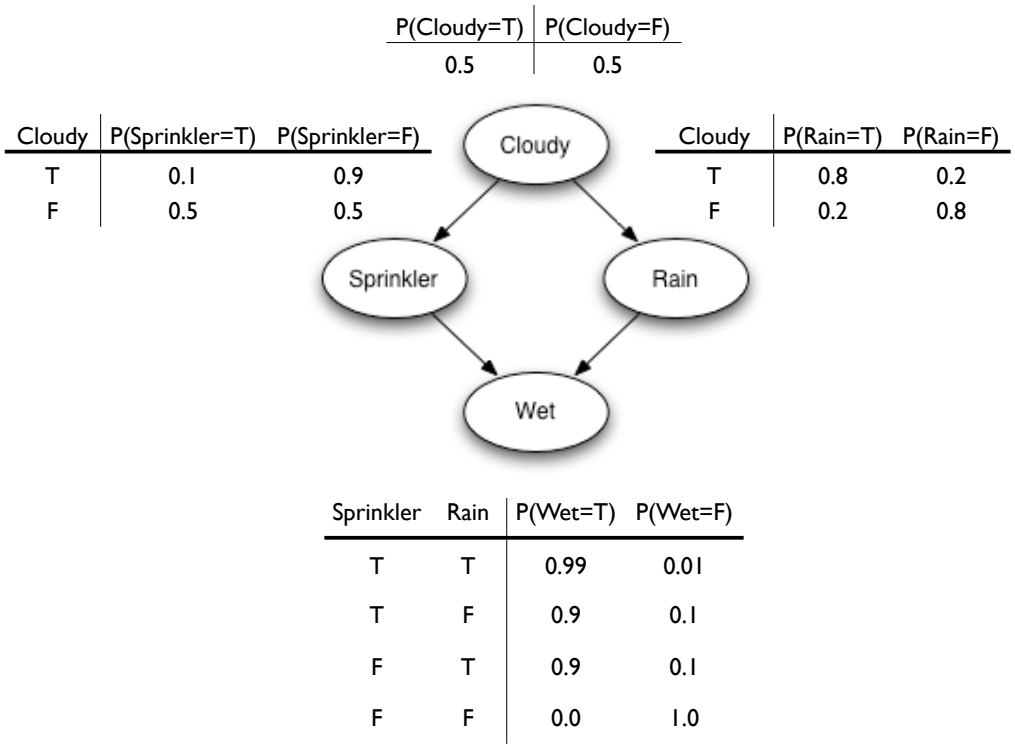


Figure A.1: The Sprinkler network, a simple example for a Bayesian Network, connecting the probability distributions of a water sprinkler (S), a bout of rain (R), the sky being cloudy (C), and the grass being wet on a lawn (W). The tables contain the conditional probability distributions of the variables given their parents.

The view of cause and effect helps in building and understanding the network. For inference, the process of calculating posterior probabilities, it is more useful to look at the conditional independences encoded by the structure of the network. In analogy to statistical independence, the following defines *conditional independence*:

> Two random variables X and Z are conditionally independent given a random variable Y (written: $X \perp\!\!\!\perp Z|Y$) iff the probability distribution of X is the same for all values of Z (and vice versa), given any value of Y.

Alternatively:

$$X \perp\!\!\!\perp Z|Y \iff P(X, Z|Y) = P(X|Y)P(Z|Y).$$

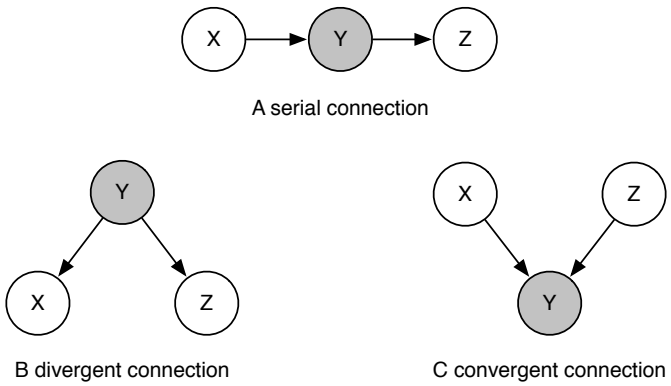


Figure A.2:

Conditioning on a variable or a set of variables can either make variables independent or induce dependencies, depending on the layout of the network. The following 3 basic situation can be distinguished:

Serial Connection: If the variable Y figure A.2 (A) is unknown, the variables X and Z are connected: The state of X influences the probability distribution of Y, which in turn influences the variable Z. If the state of Y is known, the influence of X is blocked. Hence, X and Z are conditionally independent given Y ($X \perp\!\!\!\perp Z|Y$).

Divergent Connection: In the situation depicted in figure A.2 (B), sometimes called *naïve Bayes*, the two variables X and Z again are conditionally independent given Y: A change in the conditional probabilities $P(X|Y)$ translates into a change of $P(Y|X)$ via Bayes' Rule (equation A.11), which in turn influences $P(Z|Y)$. In case the state of Y is known, the connection is blocked ($X \perp\!\!\!\perp Z|Y$).

Convergent Connection: The opposite is true in figure A.2 (C): If the state Y is unknown, changes in X only influence the probability of Y, but not of Z and vice versa. However, once the state of Y is fixed, X and Z become connected. This effect if often referred to as "explaining away": If X is confirmed as a cause of Y then the need to consider Z as an alternative or additional cause is reduced.

Applied to the sprinkler network in figure A.1 for example the following independence statements can be made:

- If we know neither the state of the lawn (W) nor of the sky (C), the probabilities of sprinkler (S) and rain (R) are connected. Knowing that it has rained makes it less likely that the sprinkler was on and vice versa. The divergent connection $S \leftarrow C \rightarrow R$ connects S and R as long as C is unknown (figure A.2 (B)). This changes as soon as we observe the sky. Knowing that the sky is clear fixes both the probabilities of sprinkler and rain, and observing one does not influence the other any more. In both cases, influence via the convergent connection $S \rightarrow W \leftarrow R$ is blocked, because W (the lawn) is unknown (figure A.2 (C)). Formally:

$$Sprinkler \perp\!\!\!\perp Rain|Sky.$$

- Knowing the state of the lawn induces a dependency between Sprinkler and Rain (figure A.2 (C), divergent connection): Knowing that a gush of rain just came down makes this the probable explanation for a wet lawn and lowers the probability that the wet lawn is also caused by the sprinkler.

In probability theory, the *chain rule of probability* permits the calculation of any member of the joint distribution of a set of random variables using conditional probabilities [104]. Given random variables $X_1, \cdots, X_n, n \in \mathbb{N}$ the joint probability $P(X_1, \cdots, X_n)$ of all n variables can be calculated as follows:

$$P(X_1, \cdots, X_n) = \prod_{i}^{n} P(X_i|X_{i-1}, \cdots, X_1) \tag{A.25}$$

For the sprinkler network that leads to the following term for the joint probability of all four variables:

$$P(C, S, R, W) = P(C) \times P(S|C) \times P(R|S, C) \times P(W|R, S, C).$$

Applying the rules for conditional independence the third fourth term can be simplified: Rain is independent of Sprinkler given the state of the sky, so $P(R|S, C) = P(R|C)$, and the state of the lawn is independent of the state of the sky given both Rain and Sprinkler, so $P(W|R, S, C) = P(W|R, S)$, hence:

$$P(C, S, R, W) = P(C) \times P(S|C) \times P(R|C) \times P(W|R, S).$$

In general, the joint probability distribution represented by a bayesian network is the

product of all the individual distributions conditioned on the parent variables.

$$\begin{aligned} P(\mathbf{Q}) &= P(q_1, \cdots, q_n) \\ &= \prod_{v_Q \in V} P(Q|\text{pa}(Q)), \end{aligned} \tag{A.26}$$

where $\text{pa}(Q)$ is the set of parent nodes of the node associated with the variable Q. Without simplification the representation of the joint probability of n binary nodes needs $O(2^n)$ factors, using the conditional independences reduces the number of factors to $O(n2^k)$, where k is the maximum number of parents a node has [76].

A.2.1 Inference in Bayesian Networks

Observing the state of a subset of variables (*evidence variables*) in a BN may change the probability distributions of other variables in the system. Calculating those *posterior distributions* is called *inference*. The most obvious way to calculate posterior distributions is using the joint probability distribution. Let $\mathbf{E}$ be a set of evidence variables $\mathbf{E} = \{E_1, \cdots, E_m\} \subset \mathbf{Q}$ that are in known states $E_1 = e_1, \cdots, E_m = e_m$, and $Q_1 \notin \mathbf{E}$ the variable the posterior probability of which we want to know ($P(Q_1|\mathbf{E})$).

$$P(Q_1|\mathbf{E}) = \frac{P(Q_1, \mathbf{E})}{P(\mathbf{E})} \tag{A.27}$$

$$= \frac{P(Q_1, E_1 = e_1, \cdots, E_m = e_m)}{P(E_1 = e_1, \cdots, E_m = e_m)} \tag{A.28}$$

Both factors can be calculated from the joint probability distribution through equation A.5 (total probability):

$$P(\mathbf{E}) = \sum_{\mathbf{q} \in \mathcal{D}(\mathbf{Q} \setminus \mathbf{E})} P(\mathbf{E}, \mathbf{q}) \tag{A.29}$$

$$P(Q_1, \mathbf{E}) = \sum_{\mathbf{q} \in \mathcal{D}(\mathbf{Q} \setminus \{Q_1, \mathbf{E}\}} P(Q_1, \mathbf{E}, \mathbf{q}). \tag{A.30}$$

For the sprinkler example, questions like "Given that the grass is wet, how likely is it that the sky was cloudy?" ($P(C = \text{true}|W = \text{true})$), or "Given that it has rained, how likely is it that the sprinkler was on?" ($P(S = \text{true}|R = \text{true})$) can be answered using the

Cloudy	Sprinkler	Rain	Wet=True	Wet=False
T	T	T	0.0396	0.0004
T	T	F	0.009	0.001
T	F	T	0.324	0.036
T	F	F	0	0.09
F	T	T	0.0495	0.0005
F	T	F	0.18	0.02
F	F	T	0.045	0.005
F	F	F	0	0.2

Figure A.3: Joint probability table for the complete sprinkler network $P(C, S, R, W)$

joint probability table of the network shown in A.3

$$P(C = \mathrm{t}|W = \mathrm{t}) = \frac{\sum_{(r,s)\in\mathcal{D}(R,S)} P(C = \mathrm{T}, W = \mathrm{T}, (r, s))}{\sum_{(r,s,c)\in\mathcal{D}(R,S,C)} P(W = \mathrm{T}, (r, s, c))} =$$
$$= 0.576$$
$$P(S = \mathrm{t}|R = \mathrm{t}) = \frac{\sum_{(c,w)\in\mathcal{D}(C,W)} P(S = \mathrm{T}, R = \mathrm{T}, (c, w))}{\sum_{(c,r,w)\in\mathcal{D}(C,R,W)} P(R = \mathrm{T}, (c, s, w)} =$$
$$= 0.18$$

This, however, requires the complete joint probability table to be calculated, which is computationally expensive. There are several more efficient ways to calculate inference, three of which will be briefly outlined in the following. A detailed introduction to all the concepts can be found in [76].

Variable Elimination is a method to solve one inference query at a time. Instead of using the product rule to calculate the joint probability (eq. A.26), new evidence is directly entered into the equation. Making use of distributivity ($a{\cdot}b{+}a{\cdot}c = a{\cdot}(b{+}c)$) the non-observed non-query variables are eliminated one by one until only the query variables remain.

Message Passing algorithms are designed to calculate exact inference on trees. Usually they are not applied directly to a Bayesian Network but to a *junction–* or

cluster tree, a modified version of the graph guaranteed to be a tree. The nodes in the junction tree are supernodes containing several variables of the original model, based on maximal cliques in the triangulated graph. New evidence is distributed to all nodes, which results in all distributions being updated to the new evidence. Inference is solved not only for one, but for all nodes simultaneously.

Sampling Methods don't calculate exact but approximate inference and are generally much more efficient than exact algorithms. New evidence is entered into the model (the probabilities of states that are impossible according to the new evidence are set to 0) and all unobserved nodes are randomly sampled repeatedly according to their (adapted) distributions. The occurrences of the different states approximate the posterior probabilities.

Variable elimination and message passing algorithms calculate inference exactly. However, the same NP-hard problem is the key to efficiency in both: To be maximally effective the order in which variables are eliminated needs to be optimal. The problem of finding the optimal order is the same as finding a way to triangulate a graph with the fewest possible edges. The construction of cluster trees also requires a triangulated graph, which has a major influence on the tree width and consequently on computational cost.

A.2.2 Types of CPDs

Representing the JPD of a Bayesian network, and the CPD of a node for that matter, is only possible if all variables follow a discrete probability distribution. Expressing conditional probabilities when continuous distributions are involved needs different representations of the CPDs in the nodes, depending on the type of the variable and the type of the parents. Table A.1, taken from [76], gives an overview of the different CDPs for the situation displayed in figure A.4.

Multinomial and *conditional multinomial* are discrete probabilities, that can be displayed by a vector ($\pi(y)$) or a matrix ($A(y,q)$) respectively. The *softmax* function is a generalization of the sigmoid function to have $|Y|$ different "outputs" instead of two (also known as multi-logistic regression). The function $\sigma(x,w,y)$ is parametrized by w, and maps a value x (the state of the continuous parent) to a multinomial distribution over the states of Y. Given a mixture of discrete and continuous parents, one parameter per state of the discrete parent Q exists.

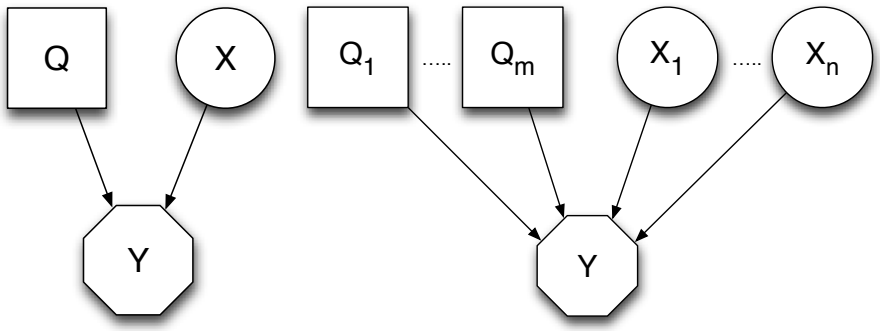


Figure A.4: *Left panel*: very simple network consisting of one discrete variable Q and one continuous variable X parenting a variable Y that is either continuous or discrete. *Right panel:* same situation with m discrete $(Q_1, \cdots, Q_m)$ and n continuous parents $X_1, \cdots, X_n$.

	Q	X	Name	CPD
Y disc.	-	-	Multinomial	$P(Y = y) = \pi(y)$
	q	-	Cond. Multinomial	$P(Y = y\|Q = q) = A(y, q)$
	-	x	Softmax	$P(Y = y\|X = x) = \sigma(x, w, y)$
	q	x	Cond. softmax	$P(Y = y\|Q = q, X = x) = \sigma(x, w_q, y)$
Y cont.	-	-	Gaussian	$P(Y = y) = \mathcal{N}(y; \mu, \sigma)$
	q	-	Cond. Gaussian	$P(Y = y\|Q = q) = \mathcal{N}(y; \mu_q, \sigma_q)$
	-	x	Linear Gaussian	$P(Y = y\|X = x) = \mathcal{N}(y; wx + \mu, \sigma)$
	q	x	Cond. lin. Gaussian	$P(Y = y\|Q = q, X = x) = \mathcal{N}(y; w_q x + \mu_q, \sigma_q)$

Table A.1: CPDs for different child/parent combinations: Y has either no discrete parent (Q= -) or one discrete parent Q, which is in state q (Q = q), and either no continuous parent (X = -) or one continuous parent that has taken the value x (X = x). The CPDs are valid for the setup displayed in the left panel of figure A.4. (Table taken from [76])

Continuously distributed variables in bayesian networks are usually modeled as Gaussians $P(Y = y) = \mathcal{N}(y; \mu, \sigma^2)$ with mean μ and variance σ^2. In the *conditional gaussian* case, a discrete variable parenting a continuous variable, μ and σ^2 depend on the values of the parenting discrete variables, and a pair of μ_q and σ_q^2 exists for each possible state of Q. In the *linear gaussian* case the mean of the child distribution depends linearly on the value x of the parenting continuous node through a parameter w. In the *conditional linear* case the weight parameter w_q, mean μ_q, and variance σ_q^2 depend on the state of the discrete parents Q.

The CPDs for cases where Y has several discrete and/or continuous parents ($\mathbf{Q} = \{Q_1, \cdots, Q_m\}$ and $\mathbf{X} = \{X_1, \cdots, X_n\}$ respectively, depicted in the right panel of figure A.4), can be generalized from the table above. Single variables (X and Q) and the states they are in, are replaced by sets of variables ($\mathbf{X}$ and $\mathbf{Q}$) and state vectors/tuples ($\overrightarrow{x} = (x_1, \cdots, x_n)$ and $\mathbf{q} = (q_1, \cdots, q_m)$, respectively). One-dimensional parameters w are replaced by $n \times n$ weight matrices W, one-dimensional means μ by n-dimensional vectors $\overrightarrow{\mu}$, and variances σ^2 by covariance matrices Σ.

A number of alternative ways of coding and parametrizing the conditional dependencies in a network exist, differing in degrees of freedom, how easy they are to interpret, and how efficiently inference can be calculated. See, for example, [76].

A.2.3 Learning in Bayesian Networks

Structure and parameters (prior distributions) of a BN can either be set by a domain expert, or learned and estimated from measured data. The theory and pertinent literature on learning in graphical networks is vast and will only be briefly outlined here. An overview of the literature and an introduction to the field can be found in [8]. Learning in BNs comes in the following four situations, in ascending difficulty:

1. Estimate parameters from a complete dataset
2. Estimate parameters from a dataset with missing values
3. Learning structure
4. Learning structure with hidden variables

For the estimation of the parameters (case 1) the two most popular approaches are *maximum likelihood estimation* (ML) and *maximum a posteriori estimation* (MAP). Let

Θ be the set of free parameters (the sufficient statistics of continuous distributions, or the different state probabilities of discrete distributions), and $\mathcal{D} = \{\mathbf{d}^1, \cdots, \mathbf{d}^N\}$ a collection of data, where each $\mathbf{d}^i = (d_1^i, \cdots, d_n^i)$ is a measured state of the network (the values of variables $X_1, \cdots, X_n$). The goal of ML estimation is to compute the set of parameters Θ^* that maximizes the (log)-likelihood of the data set, which is the probability of the model producing the measured data $\mathcal{D}$:

$$\Theta^* = \arg\max_{\Theta} P(\mathcal{D}|\Theta) = \arg\max_{\Theta} \log P(\mathcal{D}|\Theta),$$

where

$$\begin{aligned}\log P(\mathcal{D}|\Theta) &= \log \prod_{i=1}^{N} P(\mathbf{d^i}|\Theta) = \sum_{i=1}^{N} \log P(\mathbf{d^i}|\Theta) \\ &= \sum_{i=1}^{N} \sum_{j=1}^{n} P(X_j|pa(X_j), \mathbf{d^j})\end{aligned}$$

is the likelihood of the data. Hence, ML estimation boils down to estimating the parameters of each CPD given the local data. That essentially means counting occurrences of events for discrete variables, and computing sufficient statistics for continuous distributions.

Maximum a posteriori (MAP) estimation is the bayesian statistics variant of ML that includes a prior distribution on the parameters:

$$\Theta^* = \arg\max_{\Theta} \log P(\mathcal{D}|\Theta) + \log P(\Theta).$$

This has several advantages: (1) it is possible to outfit the model with prior knowledge, (2) it can prevent overfitting to the data in case the number of free parameters is much larger than the size of the dataset, and (3) it enables online learning, where instead of having the complete dataset available from the start, the dataset grows in time and the model adapts to the new data. For many types of distributions there exist prior distributions, so called *conjugate priors*, that make it possible to compute the posterior in closed form.

The case of maximum likelihood estimation in a partially observed network (case 2) – some variables being generally unobservable, or just not measured in some instances – is usually solved either through *gradient ascent* or *expectation maximization* (EM). Both return a point estimate of all parameters in the network.

The problem becomes much more difficult and computationally expensive when the structure of the network is not known and shall be learned from the data (case 3). Given a scoring function for a network structure a greedy hill climbing algorithm can be used to search a space of possible network structures. To score a network all parameters have to be estimated with ML or MAP estimation. Searching in the space of possible networks can be done, for example, by starting with a random network and changing its structure locally: adding or removing edges, or inverting their direction. If there might be hidden influences (variables) that are not in the dataset but supposed to be part of the system (case 4), the search includes adding variables, and then structurally optimizing the network.

A.3 Dynamic Bayesian Networks

The situation described above (section A.2) assumes that the variables are independent in time: the previous state of a variable does not have any influence on its current state. Dynamic Bayesian Networks (DBNs) can deal with situations that are observed over time: the states of the observed variables form a time-series, where the probability distributions over the states may not only depend on the current state of the system but also on previous events. Unlike datasets produced by a BN $\mathcal{D} = \{\mathbf{d}^1, \cdots, \mathbf{d}^N\}$, where the order of the $\mathbf{d}_i$ is not important, DBNs produce (and are learned from) sets of data sequences $\mathcal{D} = \{\mathbf{d}^1_{1:N_1}, \cdots, \mathbf{d}^k_{1:N_k}\}$, where each $\mathbf{d}^i_{1:N_i}$ is a sequence of measurements of the complete network: $\mathbf{d}^i_{1:N_i} = (\mathbf{d}^i_1, \cdots, \mathbf{d}^i_{N_i})$ with $\mathbf{d}^i_j = (d^i_{j,1}, \cdots, d^i_{j,n})$ a measured state of the network (the values of variables $X_1, \cdots, X_n$)

DBNs are represented by a BN that is duplicated over time, one instance of the network for every time step $t \in \{1, \cdots, N\}$, where N is the length of the sequence. In addition to the edges within the BN there are edges connecting the networks from one time step to the next. Normally, the model is assumed to be first-order Markov, which means that the state at time t does not depend on anything before $t-1$: $P(\mathbf{X}_t|\mathbf{X}_{1:t-1}) = P(\mathbf{X}_t|\mathbf{X}_{t-1})$, where $\mathbf{X}$ is the set of variables in the network, and $P(\mathbf{X}_t)$ is the JPD of all variables at time t.

Hidden Markov Models (HMMs) ([62, 104]) are a very simple Dynamic Bayesian Network consisting of 2 nodes: the discrete state space variable (the distribution over the states that the model can be in) and the output variable (the distribution over the possible outputs in the current state). The state space node is connected from one time-

step to the next. The goal is usually to infer the most probable sequence of (unobserved) states from the sequence of (observed) outputs, a process called *Viterbi decoding* and calculated by the *forwards-backwards algorithm*. The transition and output probabilities are learned from data sequences using *Expectation Maximization*.

A major problem with HMMs is that they are constrained to a single discrete state space variable. Casting several discrete variables into a "super-variable" distributed over the cartesian product of the domains, the domain of the super-variable has exponentially many states, and, consequently, parameters to estimate. This requires very large datasets to learn from. Also, the complexity of the forwards-backwards inference algorithm, which is exponential in the number of variables coded into the state space, becomes intractable.

Kalman Filter Models (KFMs) [104] can also be viewed as DBNs with 2 nodes and the same topology as HMMs. Both state-space and observation variables are (continuous) gaussian distributions, transition and observation functions (the way conditioning is implemented between the nodes) are linear-Gaussian (see A.2.2). Learning and inference in KFMs is conceptually very similar to HMMs. The major constraint of KFMs is that they represent a jointly gaussian distribution, which is unimodal, and therefore not appropriate for most problems. Switching KFMs introduce additional discrete variables that enable shifting between several distinct gaussian distributions.

Dynamic Bayesian Networks are a generalized view on state space models: other than the general restrictions on Bayesian Networks there are no restrictions on the state space regarding types and number of variables, or topology. This makes them a very powerful paradigm, applicable and adaptable to a variety of problems. Murphy proposed the *frontier algorithm* [76], a general message passing algorithm to calculate inference and learn dynamic bayesian networks.

Appendix B

B.1 The Magaloff Corpus

Opus		Type	Key	Tempo	Measure	Score Pages
January 16, 1989						
Op.1		Rondo	E♭ Major	Allegro	2/4	15
Op.4	Mv.1	Sonata	C Minor	Allegro maestoso	2/2	15
	Mv.2		E♭ Minor	Trio	3/4	3
	Mv.3		A♭ Major	Larghetto	5/4	3
	Mv.4		E♭ Major	Presto	2/2	18
Op.5		Rondo	F Dur	Vivace	3/4	19
Op.6	No.1	Mazurka	A Major	1/4 = 132	3/4	3
	No.2		E Major	1/4 = 63	3/4	2
	No.3		E Major	Vivace	3/4	3
	No.4		G♭ Major	Presto m. n. troppo	3/4	1
Op.7	No.1	Mazurka	B♭ Major	Vivace	3/4	2
	No.2		C Major	Vivo m. n. troppo	3/4	2
	No.3		F Minor	3/4 = 54	3/4	3
	No.4		A♭ Major	Presto ma non troppo	3/4	2
	No.5		C Major	Vivo	3/4	1
Op.9	No.1	Nocturne	B♭ Minor	Larghetto	6/4	5
	No.2		E♭ Major	Andante	12/4	3
	No.3		B Major	Allegretto	6/8	9
Op.10	No.1	Étude	C Major	Allegro	4/4	5
	No.2		A Minor	Allegro	4/4	4
	No.3		E Major	Lento ma non troppo	2/4	4

	No.4		C♯ Minor	Presto	4/4	6
	No.5		G♭ Major	Vivace	2/4	4
	No.6		E♭ Minor	Andante	6/8	3
	No.7		C Major	Vivace	6/8	3
	No.8		F Major	Allegro	4/4	6
	No.9		F Minor	Allegro molto agitato	6/8	4
	No.10		A♭ Major	Vivace assai	12/8	5
	No.11		E♭ Major	Allegretto	3/4	3
	No.12		C Minor	Allegro con fuoco	2/2	5
January 19, 1989						
Op.12		Introduction	B♭ Dur	Allegro maestoso	4/4	2
		Variations	B♭ Major	Allegro moderato	6/8	11
Op.15	No.1	Nocturne	F Major	Andante cantabile	3/4	4
	No.2		F♯ Major	Larghetto	2/4	4
	No.3		G Minor	Lento	3/4	4
Op.16		Rondo	E♭ Dur	Andante	4/4	22
Op.17	No.1	Mazurka	B♭ Major	Vivo e risoluto	3/4	2
	No.2		E Minor	Lento ma non troppo	3/4	2
	No.3		A♭ Major	Legato assai	3/4	2
	No.4		A Minor	Lento ma non troppo	3/4	4
Op.18		Waltz	E♭ Dur	Vivo	3/4	9
Op.19		Bolero	C Dur	Allegro molto	3/8	13
Op.20		Scherzo	B Minor	Presto con fuoco	3/4	18
Op.23		Ballade	G Minor	Moderato	6/8	15
Op.24	No.1	Mazurka	G Minor	Lento	3/4	2
	No.2		C Major	Allegro non troppo	3/4	4
	No.3		A♭ Major	Moderato	3/4	2
	No.4		B♭ Minor	Moderato	3/4	4
Op.25	No.1	Étude	A♭ Major	Allegro sostenuto	4/4	5
	No.2		F Minor	Presto	2/2	4
	No.3		F Major	Allegro	3/4	4
	No.4		A Minor	Agitato	2/2	3
	No.5		E Minor	Vivace	3/4	6
	No.6		G♯ Minor	Allegro	2/2	6

	No.7		C♯ Minor	Lento	3/4	4
	No.8		D♭ Major	Vivace	2/2	3
	No.9		G♭ Major	Allegro assai	2/4	2
	No.10		B Minor	Allegro con fuoco	2/2	6
	No.11		A Minor	Allegro con brio	2/2	8
	No.12		C Minor	Allegro molto con fuoco	2/2	6
March 15, 1989						
Op.26	No.1	Polonaise	C♯ Minor	Allegro appassionato	3/4	6
	No.2		E♭ Minor	Maestoso	3/4	11
Op.27	No.1	Nocturne	C♯ Minor	Larghetto	4/4	6
	No.2		D♭ Major	Lento sostenuto	6/8	6
Op.28	No.1	Prélude	C Major	Agitato	2/8	1
	No.2		A Minor	Lento	2/2	1
	No.3		G Major	Vivace	2/2	2
	No.4		E Minor	Largo	2/2	1
	No.5		D Major	Allegro molto	3/8	1
	No.6		B Minor	Lento assai	3/4	1
	No.7		A Major	Andantino	3/4	1
	No.8		F♯ Minor	Molto agitato	4/4	3
	No.9		E Major	Largo	4/4	1
	No.10		C♯ Minor	Allegro molto	3/4	1
	No.11		B Major	Vivace	6/8	1
	No.12		G♯ Minor	Presto	3/4	3
	No.13		F♯ Major	Lento	6/4	2
	No.14		E♭ Minor	Allegro	2/2	1
	No.15		D♭ Major	Sostenuto	4/4	3
	No.16		B♭ Minor	Presto con fuoco	2/2	4
	No.17		A♭ Major	Allegretto	6/8	4
	No.18		F Minor	Allegro molto	2/2	2
	No.19		E♭ Major	Vivace	3/4	3
	No.20		C Minor	Largo	4/4	1
	No.21		B♭ Major	Cantabile	3/4	3
	No.22		G Minor	Molto agitato	6/8	2
	No.23		F Major	Moderato	4/4	2

	No.24		D Minor	Allegro appassionato	6/8	4
Op.29		Impromptus	A♭ Major	Allegro assai quasi presto	2/2	6
Op.30	No.1	Mazurka	C Minor	Allegro non tanto	3/4	2
	No.2		D Major	Vivace	3/4	2
	No.3		D♭ Major	Allegro non troppo	3/4	3
	No.4		C♯ Minor	Allegretto	3/4	5
Op.31		Scherzo	B♭ Minor	Presto	3/4	23
April 10, 1989						
Op.32	No.1	Nocturne	B Major	Andante sostenuto	4/4	4
	No.2		A♭ Major	Lento	4/4	6
Op.33	No.1	Mazurka	G♯ Minor	Mesto	3/4	2
	No.2		D Major	Vivace	3/4	4
	No.3		C Major	Simplice	3/4	2
	No.4		B Minor	-	3/4	6
Op.34	No.1	Waltz	A♭ Major	Vivace	3/4	9
	No.2		A Minor	Lento	3/4	5
	No.3		F Major	Vivace	3/4	4
Op.35	Mv.1	Sonata	B♭ Minor	Grave	2/2	10
	Mv.2		E♭ Minor	Scherzo	3/4	8
	Mv.3		B♭ Minor	Marche funébre	4/4	3
	Mv.4		B♭ Minor	Presto	2/2	4
Op.36		Impromptus	F♯ Major	Andantino	4/4	8
Op.37	No.1	Nocturne	G Minor	Andante sostenuto	4/4	4
	No.2		G Major	Andantino	6/8	6
Op.38		Ballade	F Major	Andantino	6/8	10
Op.39		Scherzo	C♯ Minor	Presto con fuoco	3/4	16
Op.40	No.1	Polonaise	A Major	Allegro con brio	3/4	8
	No.2		C Minor	Allegro maestoso	3/4	8
Op.41	No.1	Mazurka	E Minor	Andantino	3/4	2
	No.2		B Major	Animato	3/4	2
	No.3		A♭ Major	Allegretto	3/4	2
	No.4		C♯ Minor	Maestoso	3/4	4
Op.42		Waltz	A♭ Major	leggiero	3/4	10
Op.43		Tarentella	A♭ Major	Presto	6/8	10

April 13, 1989						
Op.44		Polonaise	F♯ Minor	-	3/4	20
Op.45		Prélude	C♯ Minor	Sostenuto	2/2	4
Op.46		Allegro de Concert	A Major	Allegro Maestoso	2/2	22
Op.47		Ballade	A♭ Major	Allegretto	6/8	13
Op.48	No.1	Nocturne	C Minor	Lento	4/4	6
	No.2		F♯ Minor	Andantino	4/4	6
Op.49		Fantasie	F Minor	Grave	4/4	20
Op.50	No.1	Mazurka	G Major	Vivace	3/4	4
	No.2		A♭ Major	Allegretto	3/4	4
	No.3		C♯ Minor	Moderato	3/4	6
Op.51		Impromptus	G♭ Major	Tempo giusto	12/8	8
Op.52		Ballade	F Minor	Andante con moto	6/8	17
Op.53		Polonaise	A♭ Major	Maestoso	3/4	14
Op.54		Scherzo	E Major	Presto	3/4	23
May 17, 1989						
Op.55	No.1	Nocturne	F Minor	Andante	4/4	4
	No.2		E♭ Major	Lento sostenuto	12/8	4
Op.56	No.1	Mazurka	B Major	Allegro non tanto	3/4	6
	No.2		C Major	Vivace	3/4	2
	No.3		E♭ Major	Moderato	3/4	6
Op.57		Berceuse	D♭ Major	Andante	6/8	6
Op.58	Mv.1	Sonata	B Minor	Allegro maestoso	4/4	15
	Mv.2		E♭ Major	Scherzo (molto vivace)	3/4	5
	Mv.3		B Major	Largo	4/4	6
	Mv.4		B Minor	Finale (Presto non tanto)	6/8	15
Op.59	No.1	Mazurka	A Minor	Moderato	3/4	4
	No.2		A♭ Major	Allegretto	3/4	3
	No.3		A Major	Vivace	3/4	5
Op.60		Barcarolle	F♯ Major	Allegretto	12/8	12
Op.61		Polonaise	A♭ Major	Allegro Maestoso	3/4	18
Op.62	No.1	Nocturne	B Major	Andante	4/4	6

	No.2		E Major	Lento	4/4	5
Op.63	No.1	Mazurka	B Major	Vivace	3/4	4
	No.2		F Minor	Lento	3/4	2
	No.3		E Major	Allegretto	3/4	2
Op.64	No.1	Waltz	D♭ Major	Molto vivace	3/4	4
	No.2	Waltz	C♯ Minor	Tempo giusto	3/4	6
	No.3	Waltz	A♭ Major	Moderato	3/4	6

Appendix C

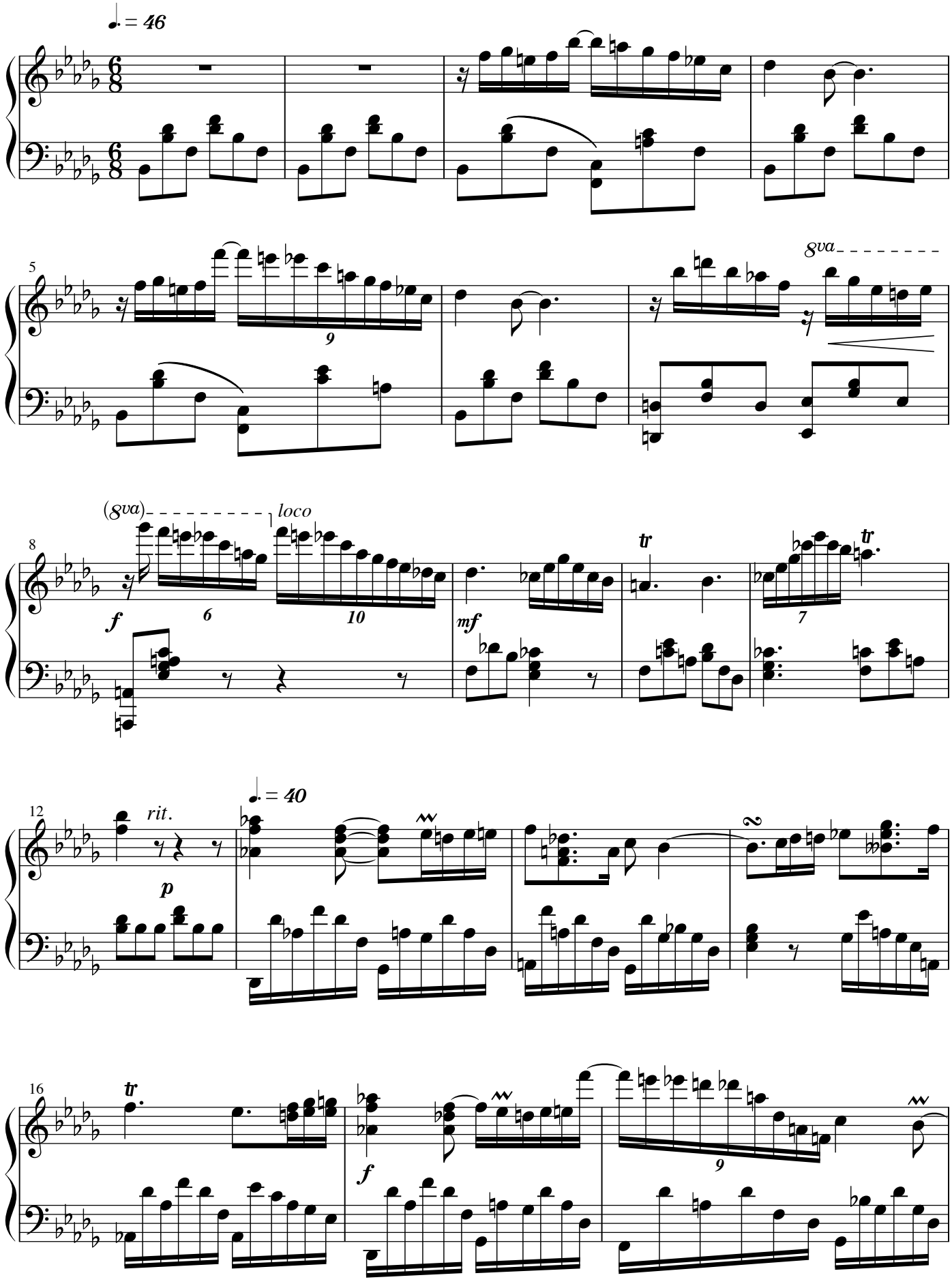
RECON 2008 & 2011 – Pieces and Awards

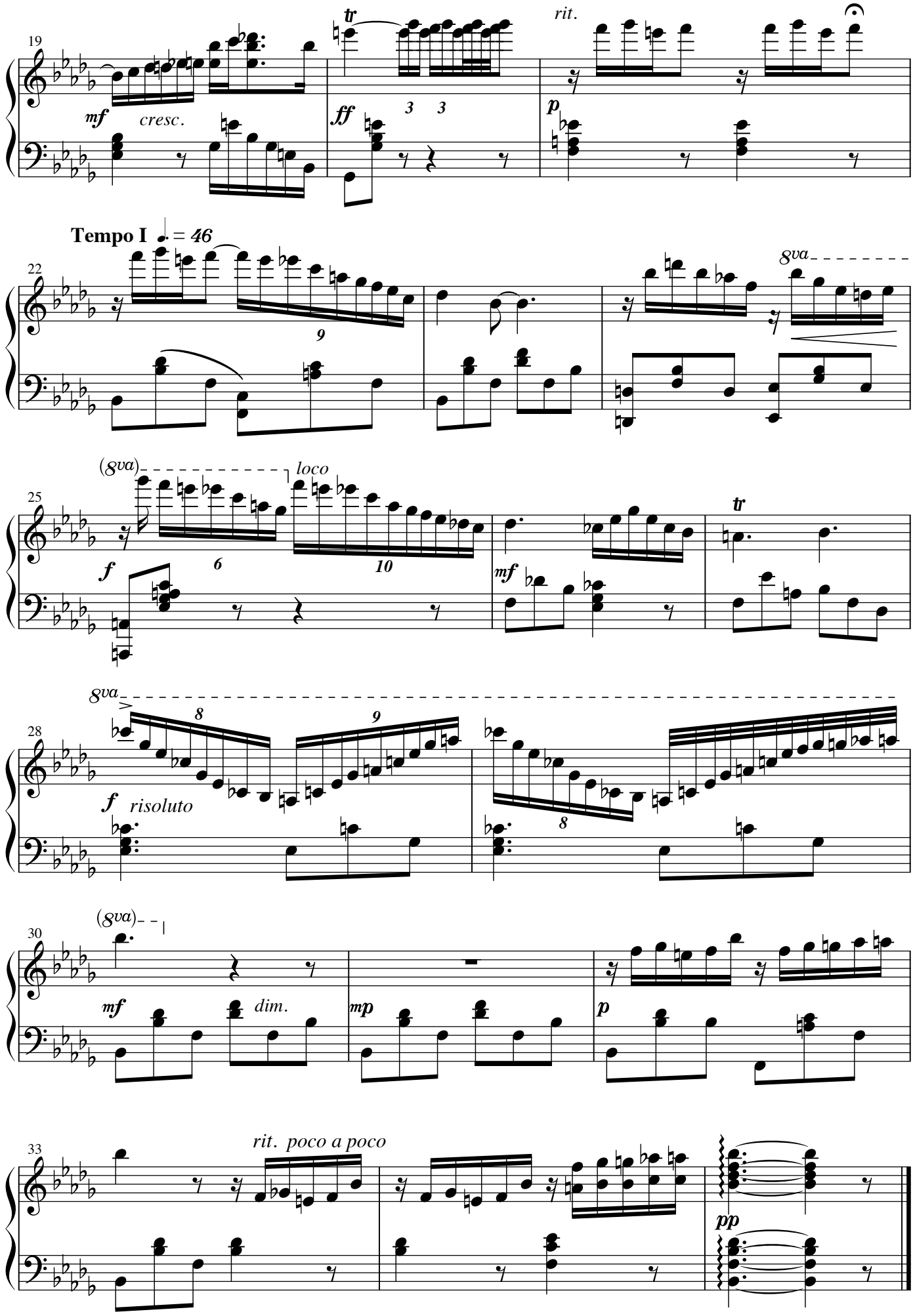
C.1 Rencon Set Pieces 2008

ICMPC-Rencon Set pieces I

My Nocturne

Tadahiro Murao

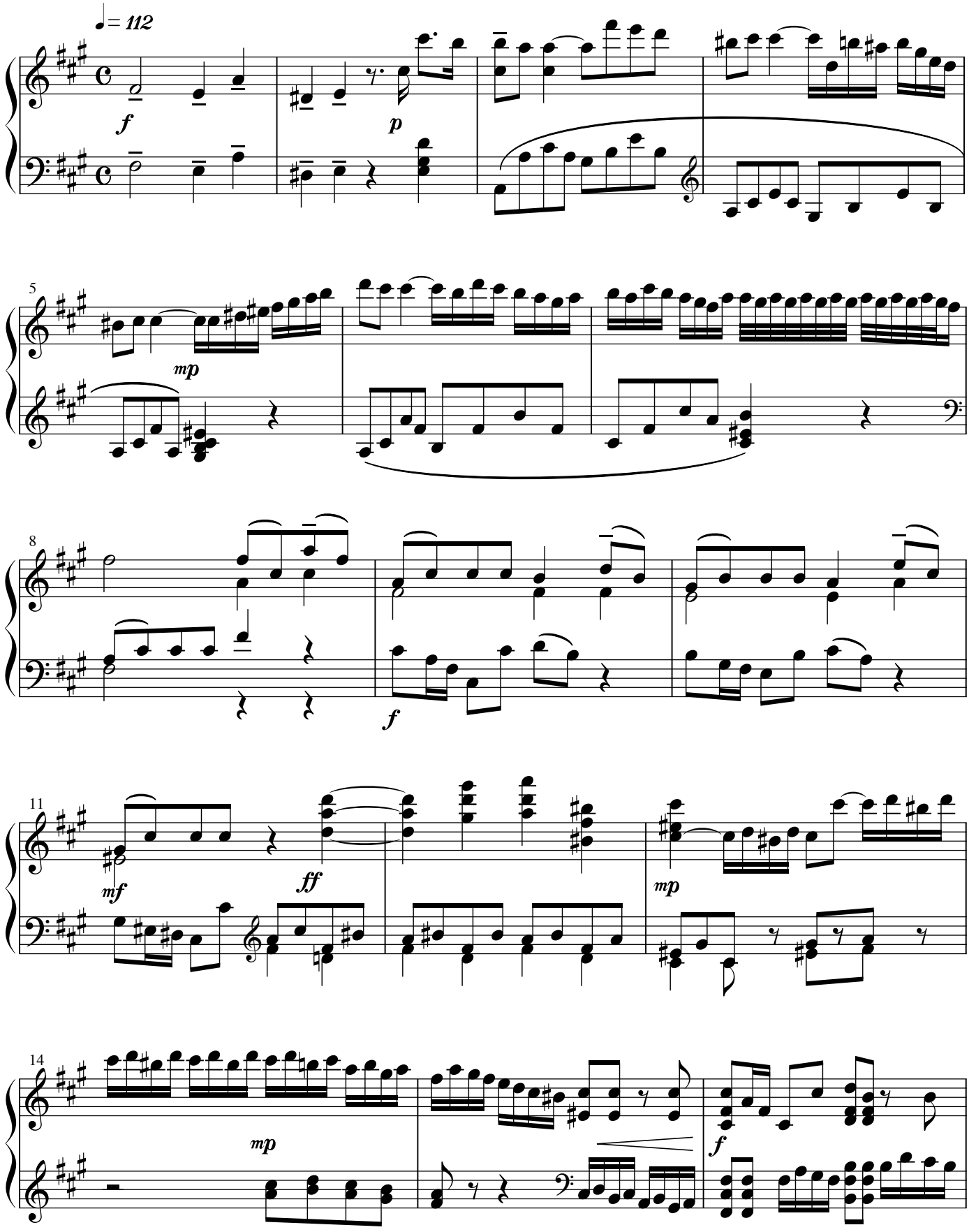


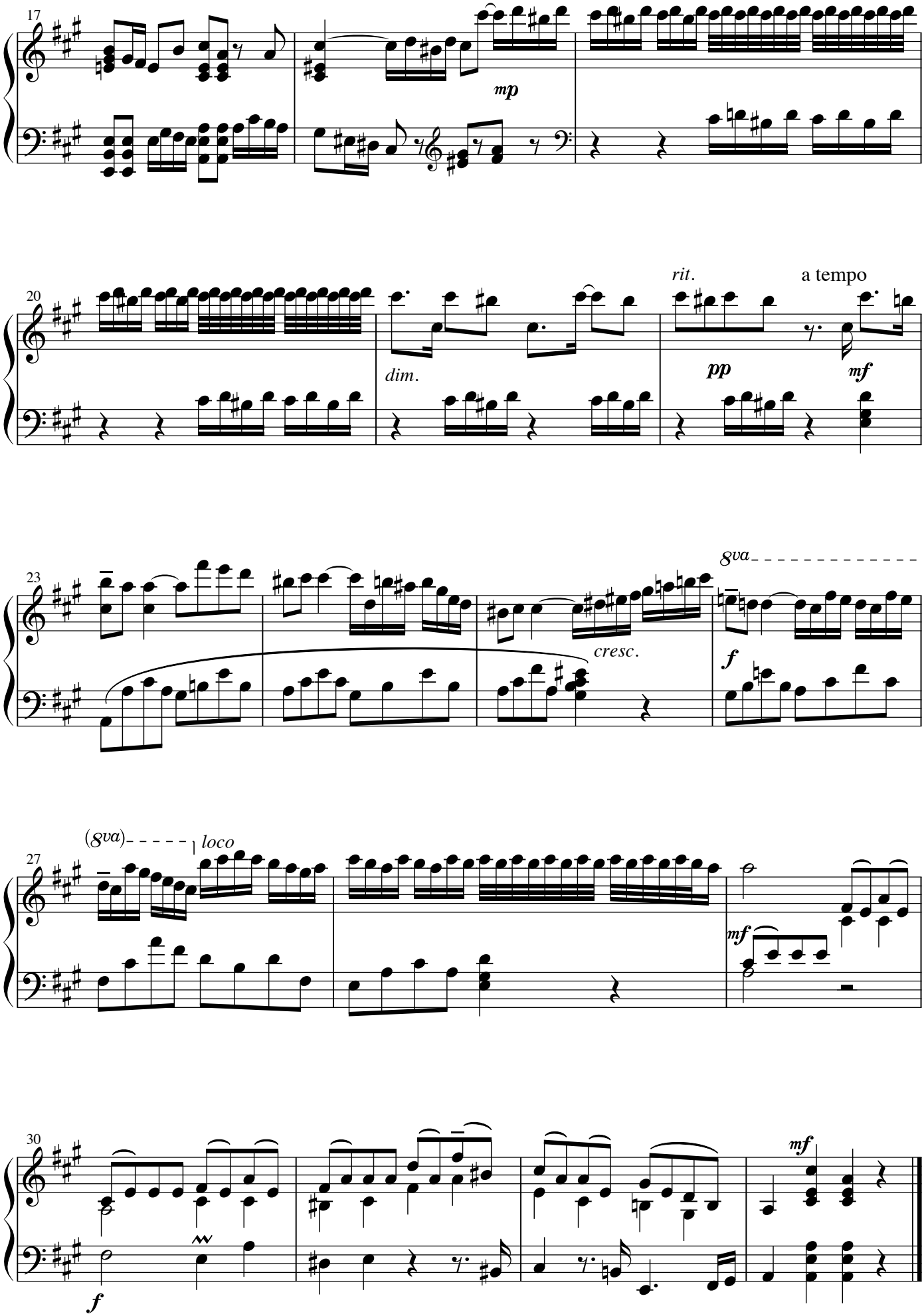

19
tr
rit.
mf
cresc.
ff
3 3
p
Tempo I ♩. = 46
22
9
8va
(8va)
loco
25
6
10
f
mf
tr
8va
28
8
9
f
risoluto
8
(8va)
30
mf
dim.
mp
p
33
rit. poco a poco
pp

ICMPC-Rencon Set pieces II

My Mozart in Sentiment

Tadahiro Murao




17
mp
20
rit.
a tempo
dim.
pp
mf
23
8va
cresc.
f
(8va)
loco
27
mf
30
mf
f

C.2 Rencon Awards 2008

Rencon

ICMPC10

Rencon Award

Rencon Steering Committee is pleased to honor

Maarten Grachten, Sebastian Flossmann & Gerhard Widmer

for the most splendid performance rendered at the autonomous section of ICMPC10 Rencon Workshop on methods for automatic music performance and their applications in a public rendering contest.

This is to certify that the performance rendered by

YQX

on August 27th, 2008 in Sapporo, JAPAN, playing

T. Murao: My Mozart in Sentiment
T. Murao: My Nocturne

has been selected as the most splended and appreciated.

August 27th, 2008.

橋田光代 蓮根

Mitsuyo Hashida

Chairperson of ICMPC10 Rencon Workshop
on methods for automatic music performance
and their applications in a public rendering contest

Rencon

ICMPC10

Rencon Technical Award

Rencon Steering Committee is pleased to honor

Maarten Grachten, Sebastian Flossmann & Gerhard Widmer

for the most splendid performance rendered at the autonomous section of ICMPC10 Rencon Workshop on methods for automatic music performance and their applications in a public rendering contest.

This is to certify that the performance rendered by

YQX

on August 27th, 2008 in Sapporo, JAPAN, playing

T. Murao: My Mozart in Sentiment
T. Murao: My Nocturne

has been selected as the most splended and appreciated.

平田圭二
Keiji Hirata

August 27th, 2008.

橋田光代 蓮根
Mitsuyo Hashida
Chairperson of ICMPC10 Rencon Workshop
on methods for automatic music performance
and their applications in a public rendering contest

Rencon

ICMPC10

Rencon Murao Award

Rencon Steering Committee is pleased to honor

Maarten Grachten, Sebastian Flossmann & Gerhard Widmer

for the most splendid performance rendered at the autonomous section of ICMPC10 Rencon Workshop on methods for automatic music performance and their applications in a public rendering contest.

This is to certify that the performance rendered by

YQX

on August 27th, 2008 in Sapporo, JAPAN, playing

T. Murao: My Mozart in Sentiment
T. Murao: My Nocturne

has been selected as the most splended and appreciated by the composer of the set pieces.

August 27th, 2008.

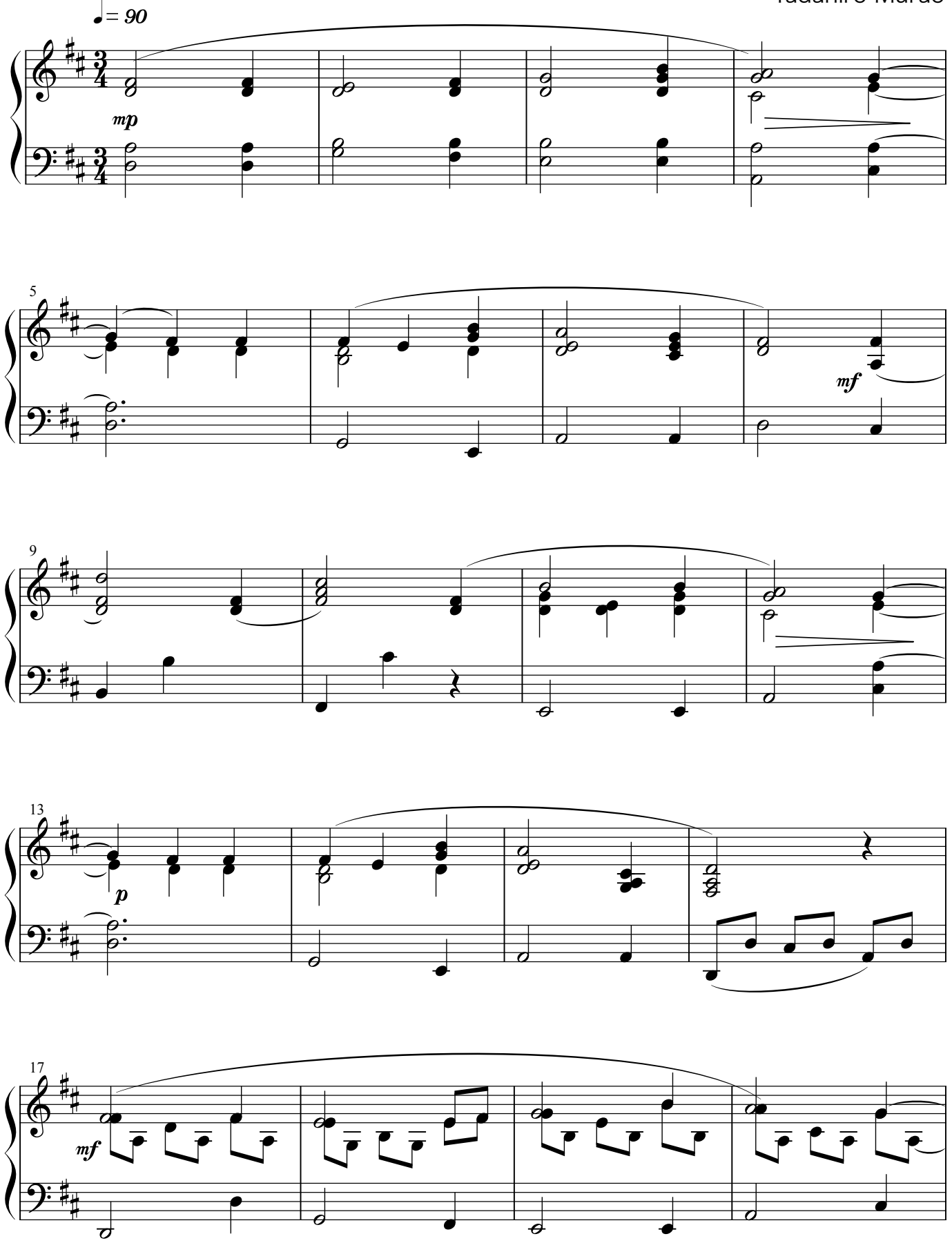
村尾忠廣
Tadahiro Murao
The contributor of the set pieces

橋田光代 蓮根
Mitsuyo Hashida
Chairperson of ICMPC10 Rencon Workshop on methods for automatic music performance and their applications in a public rendering contest

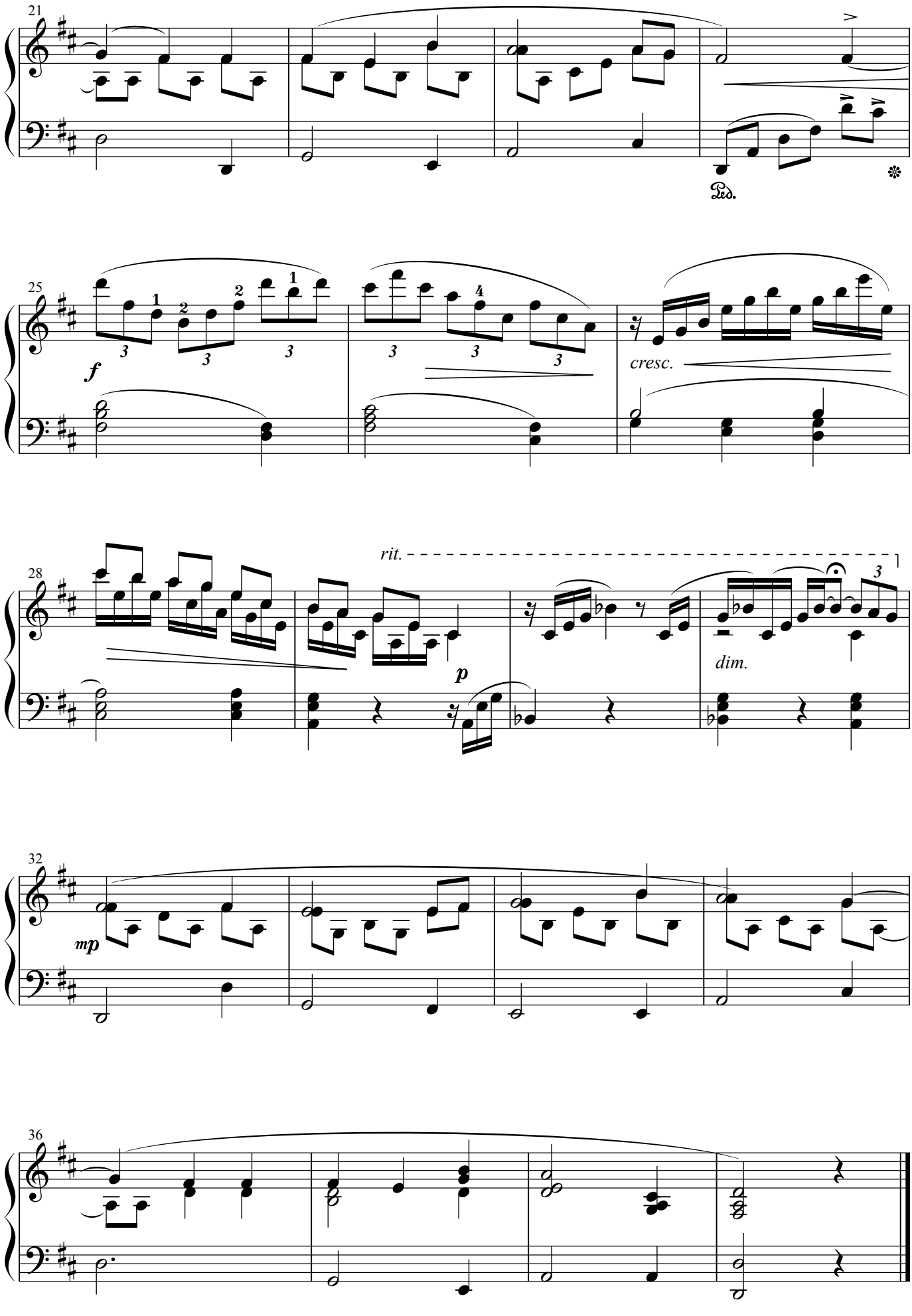
C.3 Rencon Set Pieces 2011

A Little Consolation

Tadahiro Murao



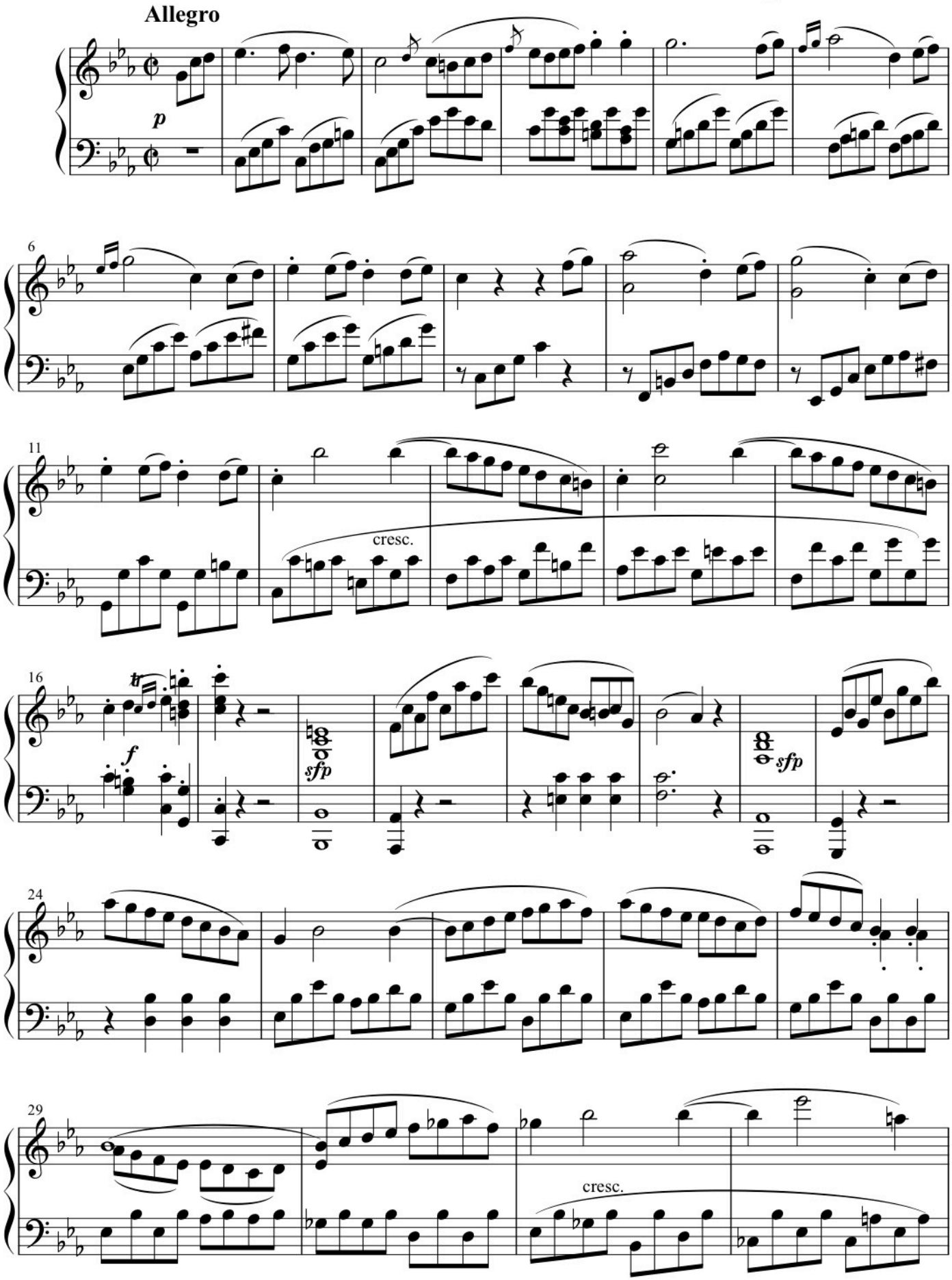
Set piece of Rencon Workshop 2011 (SMC-Rencon)

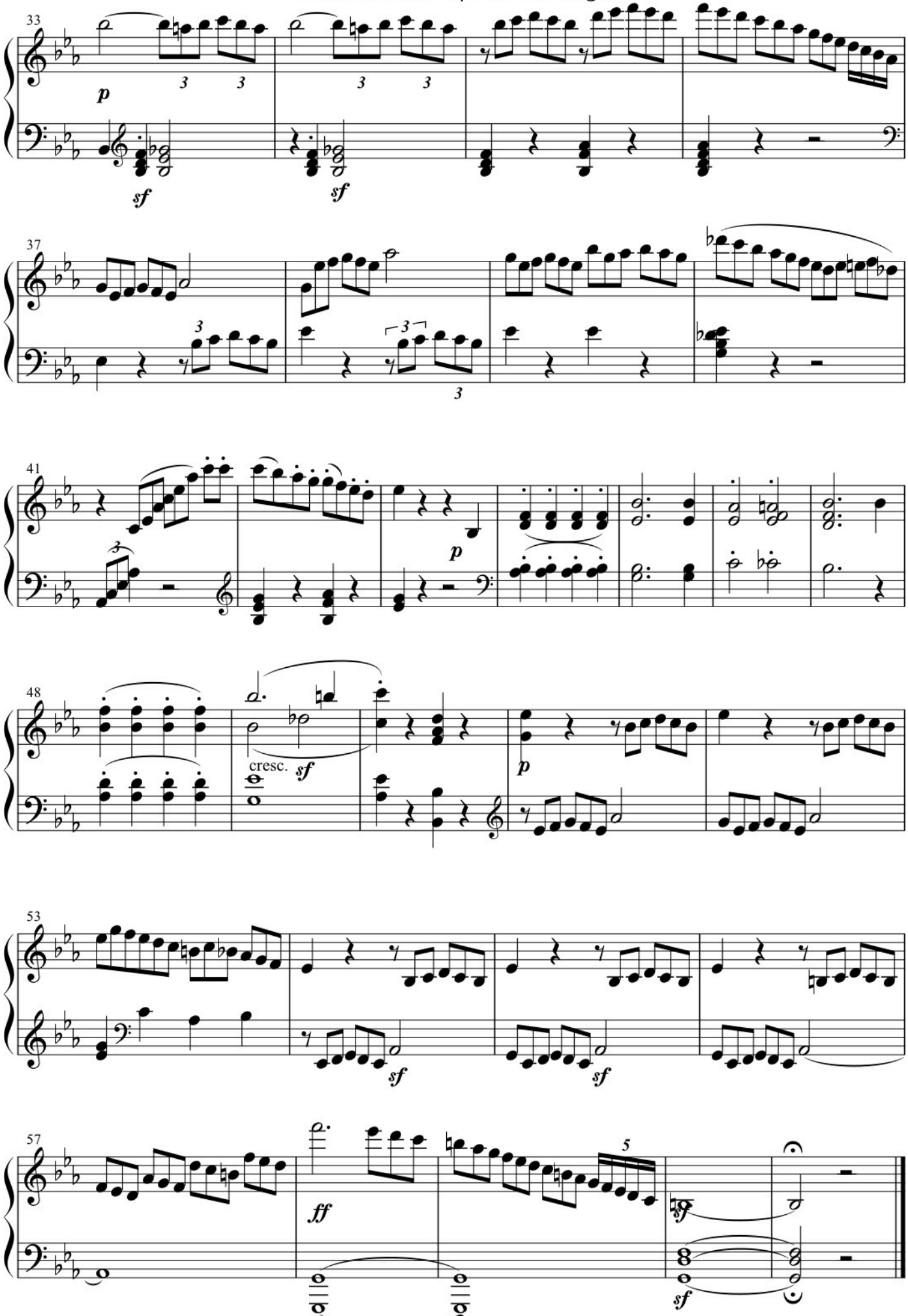

21
Ped.
25
f
cresc.
28
rit.
p
dim.
32
mp
36

No.5

Sonata No. 8 Op. 13 III. Allegro

Ludwig van Beethoven




33
37
41
48
53
57
cresc.

C.4 Rencon Awards 2011

Rencon

SMC-Rencon Award

Rencon Steering Committee is pleased to honor

Sebastian Flossmann, Maarten Grachten & Gerhard Widmer

for the most splendid performance rendered by their system at SMC2011 Rencon Workshop on methods for automatic music performance and their applications in a public rendering contest.

smc 2011

This is to certify that the performance rendered by

YQX ,

on July 6rh, 2011 in Padova, playing

Beethoven's Sonata. No8. 3rd Mov.

has been selected as the most splended and appreciated.

July 6th, 2011

橋田光代 蓮根

Chairperson of SMC-Rencon

Appendix D

Tables

Op.10 No.1		Op.10 No.2		Op.10 No.4		Op.10 No.5		Op.10 No.7		Op.10 No.8	
BI49	157	BI49	129	HA29	157	SH32	104	BI49	232	BI49	142
HA29	159	**MA77**	139	BI49	157	**MA63**	111	**MA63**	237	HA29	157
SH32	163	SG32	140	AR53	161	LO27	115	HA29	242	SH32	157
CO56	164	HEN	144	SC31	165	**MA77**	115	SC31	243	**MA63**	159
MA63	165	HA29	145	**MA63**	166	AS38	115	**MA77**	244	BA44	168
SC31	169	**MA63**	145	SH32	169	HEN	116	SH32	248	SC31	173
AS38	170	CO56	149	LO27	169	BI49	117	HEN	252	LO27	174
MA77	170	AR53	152	PO30	169	SC31	117	AR53	252	GI33	174
HEN	176	SC31	152	**MA77**	170	Gi33	118	GA30	254	**MA77**	174
PO30	178	PO30	152	GI33	174	CO56	120	LU27	256	HEN	176
LO27	179	LO27	156	AS38	174	LU27	120	LO27	256	AS38	177
BA44	179	AS38	157	CO56	175	AR53	121	CO56	263	CO56	178
LU27	180	LU27	159	HEN	176	HA29	122	AS38	264	AR53	179
GA30	190	GI33	165	LU27	179	PO30	123	PO30	266	PO30	180
GI33	191	GA30	173	BA44	191	GA30	131	GI33	271	GA30	188
AR53	196	BA44	176	GA30	197	BA44	139	BA44	285	LU27	190

Table D.1: Tempo values for selected Chopin Etudes. Entries are the first to letters of a pianists name followed by their age at the time of recording. Columns are sorted by ascending tempo.

Op.10 No.10		Op.10 No.12		Op.25 No.1		Op.25 No.2		Op.25 No.4		Op.25 No.5	
BI49	426	PO30	64	HA29	77	AS38	102	AR53	65	**MA63**	168-157-179
BA44	450	LO27	64	AS38	84	HA29	103	HEN	80	HEN	184-168-184
MA63	467	**MA63**	65	LO27	91	LO27	104	BA44	84	HA29	189-108-190
SC31	471	SC31	66	LU27	93	**MA77**	106	**MA77**	85	**MA77**	190-178-184
HEN	480	LU27	66	SO35	94	AR53	109	BI49	87	GI33	190-156-204
SH32	480	AS38	66	GA30	102	**MA63**	111	**MA63**	87	LO27	191-109-217
AR53	483	HA29	68	**MA63**	102	HEN	112	PO30	88	AR53	198-116-180
LU27	487	BA44	71	BI49	103	GA30	113	CO57	89	GA30	198-144-210
HA29	505	SH32	71	HEN	104	LU27	116	HA29	92	LU27	198-150-185
GA30	508	**MA77**	72	**MA77**	104	SO35	118	GI33	93	BI49	198-150-185
AS38	512	BI49	74	AR53	104	GI33	122	GA30	95	PO30	210-160-210
PO30	513	CO56	75	Gi33	105	BI49	123	SP35	100	AS38	210-112-226
LO27	529	HEN	76	BA44	109	PO30	125	LU27	100	SO35	211-125-195
CO56	542	GI33	77	PO30	111	CO57	128	LO27	102	BA44	218-173-203
MA77	550	GA30	87	CO57	118	BA44	138	AS38	106	CO57	243-168-242
GI33	574	AR53	88								

Table D.2: Tempo values for selected Chopin Etudes. Entries are the first to letters of a pianists name followed by their age at the time of recording. Columns are sorted by ascending tempo. (Ctd.)

Op.25 No.6		Op.25 No.8		Op.25 No.9		Op.25 No.10		Op.25 No.11		Op.25 No.12	
HEN	69	BI49	64	BI49	94	**MA77**	64-90-65	HA29	51	HA29	58
MA63	70	HA29	66	HA29	104	BI49	64-106-68	BI49	53	**MA77**	62
BI49	71	HEN	69	AR53	107	LO27	67-86-70	**MA63**	58	**MA63**	69
AR53	71	GA30	69	**MA77**	107	BA44	71-112-70	GI33	59	AS38	70
CO57	73	**MA63**	69	LU27	107	AR53	71-96-68	**MA77**	60	LO27	73
PO30	74	AR53	70	CO57	110	AS38	71-84-70	LO27	61	CO57	73
BA44	74	LO27	71	HEN	112	**MA63**	71-110-70	CO57	61	BI49	74
MA77	75	**MA77**	71	PO30	113	CO57	71-127-71	AS38	62	GI33	74
AS38	75	CO57	73	**MA63**	115	HEN	72-126-72	LU27	63	SO35	76
HA29	75	GI33	73	GI33	117	PO30	72-104-74	PO30	63	LU27	76
LO27	77	AS38	73	LO27	118	GI33	74-129-73	AR53	63	PO30	76
GI33	78	PO30	76	GA30	120	HA29	74-112-76	SO35	66	AR53	77
LU27	83	LU27	77	AS38	125	LU27	75-96-71	HEN	69	HEN	80
GA30	84	BA44	78	SO35	125	SO35	83-86-87	BA44	69	BA44	82
SO35	85	SO35	81	BA44	131	GA30	86-117-81	GA30	71	GA30	83

Table D.3: Tempo values for selected Chopin Etudes. Entries are the first to letters of a pianists name followed by their age at the time of recording. Columns are sorted by ascending tempo.

	ART			IOI			VEL		
	YQX	YQX/L	YQX/G	YQX	YQX/L	YQX/G	YQX	YQX/L	YQX/G
CHP	IR-A	PI	PI	IR-A	IR-A	PI	MinP	MinP	MinP
	PI	PI-G	PI-G	CD	PI-G	PI-G	AvMax	AvMax	AvMax
	PI-G	RC	CD	MS	LC	MS	DR	DR	AvMin
	RC	DR	MS	RC	RC	RC			MS
				DR					
	0.328	0.320	0.326	0.221	0.216	0.222	0.176	0.163	0.183
MOZ/F	IR-A	IR-A		IR-A			IR-A		
	IR-L	IR-L		IR-L	IR-L	IR-L	IR-L	IR-L	IR-L
	PI	PI	PI	PI	PI	PI	PI	PI	PI
	PI-G	PI-G	PI-G	PI-G	PI-G	PI-G	PI-G	PI-G	PI-G
		CD			CD		CD		
	LC	LC					LC		
							MaxP		
							MinP	MinP	MinP
	AvMax	AvMax		AvMax	AvMax	AvMax		AvMax	
	AvMin	AvMin		AvMin			AvMin	AvMin	
	MS	MS	MS	MS	MS	MS	MS	MS	MS
	RC	RC		RC	RC	RC	RC	RC	RC
	DR	DR		DR	DR		DR	DR	
	0.411	0.414	0.361	0.389	0.380	0.340	0.395	0.393	0.353
MOZ/S		IR-A		IR-A	IR-A		IR-A	IR-A	IR-A
	IR-L	IR-L					IR-L	IR-L	
	PI	PI				PI			
	PI-G	PI-G				PI-G		PI-G	PI-G
		CD		CD	CD		CD	CD	
		LC		LC	LC		LC	LC	
									MinP
		AvMax							AvMax
		AvMin							
	MS	MS	MS	MS	MS	MS			
	RC	RC	RC	RC	RC	RC	RC	RC	
	DR	DR							
	0.277	0.256	0.257	0.409	0.409	0.442	0.250	0.235	0.235

Table D.4: Results of Feature Selection for Simple YQX Prediction on three different data sets

	TEM			TIM			VEL-D		
	YQX	YQX/L	YQX/G	YQX	YQX/L	YQX/G	YQX	YQX/L	YQX/G
CHP	IR-A	IR-A	IR-L	IR-A	IR-A	PI-G	AvMax	IR-A	AvMax
	RC	CD	MS	CD	PI	MS	MinP	LC	AvMin
	DR	MS	RC	MS	PI-G	RC	MS	MS	MS
		RC		RC	RC			RC	RC
		DR		DR	DR				
	0.193	0.150	0.211	0.150	0.130	0.158	0.210	0.182	0.202
MOZ/F	IR-A	IR-A		IR-A	IR-A				
		IR-L	IR-L	IR-L	IR-L				
		PI		PI	PI	PI			
		PI-G	PI-G	PI-G	PI-G	PI-G			
	CD			CD	CD				
		LC		LC	LC				
		MaxP							
		MinP	MinP						
				AvMax	AvMax	AvMax			
				AvMin	AvMin				
	MS	MS	MS	MS	MS	MS			
	RC	RC	RC	RC	RC	RC			
		DR		DR	DR				
	0.268	0.381	0.399	0.368	0.360	0.289			
MOZ/S	IR-A				IR-A				
				IR-L	IR-L				
			PI			PI			
			PI-G			PI-G			
	CD								
	LC	LC							
		MinP	MinP						
			AvMin						
			MS	MS	MS	MS			
	RC		RC	RC	RC	RC			
				DR	DR				
	0.181	0.210	0.264	0.412	0.408	0.426			

Table D.5: Results of Feature Selection for Simple YQX Prediction on three different data sets

Curriculum Vitae

I was born in 1980 in Regensburg, Germany and graduated in Computer Science from the University of Passau in 2005. After one year of Musicology at the University of Regensburg in 2006, I started a 6 month internship at Gerhard Widmer's *Intelligent Music Processing and Machine Learning* Group at the Austrian Research Institute for Artificial Intelligence (OFAI) in Vienna. In late 2007, I began working as a PhD student at the Department of Computational Perception at Johannes Kepler University, Linz. I have been taking piano lessons since the age of 5, and classical music has been my constant companion and source of inspiration since then.

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Dissertation selbstständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt bzw. die wörtlich oder sinngemäß entnommenen Stellen als solche kenntlich gemacht habe.
Die vorliegende Dissertation ist mit dem elektronisch übermittelten Textdokument identisch.

Linz, am ....