Computational Modelling and Quantitative Analysis of Dynamics in Performed Music

Katerina Kosta

PhD thesis

School of Electronic Engineering and Computer Science
Queen Mary University of London

2017

Abstract

Musical dynamics—loudness and changes in loudness—forms one of the key aspects of expressive music performance. Surprisingly this rather important research area has received little attention. A reason is the fact that while the concept of dynamics is related to signal amplitude, which is a low-level feature, the process of deriving perceived loudness from the signal is far from straightforward.

This thesis advances the state of the art in the analysis of perceived loudness by modelling dynamic variations in expressive music performance and by studying the relation between dynamics in piano recordings and markings in the score. In particular, we show that dynamic changes: a) depend on the evolution of the performance and the local context of the piece; b) correspond to important score markings and music structures; and, c) can reflect wide divergences in performers' expressive strategies within and across pieces.

In a preparatory stage, dynamic changes are obtained by linking existing music audio and score databases. All studies in this thesis are based on loudness levels extracted from 2000 recordings of 44 Mazurkas by Frédéric Chopin. We propose a new method for efficiently aligning and annotating the data in score beat time representation, based on dynamic time warping applied to chroma features. Using the score-aligned recordings, we examine the relationship between loudness values and dynamic level categories.

The research can be broadly categorised into two parts. The first investigates how dynamic markings map to performed loudness levels. Empirical results show that different dynamic markings do not correspond to fixed loudness thresholds. Rather, the important factors are the relative loudness of neighbouring markings, the inter-relations of nearby markings and other score information, the structural location of the markings, and the creative license exercised by the performer in inserting further interpretive dynamic shaping.

The second part seeks to determine how changes in loudness levels map to score features using statistical change-point techniques. The results show that significant dynamic score markings do indeed correspond to change points. Furthermore, evidence suggests that change points in score positions without dynamic markings highlight structurally salient events or events based on temporal changes.

In a separate bidirectional study, we investigate the relationship between dynamic markings in the score and performed loudness using machine learning techniques. The techniques are applied to the prediction of loudness levels corresponding to dynamic markings, and to the classification of dynamic markings given loudness values. The results show that loudness values and markings can be predicted relatively well when trained across recordings of the same piece, but fail dismally when trained across a pianist's recordings of other pieces. The findings demonstrate that score features may trump individual style when modelling loudness choices. The analysis of the results reveal that form—such as the return of the theme—and structure—such as repetitions—influence predictability of loudness and markings.

This research is a first step towards automatic audio-to-score transcription of dynamic markings. This insight will serve as a tool for expression synthesis and musicological studies.

Acknowledgements

I would like to thank my supervisors Elaine Chew and Oscar Bandtlow in particular for their immense feedback and support over these years.

It has been a pleasure to work and interact with some excellent people at **C4DM** at QMUL such as Jordan Smith, Sebastian Ewert, Emmanouil Benetos, Dan Stowell, Siying Wang, Yading Song, and George Fazekas; they were close to me throughout my Ph.D. studies and formed a very fruitful and unique environment where I developed my ideas.

I am grateful that I have been surrounded by inspiring people who have had interdisciplinary research background such as Laurel Pardue, Siddhart Sigtia, Bogdan Vera, Luwei Yang, Victor Zappi, Kat Agres, Alesia Milo, Chris Heinrichs, Brecht De Man, and Matthiew Barthet. Also thanks to all **mupae** people for the fruitful meetings and the long research or life-related discussions.

I could not forget to acknowledge other people that I had the opportunity to work with during these years; these are Rafael Ramirez, Rebecca Killick, and Akira Maezawa. I am grateful to the Music Techonology Group of Universitat Pompeu Fabra and its effective guidance while pursuing my research visit in Barcelona.

I'm grateful for the financial support offered to me by Queen Mary University of London through a three-year studentship and several travel grants in the course of the program.

Finally I would like to thank my good friends Maria Panteli, Srikanth Cherla, Marilena Karanika and Ioana Dalca who have been great supporters of any kind anytime. Special thanks goes to Panos for his moral and practical support and to the Peninsula where I wrote my final papers. Last but not least, I want to thank my father for giving me strength, motivation and inspiration throughout my previous studies, and for always reminding me to follow my dreams.

Katerina Kosta,
23^{rd} January 2017

Licence

Contents

1 Introduction **13**

1.1 Problem overview 13

1.1.1 Performers' expression through music 13

1.1.2 Music interpretation through a score 14

1.2 Motivation 16

1.3 Aim 17

1.4 Contributions 17

1.5 Associated publications 17

1.6 Dissertation outline 19

2 Background **20**

2.1 The concept of loudness 20

2.1.1 Signal properties and its relationship to perceived loudness 22

2.1.2 Sone values as source of loudness information 23

2.1.3 Alternative loudness extraction techniques from music signal 25

2.1.4 Psychological studies of loudness perception in music . . . 25

2.2 The concept of score notation 26

2.3 Articulation in Chopin's works and notation 28

2.4 Related studies 30

2.4.1 Mapping between audio loudness and score 30

2.4.2 The interconnection of changes in loudness and tempo . . 31

2.4.3 Expression features used for structure extraction 32

3 Data acquisition **33**

3.1 Audio and score information 34

3.2 Audio recording to score alignment 35

3.2.1 Optimizing *the reference audio* choice 37

3.2.2 *Reference audio* detection heuristic 37

3.3 Loudness information of dynamic markings 39

4 Dynamics and relativity: Is *piano* always *piano*? 41
4.1 Performed Loudness Study . . . 43
4.1.1 Ordinal Loudness Sequence Preserved On Average? . . . 43
4.1.2 Is the Ordinal Loudness Sequence Preserved In Pairwise Instances? . . . 55
4.1.3 Analysis of different manifestations of the same dynamic markings throughout a piece . . . 58
4.2 Conclusions - Discussion . . . 64

5 *Interlude*: An analysis of outliers in performed loudness transitions 65
5.1 Outliers in individual marking pair transition . . . 66
5.2 Recordings in outlier dynamic clusters . . . 68
5.3 Discussion . . . 71

6 A change-point approach towards representing musical dynamics 72
6.1 Change-points techniques . . . 73
6.2 Change-points extraction . . . 75
6.3 Comparison results . . . 76
6.3.1 Conclusions and discussion . . . 78
6.4 Analysis of change-points locations . . . 78
6.4.1 Discussion . . . 80

7 Mapping between dynamic markings and performed loudness: A machine learning approach 82
7.1 Material . . . 83
7.2 Learning task and algorithms . . . 84
7.3 Performed loudness-level modeling . . . 87
7.3.1 Predicting loudness given other pianists' recordings of target piece . . . 88
7.3.2 Predicting loudness given target pianist's recordings of other pieces . . . 92
7.3.3 A pianist's interpretation may not be predictable based on his/her approach to other pieces . . . 94
7.3.4 Inter pianist similarity . . . 97
7.3.5 Discussion on feature analysis . . . 98
7.4 Dynamic marking prediction . . . 99
7.4.1 Predicting dynamic markings given other pianists' recordings of target piece . . . 100

7.4.2 Predicting dynamic markings given target pianist's recordings of other pieces . . . 101
7.4.3 Easily predicted markings . . . 102
7.4.4 Easy/hard to predict Mazurkas: ratio of correctly-classified markings . . . 104
7.5 Conclusions . . . 105

8 Conclusions and further work **108**
8.1 Conclusions . . . 108
8.2 Future Directions . . . 109

A Glossary of music terms **111**

B Change-points plots across the Mazurkas' recordings **114**

List of Figures

2.1 Sample equal loudness curves 22
2.2 Part of notation developed by Logothetis for his *graphic scores*. . 27
2.3 The score of Logothetis's piece *Agglomeration*. 27

3.1 Comparison of the groups constituted by the median error values per recording, implementing the Friedmann test correlation rank for the 42 recordings of Mazurka Op.6 No. 2 38
3.2 Loudness representation of a single Mazurka Op. 6, no. 1 recording 40

4.1 The Mazurka cases where only the ***f*-*ff*** pair has negative τ value 46
4.2 The repeated phrase where the ***f*** appears in Mazurka Op. 68 No. 3 46
4.3 The repeated phrase where the ***ff*** appears in Mazurka Op. 30 No. 3 47
4.4 The Mazurka cases where only the ***pp*-*p*** pair has negative τ value 48
4.5 The position of the ***pp*** marking in Mazurka Op. 33 No. 4 49
4.6 The ***pp*** markings in Mazurka Op. 67 No. 3 49
4.7 The Mazurka cases where only the ***p*-*mf*** pair has negative τ value 50
4.8 The ***mf*** marking in Mazurka Op. 50 No. 1 and its relation with the preceded ***p*** marking. 51
4.9 Box plots of the dynamic values of the markings belonging to the pairs ***p*-*f***, and ***mf*-*f*** in Mazurka Op. 41 No. 2 52
4.10 The location of the first ***f*** marking (top), and the location of the second ***f*** marking in Mazurka Op. 41 No. 2 (bottom) 52
4.11 Box plots of the dynamic values of the markings belonging to the pairs ***pp*—*ff***, ***p*—*ff***, and ***f*—*ff*** in Mazurka Op. 50 No. 3 53
4.12 Box plots of the dynamic values of the markings belonging to the pairs ***pp*—*mf***, and ***p*—*mf*** in Mazurka Op. 67 No. 1 53
4.13 Box plots of the dynamic values of the markings belonging to the pair ***pp*–*p*** and the pair ***pp*–*mf*** in Mazurka Op. 67 No. 2 54
4.14 The repeated phrase in Mazurka Op. 67 No. 2 which includes the ***pp*** marking, and the two ***p*** markings that follow up. 55

4.15 The log ratio of the loudness in the transition from a marking m_{k-1} to a marking m_k in the score sequence 57
4.16 The average standard deviation values of the log loudness ratio for each marking—transitions are from the x-axis to the y-axis. . 58
4.17 The case of the loudness values of groups of dynamic markings (***ff***'s in Mazurka Op. 6 No. 1—top, and ***f***'s in Mazurka Op. 67 No. 3—bottom) 61
4.18 Multiple comparison of the means for the groups of the same markings 62
4.19 The repeated phrase in Mazurka Op. 6 No. 3 where the markings $\boldsymbol{p}_4$, $\boldsymbol{p}_5$, $\boldsymbol{p}_6$ (repetition of $\boldsymbol{p}_4$), and $\boldsymbol{p}_7$ (repetition of $\boldsymbol{p}_5$) appear. . 63

5.1 Distribution of outlier transitions between each dynamic marking pair 67
5.2 Score excerpt from Mazurka Op. 59 No. 3 where the markings of the first pair (***f***, ***p***) appears 68
5.3 Loudness graphs for performances of Mazurka Op. 24 No. 2 . . . 69
5.4 Loudness transition clusters found for Mazurka Op. 7 No. 1 (top) and Op. 33 No. 3 (bottom) 70

6.1 An example change-point detection result from the same audio excerpt (raw data in sones) 74
6.2 The distribution of what each of the peaks at change-point positions represent in the score. 79
6.3 One phrase boundary detected in Mazurka Op. 33 No. 2 in beat 23 80
6.4 One motif boundary detected in Mazurka Op. 30 No. 1 in beat 155 80

7.1 Pearson correlation coefficient between predicted and actual loudness values for each Mazurka 89
7.2 Representation of the loudness levels on the marking positions in score time for Mazurka M24-3 for the eight pianists. 90
7.3 Pearson correlation coefficients of Mazurkas for which the worst predicted pianist's recording scored a negative r value 91
7.4 Loudness time series in score-beat time for recordings of Mazurka Op. 24 No. 1 92
7.5 Pearson correlation coefficient mean, min, max, and median values for each machine learning method 93
7.6 Euclidean distance between predicted and actual loudness values for each pianist 95

7.7 Pearson correlation coefficient values for each pianist, averaged over all machine-learning methods, for the Mazurkas having all negative 96
7.8 Average actual and predicted dynamic values at dynamic marking positions for the Mazurkas M07-1, M24-2, M24-4, and M50-3 96
7.9 Matrix displaying Pearson correlation coefficient values averaged over all machine-learning methods 98
7.10 Upper bound for average (over all machine learning methods) Pearson correlation coefficient as number of features considered increases 100
7.11 Case of the ***ff*** marking in Mazurka Op. 17 No. 4 (M17-4), correctly classified in all recordings 103
7.12 Case of the ***f*** marking in Mazurka Op. 24 No. 4 (M24-4), correctly classified in all recordings 103
7.13 Case of the ***pp*** marking in Mazurka Op. 30 No. 3 (M30-3), correctly classified in all recordings 104
7.14 Average ratio of correctly classified markings and number of markings per Mazurka over all pianists 104
7.15 Loudness values at dynamic marking positions for Mazurka Op. 50 No. 2 105

List of Tables

3.1 Chopin Mazurkas used in this research, the number of recordings for each one, and the number of dynamic markings that appear in each one 34
3.2 Dynamic and tempo markings/text that appear in the scores. . . 35

4.1 The Kendall's τ values for pairwise comparisons of the mean dynamic values that correspond to the dynamic markings for each Mazurka 45
4.2 List of marking pairs that contradict the OLS with information on Mazurkas where the pairs appear as well as their proportion with respect to the total number of pairs in that Mazurka. 56
4.3 Results for one-way (P–value 1) and two–way (P–value 2) ANOVA tests for marking groups per Mazurka 60

5.1 Pianists whose recordings had the highest proportion of outlier transitions. 67
5.2 Recordings having the highest proportion of outlier transitions. . 68
5.3 Pianists having the highest proportion of recordings in outlier clusters. 69
5.4 Pianists whose recordings co-occur in more than twenty outlier clusters. 71

6.1 The value of the parameter k in the threshold formula that has been used for all the recordings per Mazurka 75
6.2 Average and maximum F-measure, P and R values and average Hausdorff distance for change points selected. 77
6.3 Markings that have been present in the score in positions of change-point peaks 79

7.1 Pianist's name, year of the recording, and pianist ID. 83
7.2 Ranking of importance of the features for the loudness prediction task. 99

7.3 Percentage of Correctly Classified Instances (CCI) 101
7.4 Markings that have been predicted correctly for all recordings of the Mazurka containing that marking 102

Chapter 1

Introduction

1.1 Problem overview

1.1.1 Performers' expression through music

Musical expression is an important part of music performance analysis. Palmer and Hutchins [2006] underscore that performers "manipulate the sound properties, including frequency (pitch), time, amplitude, and timbre (harmonic spectrum) above and beyond the pitch and duration categories that are determined by composers." These manipulations define the term "musical expression" which is partially achieved by altering the variables for stress, rhythm, accent, and intensity contour. Another description of what expressiveness means is posed by Fabian et al. [2014] in their letter to the authors of their book; however they mention that it "could be a point of departure rather than a prescription":

"Expressiveness:
1. refers to the effect of auditory parameters of music performance (loudness, intensity, phrasing, tempo, frequency spectrum, etc.)–covering acoustic, phychoacoustic, and/or musical factors
2. refers to the variation of auditory parameters away from a prototypical performance, but within stylistic constraints (e.g. too much variation is unacceptable, and does not fall within the gamut of expressiveness)
3. is used in the intransitive sense of the verb (no emotion or mood or feeling is necessarily being expressed; rather the music performance sounds "expressive" to differing degrees)." ([Fabian et al., 2014, pg.xxi])

From the standpoint of the performer, Benetti Jr. [2013] showed that pianists in particular employ procedures such as melodic shaping, emphasis on char-

acter, articulations, structural rules for applying *rubato*, and large dynamic contrasts. These results demonstrate that performers generally conceptualise expressiveness in relation to aesthetic trends. However one should remember that performers reflect composers' intentions. [Cook, 2000, p.14] highlights that classical music scholars distinguish between authors (i.e. composers) and reproducers (i.e. performers). The latter are often criticised when they obscure the original music through "over-interpretation or gratuitous virtuosity" ([Cook, 2000, p.14]).

Uncertainty remains with respect to whether "the expressive deviations measured are due to deliberate expressive strategies, music structure, motor noise, imprecision of the performer, or even measurement errors." (Langner and Goebl [2003]). Indeed, while a significant part of the performer's aim is to communicate the composer's intentions, nevertheless, performers bring their personal, cultural, and historical viewpoints to the fore when subjectively understanding expression [Fabian, 2014, p.61].

Of particular interest, in our research we use recordings from many pianists playing the same Mazurka pieces by Frédéric Chopin. He was both a composer and a performer, such that his personal performing style was eventually imitated internationally, although he may not have had the intention to emphasize his own originality ([Einstein, 1975, p.215]). This is important to be taken into account when analysing the variations across different interpretations of his pieces.

1.1.2 Music interpretation through a score

The problem of music interpretation is a complex one. The score is a representation of the composer's intentions. The same symbol—be it a note, dynamic marking, indication of articulation, or phrase grouping—can have a variety of possible interpretations. The original score is refracted through the performer [Rink, 2002, p.59], who can choose to render the symbols in unique ways. The auditive effect of the refracted work is communicated to the listener, who in turn perceive it as a subjective psychological experience.

The problem is made more complex as "even well-known scores that appear to have widely-recognized meanings are changing all the time, not simply as general performance style changes but in their characterization, leading to a change in their perceived nature" Leech-Wilkinson [2012]. Possible deviations from the score occur, mostly playing notes in different octave as written or applying rubato in position that is not indicated. More general, as it is described in Cook et al. [2009]: "The practice of working from a score-based analysis to a recording basically declares off limits all those aspects of performance that

cannot be directly related to notational categories; it eliminates most of what there is to study before you even start, including all the rhetorical, persuasive, or expressive effects that contribute so much to the meaning of music as performance yet have little or nothing to do with structure as the music theorist sees it."

The focus of this thesis is on concepts related to dynamic levels in musical expressivity which are represented in the score by markings such as ***p*** (piano, meaning soft) and ***f*** (forte, meaning loud). These symbols are interpreted in performance and communicated through the varying of loudness levels.

Musical dynamics can be conceptualized to exist on two levels: "Primary dynamic shadings" are the main dynamic levels associated with notated marking, and "inner shadings" are dynamic variations associated with the foreground level, which may not be notated (Khoo [2007]). This is analogous to the analysis of musical meter by Volk [2008], where the notated time signature indicates the outer meter, while the local grouping of beats defines the inner meter. Inner and outer meters sometimes conflict; similarly, the absolute loudness of dynamic markings may be superseded by their local context so that a ***p*** dynamic might objectively be louder than a ***f*** at another part of the piece, as shown in our preliminary study (Kosta et al. [2014]).

One target is to understand the connection between loudness levels in recordings and printed dynamic markings in the score. However, two important considerations concern score markings: first, dynamic markings may simply reinforce the natural predisposition of a music part to be expressed in a particular way based on its structure; second, precise annotation by the composer can be present at a non-obvious place (Grachten and Widmer [2012]). Together with the large amount of information and performance related parameters (Bisesi and Parncutt [2010]), these challenges make the modelling of music expression one of today's most important unsolved problems in the description of music audio.

1.2 Motivation

Information abstracted from music signals can serve as an important source of data for multimedia content analysis. A big challenge in creating these data resources lies in the creation of meaningful and salient features for performed music, which describes almost all the music that we hear.

The extraction of performance features, especially those related to expressivity, remains unelucidated today. In the past decade, several research groups have developed algorithms for analysing musical expression in order to understand or predict audio characteristics that determine expressiveness. They mostly derive information from the analysis of features such as timbre and roughness or from the analysis of basic concepts such as tempo and audio amplitude.

The origin of our research is to analyse the musical dynamics in isolation and beyond the concept of amplitude variations. The change in perceived loudness as music unfolds is indispensable in expressive music performance. Surprisingly, despite its prominence in musical culture, this research topic has received little scholarly attention. Main keys to enforce the importance of the topic include the fact that the complexity in ability to modulate dynamics in an orchestra is an index of music evolution throughout the ages and it is one of the key roles of the conductor. The same happened in instrument making which evolved with the ability to manipulate dynamics. An example of the above is the transfer from baroque to the romantic period. Also the manipulation of dynamics is a crucial skill to achieve for expert musicians and it is key for music critics and audiences alike when deciding whether the player performed as composer intended.

Also we want to better understand the parameters which define the performers' responses to dynamic markings. However, the connection between loudness levels in performance and the printed dynamic markings in the score is still poorly understood. While the meaning of prosodic inflections given the same words have been widely studied in speech and linguistics[1], very little work exists that addresses different meanings of the same symbolic representations of expressive nuances in music.

Another challenge of conducting systematic research on expressive parameters is the lack of audio data and annotations for analysis and for training; all the above were the motivation for this thesis.

[1]For example, studies have examined the meanings of nuances in utterances of the words "whatever" Benus et al. [2007] and "okay" Gravano et al. [2007].

1.3 Aim

The aim of this work is to develop automatic methods that can more accurately and meaningfully map musical expression in recorded music to an ontology for music dynamics, beyond simply extracting a direct dynamic level; such techniques will also be valuable for music transcription. On the synthesis side, the analysis will be useful for generating expressive renderings of notated scores. It also serves as a tool for musicological studies.

1.4 Contributions

This thesis advances the state of the art in the analysis of perceived loudness by modelling dynamic variations in expressive music performance and by studying the relation between dynamics in piano recordings and markings in the score. In particular, we show that:

Dynamic changes:
a) depend on the evolution of the performance and the local context of the piece;
b) correspond to important score markings and music structures; and,
c) can reflect wide divergences in performers' expressive strategies within and across pieces.

We introduce change-point detection as a means toward extracting conceptual dynamic levels, generally represented in the score by markings. Change-point algorithms have been applied to domains such as climatology, bioinformatics, finance, oceanography, and medical imaging (see Killick and Eckley [2014]). To our knowledge, audio dynamics represents a novel application area.

Also we introduce a novel problem in machine learning by investigating machine-learning techniques for the prediction of loudness levels corresponding to dynamic markings, and for the classification of dynamic markings given loudness values.

The data that has been created for the purpose of the thesis is publicly available at the following address: https://github.com/katkost/MazurkaBL.

1.5 Associated publications

Portions of the work detailed in this thesis have been presented in national and international scholarly publications, as follows (journal submissions and publications highlighted in bold):

– Kosta, K., O. F. Bandtlow, E. Chew (2014). Practical Implications of Dynamic Markings in the Score: Is piano always piano? In Proceedings of the 53rd Audio Engineering Society (AES) Meeting on Semantic Audio, Jan 26-29, 2014, London, UK.

– Kosta, K., O. F. Bandtlow and E. Chew (2014). A Study of Score Context-dependent Dynamics in Piano Performance. In Proceedings of the Performance Studies Network International Conference (PSN3), Jul 17-20, Cambridge, UK.

– **Kosta, A., Li, S. (2014). 2013 Performance Studies Network International Conference. Computer Music Journal, 38(2): 78-80.**

– Kosta, K., O. F. Bandtlow, E. Chew (2015). A Change-point Approach Towards Representing Musical Dynamics. In T. Collins, D. Meredith, A. Volk (eds.): Mathematics and Computation in Music: 5th International Conference, MCM 2015, London, UK, June 22-25, 2015, Proceedings, pp. 179–184, Lecture Notes in Computer Science 9110, Berlin: Springer.

– Kosta K., R. Ramirez, O. F. Bandtlow, E. Chew (2015). Predicting loudness levels and classifying dynamic markings in recorded music. In Proceedings of 8th International Workshop on Machine Learning and Music (MML2015), Machine Learning for Music Generation, Vancouver.

– Kosta, K., O. F. Bandtlow, E. Chew (2016). Outliers in Performed Loudness Transitions: An Analysis of Chopin Mazurka Recordings. In Proceedings of the 14th International Conference for Music Perception and Cognition (ICMPC), pp. 601-604, July 5-9, 2016, San Francisco, California, USA.

– **Kosta, K., R. Ramirez, O. F. Bandtlow, E. Chew (2016). Mapping between dynamic markings and performed loudness: A machine learning approach. Journal of Mathematics and Music, 10(2): 149–172.**

– **Kosta, K., O. F. Bandtlow, E. Chew (submitted to Journal of New Music Research). Dynamics and Relativity: Practical Implications of Dynamic Markings in the Score.**

– Kosta, K., S. Ewert, O. F. Bandtlow, E. Chew (to be submitted in ISMIR2017). MazurkaBL: A dataset of loudness and beat-position information for 2000 Chopin Mazurka recordings.

1.6 Dissertation outline

The thesis includes the following chapters:

Chapter 2 surveys in more detail the main concepts for this thesis and explores existing relevant studies.

Chapter 3 describes the data that has been created for the thesis.

Chapter 4 presents the first study of the thesis which introduces the approach from dynamic marking to performed loudness.

Chapter 5 expands the study presented in Chapter 4 by focusing on the data outliers.

Chapter 6 presents the second study of the thesis which introduces the approach from performed loudness to score features.

Chapter 7 presents the third study of the thesis which introduces methods for the prediction of loudness levels corresponding to dynamic markings and for the classification of dynamic markings given loudness values.

Chapter 8 concludes the thesis and sets future directions.

Appendix A presents a glossary of the music terms used in this thesis.

Appendix B includes plots from the results derived from part of Chapter 6.

Chapter 2

Background

In this chapter we start by reviewing the two main subjects of the thesis which are loudness measurement (Section 2.1) and musical notation (Section 2.2). This will be useful in our understanding of the way we extract information from the audio signal and in our understanding of how this information can be represented as a musical concept, respectively. More details on the interpretation of Chopin's pieces are presented (Section 2.3). We then review existing work in the area (Section 2.4).

2.1 The concept of loudness

"A painter paints pictures on canvas.
But musicians paint their pictures on silence."
—Leopold Stokowski

Loudness is a subjective sensation of the magnitude or strength of sound. Intensity, on the other hand is an objectively measured physical property, an amount of power. It is expressed in power units per unit area, as in watts per square meter. Sound intensity is the major physical determinant of loudness, but the relationship between intensity and loudness is very intricate. The fact that the intensity-loudness relationship is confounded by frequency is especially important for music: Because the ear's sensitivity varies with frequency, equal intensities do not elicit equal loudness across all audible frequencies. The unit for sound intensity is watts per square meter. $1\mathrm{W/m}^2$ is a rather intense sound which approaches the upper limit of hearing. The threshold of hearing at 1000 Hz is around $10^{-12}\mathrm{W/m}^2$.

Sound intensity is usually expressed in terms of intensity level rather than pure intensity. Intensity level implies a comparison of a particular intensity to

some baseline value. The comparisons implied by intensity levels of sound are ratio comparisons which use the unit known as "decibel". Decibels are measures of power ratios and they are best defined in terms of the formula by which they are computed. For sound intensity,

$$D = 10[\log(I/I_0)], \tag{2.1}$$

where D is the number of decibels, I is the particular intensity which the decibel value places on a level, and I_0 is the baseline value. The baseline that is used is the intensity at the threshold of hearing.

It is important to recognise two consequences of the definition of decibels as the logarithm of a ratio. One, a relative small range of decibels encompasses a relatively large range of acoustic powers. Two, one cannot simply combine decibels to obtain an intensity level for tonal combinations. The decibel scale is logarithmic; equal decibel amounts stand for equal ratios.

Sound Pressure Level (SPL) is a more common stimulus measure than intensity level because it is easier to measure and it relates logically to the concept of pressure variations being responsible for sound. It is expressed in decibels which are computed differently in relation to a different baseline. The formula is

$$D = 20[\log(P/P_0)], \tag{2.2}$$

where D is the number of decibels, P is a particular pressure value, and P_0 is a baseline pressure. A common baseline is .00002 newtons per square meter. Most decibel measures in psychoacoustic literature are indicators of SPL.

Volume is the apparent size or extensity of a sound. Volume's existence as an independent tonal attribute is questionable. It has no parallel relationship with any single physical attribute in the way that loudness is related to intensity or pitch is related to frequency.

As referred by [Radocy and Boyle, 2003, pg.52–65], experiments by Stevens [1934] show that an increase in intensity level brought a more rapid volume increase at higher than at lower frequencies. In later research Stevens and Terrace [1962] plotted frequency-intensity combinations used for equal volume judgements and formed a "vol" scale of measurement in which 1 vol equals the apparent volume of 1000 Hz at 40 decibels of sound pressure level. They found that regardless of frequency, at around 140 decibels all tones sounded equal in volume. In general, higher frequencies required greater intensities for equivalent volume, and the rate of growth in volume (as intensity increased) varied with frequency.

Studies by Fletcher and Munson [1933] yield equal loudness contours that are represented by the Munson-Fletcher curves shown in Figure 2.1. Each point

on any curve represents a particular frequency-intensity information and any point on an equal loudness curve is as loud as any other point on the same curve. The curves represent the auditory behaviour, but they enable loudness comparisons only at the less than, equal to, or greater than level.

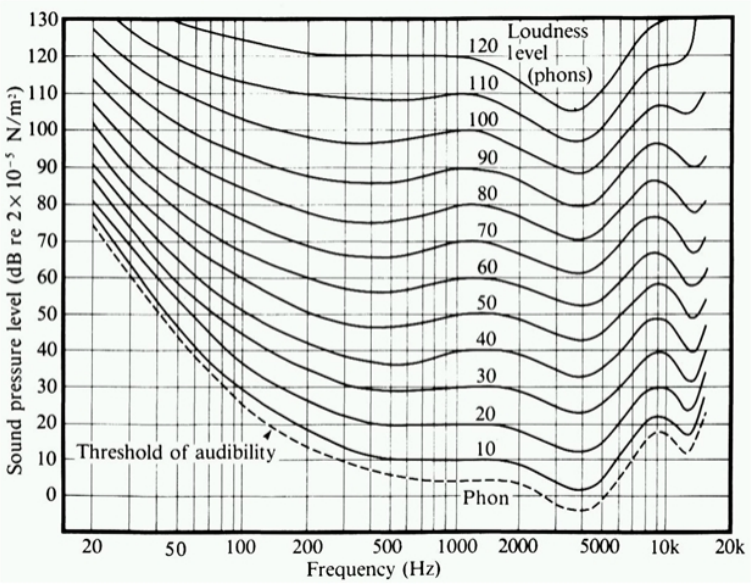


Figure 2.1: Sample equal loudness curves. The curves connect frequency-intensity level combinations which are judged to be equally loud as a 1000Hz standard of a given intensity level ([Radocy and Boyle, 2003, pg.58]). Figure derived from website `https://blogs.msdn.microsoft.com/audiofool/2007/02/07/louder-sounds-better/`, accessed 20th December 2016.

Each curve has a phon value. Phons is a measurement unit for loudness level (not loudness), thus are measures of equivalence to the intensity level of some standard. Phons are arbitrarily set to be equivalent to decibels at 1000 Hz; e.g., a 1000 Hz tone of 60 decibels is also 60 phons. Any other frequency judged as equal in loudness to the 1000 Hz–60 decibels standard is also 60 phons.

2.1.1 Signal properties and its relationship to perceived loudness

Apart from the integration along a perceptual frequency axis as described above, perceived loudness is related to a few other essential properties of the auditory system in either frequency or time domain, as presented by Nielsen and Skovenborg [2004]. The properties that are related to the frequency domain include the phenomenon of masking, that is the affection of the perception of one sound when a louder sound is present at the same time. Also, the phenomenon of spreading, closely related to masking, describes a complex interaction between

different frequencies that are sounded simultaneously: not only the actual frequencies of a signal contribute to the loudness, but also to some extent the higher and lower neighbouring frequencies.

Another property of loudness perception is the spectral loudness summation which is associated with the concept of critical bandwidth; the loudness of a signal within one critical band is independent of its bandwidth, however when the signal's bandwidth grows outside the critical band there is an increase in loudness despite the consistency of the total level.

Applying varying degrees of multi-band dynamics compression affects on loudness which is contrary to the resulting fluctuation depth as Moore [2005] describes in his study to speech signals. Nielsen and Skovenborg [2004] points out that "one of the interesting conclusions was that the long-term loudness, for a given signal, increases with the degree of dynamics compression applied, and with the RMS value held constant", where RMS is the Root Mean Square of the audio signal that is obtained by squaring the amplitude at each instant, acquiring the average of the squared values over the interval of interest, and then taking the square root of this average.

Filtering processes in hearing that influences the loudness perception include the direction dependent filtering, the phenomenon of binaural loudness summation which is the result of having two sensors of sound (ears), and the altering of the sound signal by certain characteristics of the environment, for example the reverberation characteristic.

Properties that are related to the time domain affect the way we perceive loudness. There is a delay between the appearance of an audio signal and the reaction to it by human hearing. Also the post-masking effect is the phenomenon of the perception of a sound fading out, although its source instantaneous stopped producing it. The hearing ability is reduced for a period of time when we are exposed to loud sound for a certain duration. Both the bandwidth and the duration of a burst or tone affect the perceived loudness, however "no simple time constant can be specified for the temporal loudness integration." (Nielsen and Skovenborg [2004])

2.1.2 Sone values as source of loudness information

Laboratory attempts to measure loudness in terms of how much louder one sound is than another have used observers' productions and/or estimations of various loudness distances or ratios. A sone scale of loudness, based on estimates of apparent ratios, and a lambda scale, based on equal-appearing intervals and ratios, are alternative ways of describing apparent loudness relationships (Beck and Shaw [1967]). The reference of the sone scale in particular is that a 1000

Hz sine tone with a level of 40 dB SPL (decibels in reference to sound pressure level) has a loudness of one sone and it is defined by the following equation:

$$\Psi(sones) = \frac{1}{15.849}\Big(\frac{I}{I_0}\Big)^{0.3}, \tag{2.3}$$

where $I_0 = 10^{-12}\mathrm{W/m}^2$ (Hartmann [2004]).

In our research we have used the sone values at the loudness extraction process from the audio signals. One main reason was the fact that the sone scale is linear, so in this way we can normalise in a more accurate and easy way across different recorded environments. Another equally important reason is that we wanted to train our models on audio that is pre-processed based on the basic psychoacoustic concept of the equal loudness curves, and not to make any other compression or modification. If we want a *machine* to *listen* like a human, then it needs a similar input to the one that a listener has.

For the experiments that are presented in this thesis, the loudness information is extracted from each audio signal using the `ma_sone` function in Elias Pampalk's Music Analysis toolbox[1]. The specific loudness sensation in sones per critical-band is calculated by following the process as explained by Pampalk et al. [2002]:

We calculate the power spectrum of the audio signal using a Fast Fourier Transformation. We use a window size of 256 samples, hopsize of 128, and a Hanning window with 50% overlap. The frequencies are bundled into 20 critical-bands and these frequency bands "reflect characteristics of the human auditory system, in particular of the cochlea in the inner ear" (Pampalk et al. [2002]). Then we calculate the Spectral masking effects, based on the research by Schroeder et al. [1979]. Then we calculate the loudness as dB-SPL unit and from these values we calculate the equal loudness levels in Phon. From these values we detect the values in sones, following the calculation described by Bladon and Lindblom [1981]: if loudness level in phons is L, then it is converted to loudness level in sones S following the formula:

$$S = \begin{cases} 2^{(L-40)/10}, & P \geq 40 \\ (L/40)^{2.642}, & P < 40. \end{cases} \tag{2.4}$$

The extracted loudness time series have been smoothed by local regression using a weighted linear least squares and a 2nd degree polynomial model (the "loess" method of MATLAB's `smooth` function[2]). The loudness time series for each recording is normalised by dividing with its maximum loudness value. In

[1] `www.pampalk.at/ma/documentation.html`, accessed 20 February 2016.

[2] `http://uk.mathworks.com/help/curvefit/smooth.html?refresh=true`, accessed 20 February 2016.

this way we are able to compare different recording environments.

2.1.3 Alternative loudness extraction techniques from music signal

Below we present two techniques that exist and are used for extracting the loudness from a music signal. The first one is called "Power curve"[3] from the plugins that have been created through the Mazurka project[4]. At the synopsis they describe it as measuring the power over time of a signal. They get the average power in decibels for a region of audio data following the equation

$$P_{avg} = 10 \log_{10} \Big(\frac{1}{N} \sum_{n} x_n^2 \Big), \tag{2.5}$$

where N is the number audio samples to average over, and x_n are the individual audio samples. In addition the plugin can provide a more sophisticated outcome, such as a time-symmetric exponentially smoothed function of the average raw power for each input signal region block, a smoothed power slope of the power curve, as well as a scaled power slope of the smoothed power curve.

The second one is the "intensity" plugin from the BBC audio plugins[5]. It calculates the intensity of a signal following the work by Lu et al. [2006]. As it is described at the plugin documentation, firstly the signal is divided into i subbands with the following frequency ranges:

$$\Big[0, \frac{\omega_0}{2^i}\Big), \Big(\frac{\omega_0}{2^i}, \frac{\omega_0}{2^{i-1}}\Big), \ldots, \Big(\frac{\omega_0}{2^2}, \frac{\omega_0}{2^1}\Big], \tag{2.6}$$

where ω_0 refers to the sampling rate. The intensity of each signal frame n is calculated by:

$$I(n) = \sum_{k=0}^{\omega_0/2} A(n,k),$$

summing the magnitude A of each frequency bin k.

2.1.4 Psychological studies of loudness perception in music

Several studies included in psychoacoustical literature have been attempted to obtain relationships between the subjective magnitude of loudness change and the physical magnitude of intensity change using pure tones or noise (see, for

[3] `http://sv.mazurka.org.uk/MzPowerCurve/`, accessed 18 October 2016
[4] `http://mazurka.org.uk/`, accessed 18 October 2016
[5] `https://github.com/bbc/bbc-vamp-plugins`, accessed 18 October 2016

example Moore et al. [1997] and references therein). However studies using music stimuli in their experiments (see, for example Geringer [1993] and references therein) showed that music stimuli in context were perceived somewhat differently than were the pure tone and noise-band stimuli of previous research.

Following the concept of dynamic changes using music stimuli, the study contacted by Geringer [1995] investigated listener perception of dynamic change within a musical context, regarding both stimulus presentations and subject-response procedures. Musician and non-musicians responded continuously during excerpts using the Continuous Response Digital Interface (CRDI) to indicate perceived loudness levels. Excerpts were selected from different music genres, including instrumental jazz, popular vocal, symphonic orchestra, solo piano, and synthesizer examples. The results revealed that musician subjects indicated a significantly smaller magnitude of dynamic change than non-musician subjects did.

2.2 The concept of score notation

"We call it music, but that is not music:
that is only paper."
—Leopold Stokowski

The concept of music notation, that is the methods of writing down music so that it can be performed, was not involved for a long period in the world's music history. However, it was one of the key points for the development of the European classical music.

An early system of notation was the neumatic notation. Neumes, from the Greek word "neuma", meaning "gesture" or "sigh", was the system of musical notation used from the 7^{th} to 14^{th} century that evolved from representing grave and accute accents to forming a system of a precise indication of pitch for singing (Rutherford-Johnson et al. [2012]). [Cook et al., 2009, pg.11] refers to a conclusion for neumatic notation from a study by Sam Barrett (Barrett [2008]), mentioning that "'music is in other words conceived platonically, as an abstract and enduring entity that is reflected in notation, with the notation itself being reflected in singing (since mistakes in singing can be corrected by reference to the notation)."

During the development of the notation to the form we know today, there were changes mostly on the representation of the duration and the pitch of the notes that are sounded. Innovations included the development of notational symbols for different playing techniques and performance actions. The composer who specified dynamics for the first time in a score was Giovanni

Gabrieli (1954-1612) in the *Sonata pian e forte* from the Sacrae symphoniae (1597) ([Sadie, 1998, pg. 28–29]). Annotation of dynamics, such as *piano*, "has remained relatively constant, although contemporary composers have explored its extremes." (Rutherford-Johnson et al. [2012])

An example of the latter statement could be the notation that has been established by the composer Anestis Logothetis for his avant-garde music[6]. Figure 2.2 that shows the notation for changes in expression and Figure 2.3 that shows his composition *Agglomeration* are attached in order to display his idea.

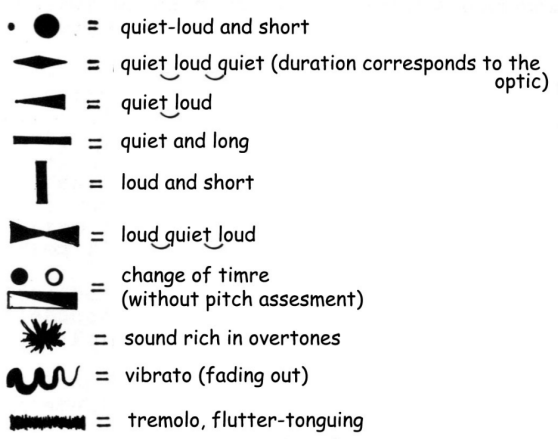


Figure 2.2: Part of notation developed by Logothetis for his *graphic scores*.

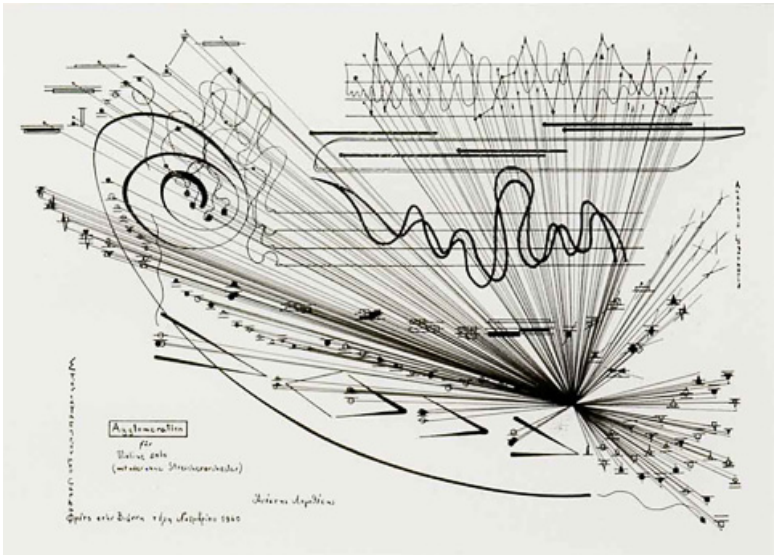


Figure 2.3: The score of Logothetis's piece *Agglomeration*.

[6] http://anestislogothetis.musicportal.gr/the_graphic_notation/?lang=en

The concept of score notation can reveal an extra dimension, the one corresponding to the time the score is created. One example is the Instant Composers Pool (ICP) orchestra [7], a project where each of the musicians who participate in an improvisation session creates the score of the instruments he or she plays in a special annotation system. The pieces that are created following this process are treated as "found objects" to be used in new and creative ways during performances. As Schuiling [2013] states, "this provides an opportunity to rethink the concept of compositions as actors in the musical process, and simultaneously to see the practical aspects of creativity and the importance of practically engaging with one's working material."

Considerable efforts have been devoted to finding new ways of representing western music score notation making it suitable for parsing and recognising by computer software. The most popular digital formats it can be converted to are the music form of XML (eXtensible Markup Language) and the kern format.

2.3 Articulation in Chopin's works and notation

"Put all your soul into it,
play the way you feel!"
—Frédéric Chopin

Elements of Chopin's compositions can be understood by giving a nuance of his core inspirations, the prime one being traditional Polish music; even in solo piano works, the dance impulse can be found in his Mazurka or Polonaise pieces (Thomas [1994]). Eigeldinger [1994] states that Chopin has been affected by late baroque and pre-classical composers, however his great influencer J. S. Bach imprints on Chopin's later works: contrary motion between parts, the sharing of the melody between upper part and bass, the octave imitation of an upper part by the bass are some of the common elements found in Ballade Op. 47 and Mazurka Op. 56 No. 3 among other pieces. Also Tristanesque harmonies, popular during the Baroque period, appear in Ballade Op. 47.

Shaffer [1994], searching for the characteristics that make a performance 'musical', analyses piano recordings of Chopin's Prelude Op. 28 No. 8 in F# minor and looks through the structural tension as well as the tempo and dynamics variations whether the performance "conveys an insight into the musical meaning."[p.184] For him, the combination of melodic, harmonic and rhythmic processes identify structure, while they operate on different levels or within or across levels. The results of the study show the use of a phrasing gesture where

[7] `http://www.icporchestra.com/`

there is an acceleration and increase in dynamics into a musical unit and the respected deceleration and decrease towards its boundary. Focusing on the expressive intentions that go beyond simply conveying phrase grouping, we see that related features include chord progressions, melody alterations among the phrases, and even a repeat of the same harmonization in positions where ***ff*** and ***p*** markings appear, which helps emphasise the dynamic contrast.

In terms of dynamics, Thomas [1994] refers to accents and dynamic contrasts in Mazurka pieces as emphasising the "foot-stamp or heel-clicking leap"; if they are located on the first beat they may emphasise a long-breathed four-bar phrase or a short-breathed two-bar phrase. If they are located on the second bar they are usually combined with expressive harmonic or melodic stresses or with combined with the case of having accompaniment rests on the first beat. Finally if they are located on the third bar they may either give a quiet understatement of the third movement–an example being the accompaniment rests on the first and second beat followed by a chord on the third one in Mazurka Op. 63 No.1–or emphasising the opening of a new section.

Chopin himself preferred pianos capable for depicting refined nuances rather than ones constructed based on providing sharpness and high intensity (Methuen-Campbell [1994]). Although markings such as ***ff*** and ***fff*** have been illustrated in his works, we acknowledge that "all his contemporaries agree in reporting that his dynamics did not exceed the degree of *forte*, without however losing a single bit of shading" ([Einstein, 1975, p.215]).

Other aspects of articulation have to do with pedaling and timing. A feature found in many of the Mazurka pieces is the use of one pedal-point joining usually four-bar chord progressions which produced the sense of a dominant introduction (Thomas [1994]). In the case of features related to timing, we should consider that a characteristic of Chopin's music is that it draws inspiration from singing, which translates to a style of piano playing ([Rink, 1994, pg. 216]). This style brings the strong sense of rubato, by keeping a more steady rhythm with the left hand while freeing the other by pushing forward or holding back. Carl Mikuli, One of his pupils, "complimented Chopin's rubato for its naturalness and its 'unshakeable emotional logic'" ([Khoo, 2007, p.91]).

2.4 Related studies

In this section we present a review of studies related to our aim. In order to better distinguish them, we have created the subcategories containing studies that mapped loudness with score information (Section 2.4.1), or dealt with the interconnection of changes in loudness and tempo (Section 2.4.2), or explored how expression features interact with the structure of the musical piece (Section 2.4.3).

2.4.1 Mapping between audio loudness and score

Precious little work exists that examines the mapping from score to audio loudness and vice versa, an exception being our preliminary study on the meanings of dynamic markings in performances of five of Chopin's Mazurkas Kosta et al. [2014].

A dynamic-contrast driven study followed score dynamic annotation has been presented by Geringer [1993]. This study analysed data sampled from 60 commercial recordings of choral, orchestral, and piano compositions mostly from the 19th century and from composers including L. W. Beethoven, F. Chopin, and R. Schumann among others. The dynamic changes within the music samples was recorded using a Bruel and Kjaer Graphic Level Recorder which allows the continuous measurement of relative intensity changes. Across the music excerpts sampled, the results indicated a larger dynamic range from ***p*** to ***f***, averaged to 13.42 dB change, than from ***f*** to ***p***, averaged to 11.97 dB change.

Another example is the MUDELD (MUsic Dynamics Extraction through Linguistic Description) algorithm proposed by Ros et al. [2016] that uses linguistic description techniques to categorise the dynamic label of separate musical phrases into three levels labeled as *piano* (***p***), *mezzo* (***m***), and *forte* (***f***). The process that was followed in this study was to extract the loudness information from the audio input by computing the Root Mean Square (RMS–short description in 2.1.1) of the signal and normalise the values in the range of $[0, 1]$. Then the music signal was manually segmented where each segment is represented by its average loudness value. Then, "disagreement values" were computed in order to form the distance between the segment's value label and the reference label, i.e. the music score, as well as the relevance between segments' labels.

As an initial experiment, they retrieved eight commercial recordings of Debussy's Syrinx. The musical piece was manually segmented into phrases. The outcome of their experiment indicates that the resulted labels were similar to the score indications.

Other related studies are the following two by Grachten et al. The first study is devoted to the creation of a linear basis framework (LBM) designed

to account for expressive variations as well as to model the effect of expressive markings in the score on music performances (Grachten and Widmer [2012]). Their approach is described by the creation of a number of *hand crafted* numerical descriptors, or "basis functions", to encode certain structural aspects of the score. Each of the basis functions convey one score marking and represents the activation of the marking in the scale of real values from zero–non- active, to one–active. The basis functions can be indicators for note attributes, such as *stacatto*, polynomial models for fitting the dynamics-note pitch relationship and indicators of expressive markings, either for gradual changes (e.g. *crescendo*) or constant changes (e.g. *piano*).

Essentially, LBM provides a way to determine the optimal influence of each of a set of basis functions with a set of weights to better approximate the examined expressive parameter. Two experiments were formed; the first examined how accurately expressive dynamics can be represented and predicted implementing LBM, using as dataset the Magaloff corpus (Flossmann et al. [2010]) which is constituted by recordings of various pieces by the same pianist on a Bösendorfer piano. The results showed that for both representation and prediction, the correlation ranges from weak to medium for many combinations of various number of basis functions and for both cases of forming the weights globally (same in all pieces) or locally (different for each piece), however having better results in the latter case. Also, the variance of the prediction is substantially lower than that of the observation, "meaning that expressive effects in the predicted performance are less pronounced." (Grachten and Widmer [2012])

The second experiment examined how well the LBM framework can reveal differences between performers by using loudness curves of commercial recordings of various pieces played by different pianists. The results showed that across pieces the variance of coefficients from the basis functions is too large to indicate direct relationships between performers and coefficients, in contrast of the observations detected within pieces. The last observation can be associated with our results presented in Chapter 7.

The second study is a machine learning approach concerned with the score-based prediction of note intensities in performed music (Grachten and Krebs [2014]). Our approach to the issue of dynamic variations is categorical, meaning that we analyse the relative dynamic changes in a musical piece as in forming a loudness level representation.

2.4.2 The interconnection of changes in loudness and tempo

Much research has focused on proposing and establishing the relationship between tempo and loudness (see, for example Widmer and Goebl [2004] and

references therein.) Simple rules for this kind of interconnection can be found in the literature, an example being "the louder the faster" by Repp [1996]. A representation of the performance changes in tempo (as Inter Onset Interval measure) and dynamics (as sound pressure level, measured in decibels) is the "performance worm", a real time implementation which displays the changes in a two dimensional space over time (Dixon et al. [2002]).

Assuming that timing and amplitude are the principal parameters the performer can vary in a piano performance, Bhatara et al. [2011] have examined which one of the two parameters serves to communicate emotion more. Their stimuli was piano performances of six nocturnes by Chopin. In an experiment where timing and amplitude were manipulated independently, they showed that timing and amplitude variations both affect emotionality judgements: "Generally, more systematic (but not random) variation in timing or amplitude translates to greater subjective ratings of emotionality."

2.4.3 Expression features used for structure extraction

Other trends for conceptualising expressiveness from the performer's perspective are based on the fact that there are many ways of interpreting a music piece utilising both compositional and expressive parameters, with respect to structural aspects such as phrase structure (Bennett [1993]). Following this direction, various techniques are used by composers as well as by performers to introduce articulation to music in order to create hierarchical grouping structures (Large and Palmer [2002]).

A few models exist to extract phrases by taking information from expressive parameters. The models include the work by hua Chuan and Chew [2007] Stowell and Chew [2012] where the changes in tempo were used to examine how well they correspond to phrase boundaries; the tempo shapes were modelled by a quadratic model and a spline curve in the first study and as a sequence of tempo arcs in the second one.

The work by Cheng and Chew [2008] presents the Local Maximum Phrase Detection method which employs information from tempo and loudness variations creating multiple layers of boundary measures. One of their findings was that loudness is more consistent as indicator of phrasing strategies than tempo.

Chapter 3

Data acquisition

This chapter describes the music material and the computational methods that form the basis of the data used for this thesis. In order to understand the range of expressive possibilities for each dynamic marking, we have required a significant number of recordings for each piece. For this study, we have drawn data from the CHARM[1] Mazurka project, which consists of known recordings of Chopin's Mazurkas amassed for the purpose of studying performance styles. For the majority of pianists, the romantic repertoire remains the chief resource for expressive possibilities (Benetti Jr. [2013]). Hence, our choice to focus on Chopin's Mazurkas.

We have created a score-beat loudness mapping from the data, that is, we have matched positions of score markings with their corresponding loudness levels in recordings. In order to obtain reliable measurements and scalable analyses, we rely on computational audio analysis tools, which despite their imperfections are becoming part of the standard equipment for empirical musicologists [Goebl et al., 2014, p. 225–233]. To speed up the labour-intensive process of annotating beat positions for each recording, we developed a heuristic for automating the alignment of multiple audio files. Loudness values are extracted and analysed using standard audio processing and statistical techniques.

The data that has been created for the purpose of the thesis can be found here: https://github.com/katkost/MazurkaBL. The files include information of score-beat time position and corresponding dynamics of the Mazurka recordings, as well as information of the position of score markings.

The chapter is organised as follows: Section 3.1 describes the dataset that corresponds to the audio and score information, Section 3.2 presents the multiple alignment strategy used to synchronize audio and score, and Section 3.3 analyses the dynamic marking information.

[1] `http://www.charm.rhul.ac.uk/index.html`, accessed 9 October 2016

3.1 Audio and score information

For the purpose of this thesis, we have focused on a part of the Mazurka dataset containing 2000 audio recordings of performances of forty-four pieces. The final number of recordings selected from each Mazurka is presented in Table 3.1. We have included the audio recordings that follow the repetitions in the score, and we have excluded the ones that do not, as well as the noisy recordings. By noising recordings, we include either the ones that have got distortions of the signal at the post-production process and old recordings as most of the cases, or the ones that are live recordings and have got additional sounds from the audience that could not be extracted.

For our main analysis, we focus on the behaviour of the following markings:

$$S = \{\, \boldsymbol{pp},\ \boldsymbol{p},\ \boldsymbol{mp},\ \boldsymbol{mf},\ \boldsymbol{f},\ \boldsymbol{ff} \,\}. \tag{3.1}$$

We note that ***pp*** occurs 63 times, ***p*** 234, ***mf*** 21, ***f*** 169, and ***ff*** 43 times, making a total of 530 markings. The ***mp*** marking does not occur in the data we analyse. The number of markings that appear in each piece individually is shown in Table 3.1.

Mazurka index	M06-1	M06-2	M06-3	M07-1	M07-2	M07-3	M17-1	M17-2	M17-3	M17-4	M24-1
# recordings	34	42	42	41	35	58	45	50	36	67	46
# markings	18	13	22	13	13	18	7	6	9	7	5
Mazurka index	M24-2	M24-3	M24-4	M30-1	M30-2	M30-3	M30-4	M33-1	M33-2	M33-3	M33-4
# recordings	56	39	54	45	50	54	55	48	50	23	63
# markings	12	7	33	8	14	25	18	5	16	4	12
Mazurka index	M41-1	M41-2	M41-3	M41-4	M50-1	M50-2	M50-3	M56-1	M56-2	M56-3	M59-1
# recordings	35	42	39	33	45	40	67	34	48	51	41
# markings	12	5	6	7	15	14	17	14	7	16	8
Mazurka index	M59-2	M59-3	M63-1	M63-3	M67-1	M67-2	M67-3	M67-4	M68-1	M68-2	M68-3
# recordings	56	56	42	62	35	31	40	42	38	48	42
# markings	8	11	9	4	17	10	13	11	12	21	8

Table 3.1: Chopin Mazurkas used in this research, the number of recordings for each one, and the number of dynamic markings that appear in each one. Mazurkas are indexed as "M<*opus*>-<*number*>."

We are aware of the fact that all the recordings are from different time periods, ranging from 1902 to the early 2000's, and that information of the score edition used by each performer is unknown. As mentioned in [Bailie, 1998, p.56], "since most of his [Chopin's] works were published in simultaneous 'first' editions in France, Germany and England, and since he also made alterations in the scores of various pupils, there are inevitably many discrepancies."

Tracing the actual score used in the preparation of each performance is unrealistic. Multiple editions of Chopin's Mazurkas exist, and the particular

edition used in each recording is not known. For the purposes of obtaining score-based dynamic markings as a basis for this study, we used the edition by Paderewski, Bronarski and Turczynski, as it is one of the most popular and readily available editions. We created an XML version of the scores and the information concerning the exact location of each dynamic marking was extracted automatically using the Music21 software package (Cuthbert and Ariza [2010]). In addition, we extracted the location of the score notations that are presented in Table 3.2.

Dynamics
Marking: p, pp, mf, f, ff, sf, fz, accent (>), crescendo, decrescendo
Text: sotto voce, dolce, dolcissimo, con anima, con forza, calando, espressivo, risoluto, leggiero, perdendosi, maestoso, gajo, smorzando
Tempo
Marking: fermata
Text: ritenuto, a tempo, Tempo I., lento, vivo, Allegro ma non troppo, Allegro, legato, legato assai, legatissimo, moderato, animato, rubato, scherzando, stretto, agitato, rallentando, tenuto

Table 3.2: Dynamic and tempo markings/text that appear in the scores.

A long '>' appears in the score edition mentioned above, which serves as an indication of an "agogic" accent: "an emphasis created by a slight *lengthening* rather than dynamic emphasis on a note or chord" Bailie [1998] (p.53). However this marking could not be in our XML edition as it is not supported by the Music21 software.

In the next section we explain how we dealt with the problem of linking the beat positions in the score with the corresponding ones in each recording.

3.2 Audio recording to score alignment

Since dynamic markings are notated in the score, their positions can be specified using the musical time axis of beats and measures. To study how a specific pianist realizes a given dynamic marking in a performance, we need to locate its corresponding position in the recording in seconds. A common way to do this it to manually annotate the position of each musical beat in each available recording, either by tapping while listening to the music (Sapp [2011]) or by using specialized tools such as Sonic Visualiser[2]. While manual annotations are typically quite reliable and accurate, creating them is a very time consuming process. For example, for this research, the manual annotation and correction

[2] `http://sonicvisualiser.org/`, accessed 23 September 2016

by inspecting the spectrogram of a single recording of Mazurka Op.6 No. 2 took 35 minutes on average.

To automate much of this annotation process, one can employ computational music alignment methods. Given a beat position in one version of a piece of music, such synchronization methods automatically locate the corresponding position in another version. In this way, for each piece, we only need to annotate a single recording, as we can use the automatically computed alignments to find, for each beat position in the annotated recording, the corresponding position in another recording. We call this recording as the *reference audio* for each Mazurka. Then the beat positions have been transferred automatically to the remaining recordings using a multiple performance alignment heuristic that is described in Section 3.2.2. The multiple performance alignment heuristic employs the pairwise alignment algorithm by Ewert et al. [2009], which is based on Dynamic Time Warping (DTW) applied to chroma features. This pairwise alignment technique extends previous synchronization methods by incorporating features that indicate onset positions for each chroma. The authors report a significant increase in alignment accuracy resulting from the use of these chroma-onset features and the average onset error for piano recordings is 44 milliseconds.

While alignment errors and corresponding inaccuracies in the derived annotations cannot be completely avoided, the synchronization enables the re-use of manually created annotations for a relatively small number of recordings to efficiently mass-annotate large databases. Some audio files may result in better alignment accuracies than others though. In order to achieve the best choice of a *reference audio*, we created our ground truth database, which is the manually detected score beat positions of Mazurka Op. 6 No. 2, find the best *reference audio* candidate, then create a heuristic method to automatically detect it so that we can apply the same method on the other Mazurkas.

As it has been mentioned in 2.3, the melody may be displaced from the corresponding beat position in the accompaniment so as to convey fluidity of expressive timing. As a rule, in our manual annotations, we have chosen to follow the melody line so as to capture the lyricism of the rubato.

The goal of the multiple performance alignment heuristic is to optimize the choice of a *reference audio* with which we can obtain better alignment accuracies than with another audio file. In order to understand the characteristics of such an audio, in Section 3.2.1 we present the analysis of the *reference audio* identification, and in Section 3.2.2 we present the heuristic method to detect it.

3.2.1 Optimizing *the reference audio* choice

For this section, we use as ground truth the manual annotations of score-beat positions in forty-two recordings of Mazurka Op. 6 No. 2, and our goal is to detect the audio file (*reference audio*) that predicts most accurately the score-beat positions of the other recordings.

The following steps have been implemented. We removed silences in the beginnings and ends of the recordings by discarding any audio at the beginning and end for which the amplitude value was < 0.002. Next, we located 177 beat positions per recording, excluding the beats where no notes were played, which were 11 in total. We implemented the algorithm described in Ewert et al. [2009] for audio-to-audio alignment and the annotations from each candidate reference audio recording were used for detecting the positions for the others in a pairwise fashion.

Let n be the number of recordings. We get in total $n \times (n-1)$ new annotations generated from all the candidate reference audio files. To determine the audio that performed best in providing the alignments with the lowest error in accuracy, we compared the alignment results to our ground truth. The Jarque-Bera test showed that not all the groups of values that included the alignment errors for each pair of recordings followed the normal distribution, hence every alignment result is described by the median error for each alignment pair. For each recording, we thus arrive at $n-1$ median error values. Then, in these new groups of values we implemented the non-parametric Friedman test, where the small p-value ($\mathrm{p} = 3.1546 \cdot 10^{-31}$) indicates that at least one column's sample median is significantly different from the others. The multiple comparison test reveals the audio with the lowest median error value for being the best *reference audio* (bold vertical line in Figure 3.1); as shown in Figure 3.1 Sztompka's recording has the lowest median error value, followed very closely by the median error value of Kiepura's recording.

3.2.2 *Reference audio* detection heuristic

The pairwise alignment algorithm creates a match between two audio files, say i and j, using dynamic time warping. The matched result is presented in the form of two column vectors $\mathbf{p}_i$ and $\mathbf{q}_j$, each with m entries where m depends on the two recordings chosen, i and j. Each vector presents a nonlinear warping of the chroma features for the corresponding audio file, and represents the timing difference between the two recordings. A pair of entries from the two vectors gives the indices of the matching time frames of the two audio files. We compute the Euclidean distance between each pair of the dynamic time warped audio files

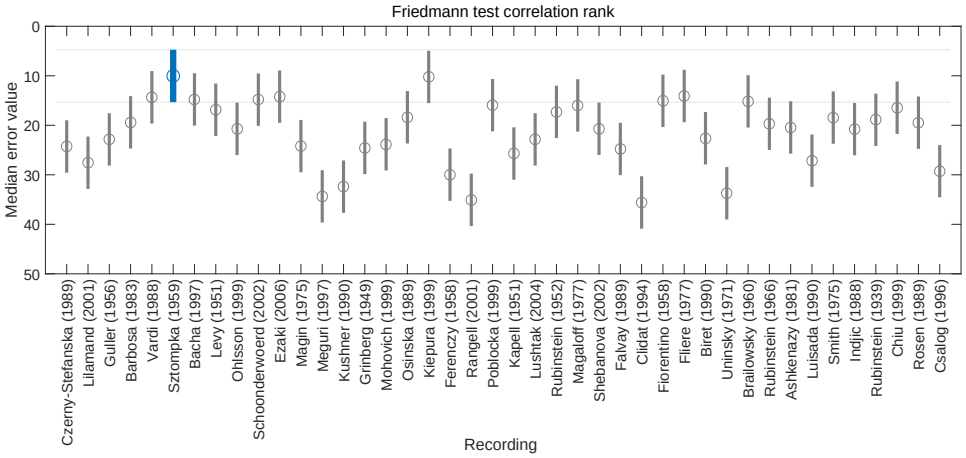


Figure 3.1: Comparison of the groups constituted by the median error values per recording, implementing the Friedmann test correlation rank for the 42 recordings of Mazurka Op.6 No. 2. Each point of the x-axis presents one recording–"*name of pianist (year of recording)*". The recording with the lowest correlation rank is highlighted in bold (pianist: Sztompka, year of recording: 1959) and it is identified as the *reference audio* for the specific Mazurka.

as follows:

$$d_{i,j} = \sqrt{\sum_{k=1}^{m}(q_{j,k} - p_{i,k})^2}, \qquad \forall i \neq j, \tag{3.2}$$

where $m \in \mathbb{N}$ is the size of the vectors. In this way, each audio has a profile corresponding to its alignment to all other audio recordings which is $\mathbf{d}_i = [d_{i,j}]$. The average value of all the alignment accuracies for the i^{th} recording in relation to the remaining ones is $\overline{\mathbf{d}}_i$.

We consider the best reference file to be one that minimizes the average distance to other audio files and without extreme differences from more than two other audio recordings as measured by the norm distance; in this way, after exploring alternative values of outliers, a quick test using the method provided above on Mazurka Op. 6 No. 2 found the same reference audio as did the formal analysis in Section 3.2.1. Mathematically, the problem of finding the reference audio can be expressed as one of solving the following problem:

$$\begin{aligned} & \min_i \overline{\mathbf{d}}_i \\ \text{s.t.} \quad & \#\left\{j : |d_{i,j}| > q_3(\mathbf{d}_i) + 1.5[q_3(\mathbf{d}_i) - q_1(\mathbf{d}_i)]\right\} \leq 2, \end{aligned}$$

where $q_\ell(\mathbf{d}_i)$ is the ℓ-th quantile of $\mathbf{d}_i$, and the left hand side of the inequality uses an interquartile-based representation of an outlier. The *reference audio* is then given by $\arg\min_i \overline{\mathbf{d}}_i$.

We manually adjust the beat positions for the *reference audio* and we infer the beat positions of the remaining recordings by using the alignment method as mentioned above. In order to evaluate the method, we compared the derived beat positions with the manually annotated positions of the *reference audio* for forty-four recordings of the Mazurka Op. 6 No. 2. The average error was 37 milliseconds.

3.3 Loudness information of dynamic markings

In Section 2.1.2 we address the way we get the loudness information from the audio recordings, which includes the extraction and normalisation of the sone values along the music pieces. In the dataset described in Section 3.1, ***pp*** occurs 63 times, ***p*** 234, ***mf*** 21, ***f*** 170, and ***ff*** 43 times, giving a total of 530 dynamic markings. The loudness value corresponding to each marking is estimated as the average sone value at the aligned beat position and the two subsequent beats. More formally, if $\{y_n\} \in \mathbb{R}$ is the sequence of loudness values in sones for each score beat indexed $n \in \mathbb{N}$ in one piece, then the loudness value associated with the marking at beat b is

$$\ell_b = \frac{1}{3}(y_b + y_{b+1} + y_{b+2}). \tag{3.3}$$

Sometimes the actual change in response to a new dynamic marking does not take place immediately, but is rather observed in subsequent beats. It is clear from the data that loudness varies considerably between transitions of dynamic markings. Thus, we additionally aim to have the smallest window possible to capture dynamic changes that occur in response to a marking. Consequently, we have chosen a three-beat window as the loudness time frame for this study, because it corresponds to the ternary time signature of 3/4 found in Mazurkas. The markings that are analysed are marked in Fig. 3.2—bottom as black x's.

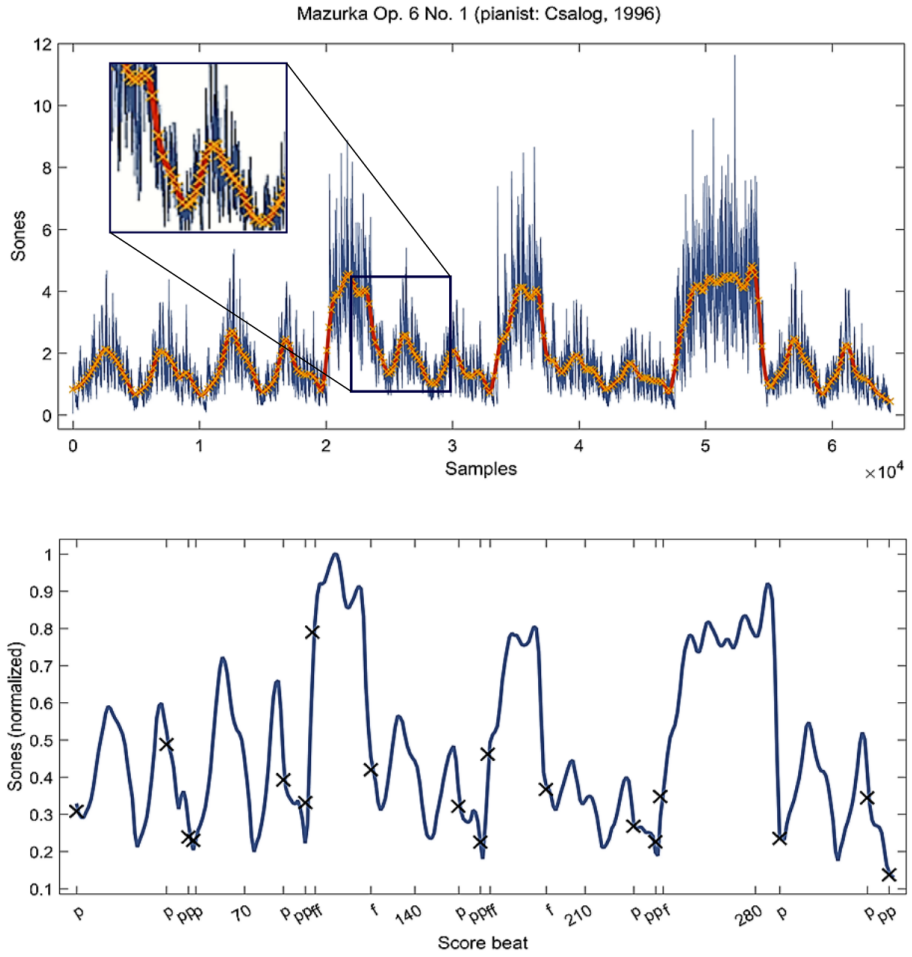


Figure 3.2: Loudness representation of a single Mazurka Op. 6, no. 1 recording (pianist: Csalog, recording year: 1996). Top: Smoothed curve (orange) of raw data of sone values (blue); the score beat positions are highlighted by yellow x's (zoomed-in representation of excerpt at top-right). Bottom: Normalized score beats curve; the averaged final loudness values for each marking are highlighted by x's.

Chapter 4

Dynamics and relativity: Is *piano* always *piano*?

Summary: This chapter focuses on the meaning and practical manifestations of expressive markings in the music score, specifically those markings that correspond to loudness levels—such as ***p*** (*piano*), ***mf*** (*mezzo forte*), and ***ff*** (*fortissimo*). We present results showing how the absolute meanings of dynamic markings change, depending on the intended (score defined) and projected (actual performed) dynamic levels of the surrounding musical context. The analysis of recorded performances shows different realisations of the same dynamic markings throughout a recording of a piece of music. Reasons for this phenomenon include the score location of the markings, such as the beginning of a piece, and the marking's location in relation to that of previous ones. Observations show that more often than not, the transition between the two markings is more consistent when pianists move from a louder to a softer marking, or move between markings that both represent a high intensity, or when the markings show high levels of contrast. For markings that appear in the score more than once, most of the time, there is a significant difference in the ways they are interpreted in a recording.

Dynamic levels in musical expressivity are represented in the score by markings such as $\boldsymbol{p}$ (*piano*, meaning soft), and $\boldsymbol{f}$ (*forte*, meaning loud). These symbols are interpreted in performance and communicated through the varying of loudness levels. If we were to order the set of dynamic markings in increasing loudness, yielding the following sequence, termed *Ordinal Loudness Sequence* (OLS) in the following,

$$\boldsymbol{pp} < \boldsymbol{p} < \boldsymbol{mp} < \boldsymbol{mf} < \boldsymbol{f} < \boldsymbol{ff}, \tag{4.1}$$

we would assume that the loudness level corresponding to a low-rank marking in the OLS is lower than the level corresponding to a higher-rank marking in the OLS. As we shall see this is not the case. The expected loudness of dynamic markings may be superseded by their local context so that a $\boldsymbol{p}$ dynamic might objectively be louder than a $\boldsymbol{f}$ at another part of the piece.

In this chapter we examine the dynamic markings in the OLS and we consider two types of relativity in the interpretation of these markings, one dealing with the overall loudness level of a particular marking throughout the times that is present in the score of one piece, and the other dealing with the distinct loudness level of a particular marking at a single position in the score. In order to gather the information we need for the first type of relativity, for each Mazurka, we compute the average loudness levels over all recordings for all markings that are the same, and compare the resulting averages. About the second type of relativity, for each Mazurka, we compute the change in loudness for each pair of different consecutive markings. For example, assume that a piece has the sequence of markings ($\boldsymbol{p}_1$, $\boldsymbol{p}_2$, $\boldsymbol{f}_1$, $\boldsymbol{ff}$, $\boldsymbol{p}_3$, $\boldsymbol{f}_2$), in the order as they appear in the score. Then for each performance recording we get the loudness values that correspond to each of the markings. For the first type of relativity, each recording has a loudness level of the $\boldsymbol{p}$'s which is computed by taking the average of the loudness values that correspond to the three p positions, and the same process is followed for all the remaining markings. Then these average values are compared (more details about the comparison method in Section 4.1.1.) For the second type of relativity, for each recording we compare the loudness change between the different consecutive markings in pairs ($\boldsymbol{p}_2$, $\boldsymbol{f}_1$), ($\boldsymbol{f}_1$, $\boldsymbol{ff}$), ($\boldsymbol{ff}$, $\boldsymbol{p}_3$), and ($\boldsymbol{p}_3$, $\boldsymbol{f}_2$), individually.

Additionally we analyse how the distance of the markings in OLS affects the transition from one loudness level to another, and finally we observe the different patterns of the manifestation of the same markings as they appear in a score sequence of one piece.

More precisely, we seek to answer the following questions:

(i) Do the aggregate dynamics of pairs of dynamic markings conform to the expected order?
(ii) Do pairs of individual instances of dynamics at markings conform to the expected order?
(iii) How does the average loudness change vary related to the ordinal distance between dynamic markings?
(v) How are the same dynamic markings realised as they recur throughout the piece?

Another challenge of conducting systematic research on expressive parameters is the lack of audio data and annotations for analysis and for training. In order to understand the range of expressive possibilities for each dynamic marking, we required a significant number of recordings for each piece. For this chapter, we have used the data set presented in Chapter 3, which is the creation of a score-beat loudness mapping from the audio recordings by matching positions of score markings with their corresponding loudness levels.

This chapter is organized as follows: Section 4.1 presents the results of the studies where each subsection corresponds to each research question displayed above, respectively. Section 4.2 presents the conclusions and further discussion.

4.1 Performed Loudness Study

4.1.1 Ordinal Loudness Sequence Preserved On Average?

In this section we explore the general behaviour of pianists' responses to the dynamic markings of the set S. More specifically, this section addresses the issue of whether the dynamic level of a softer marking could be higher on average in a recording than that of a louder marking. We define the expected rank order sequence for loudness behaviour, and use Kendall's tau rank correlation coefficient test to show that the average dynamic levels in each recording does not always abide by the expected rank ordering.

In a recording, each dynamic marking is represented by a dynamic value (in sones), as described in Section 3.3. Some markings appear in a score more than once. We compute the aggregate value for each marking in a recording by taking the average of the dynamic values whenever the same marking appears in the score. For a symbol $s \in S$ that appears n times in the score at the beat positions $\{s_1, s_2, \ldots, s_n\}$,

$$\mathrm{E}(s) = \frac{1}{n} \sum_{i=1}^{n} \ell_{s_i}. \tag{4.2}$$

We first consider for each recording the average dynamic levels E(***pp***), E(***p***), E(***mp***), E(***mf***), E(***f***), and E(***ff***) for the markings ***pp***, ***p***, ***mp***, ***mf***, ***f***, and ***ff***, respectively. It so happens that ***mp*** does not occur in the scores we consider in this study.

The results of Kendall's tau rank correlation coefficient test are summarized in Table 4.1 for a given Mazurka. Each τ value shows the agreement between pairwise mean dynamic values and their expected ordering according to the OLS. A value of 1 indicates perfect rank agreement with the OLS, and a value of -1 indicates the converse. Patterns contrary to the expected are analysed in the following subsections. More specifically, in the next subsections we further examine the values in bold (and red) that are those for which $\tau < 0$.

The graphs below show the dynamics at all beats, not only the ones at the dynamic markings. The variety of curves in the background shows the diversity of dynamic responses from the pianists at the same score beats (x axis). The graphs also show how the responses vary for each marking. Wherever a symbol in a pair of markings is present, a boxplot indicates the spread of loudness values at that position. The eye in the middle of each boxplot marks the median, and the top and bottom edges indicate the 75^{th} and 25^{th} percentiles; the whiskers extend to the most extreme data points excluding the outliers.

Mazurka cases where E(*f*) > E(*ff*)—marking pair *f*-*ff* is only one with negative τ value.

There are two cases, Mazurka Op. 30 No. 3 in D♭ major, and Mazurka Op. 68 No. 3 in F major, where there is only one negative τ value, which occurs at the marking pair ***f***-***ff***, meaning that a significant number of individual recordings have E(***f***) > E(***ff***). Fig. 4.1 shows the dynamic values for all recordings of these Mazurkas in score time, and the distribution of the pianists' responses to the markings f and ff in particular.

In Mazurka Op. 68 No. 3, we notice that the average dynamic of the first ***f*** is louder than that of any other marking, including that of ***ff***. One reason for this much higher first ***f*** may be its location on the first beat of the piece, the result of giving extra emphasis to the beginning.

In Figure 4.2, the first two bars show the score area where a f marking is located, while the last two bars show another ***f*** marking; it is noticeable that although both locations belong to the start of effectively the same two-bar phrase, the slight difference in the patterns makes pianists use different fingerings, which may add to the diverse response. Another point of note is that before the second ***f*** marking there is a twelve-bar new phrase at the key of B♭ major (not shown).

In Mazurka Op. 30 No. 3 the average loudness value of the three ***ff***'s is

Mazurka \ Pair	{*pp*,*p*}	{*pp*,*mf*}	{*pp*,*f*}	{*pp*,*ff*}	{*p*,*mf*}	{*p*,*f*}	{*p*,*ff*}	{*mf*,*f*}	{*f*,*ff*}
M06-1	0.941		0.941	1		0.353	1		1
M06-2						1			
M06-3	0.952		1	1		1	1		0.714
M07-1	0.707		1	1		1	1		1
M07-2							1		
M07-3	1		1	1		0.966	0.966		0.828
M17-1						0.956			
M17-2	0.480		0.880			0.960			
M17-3					0.389				
M17-4	0.075			1			1		
M24-1	1	1			0.696				
M24-2	0.964		0.964			0.536			
M24-3					0.536				
M24-4	0.889		1	1		0.926	1		1
M30-1						1			
M30-2						0.920			
M30-3	0.407		1	1		0.963	0.815		**-0.333**
M30-4	0.636		1	1		1	1		0.927
M33-1						0.917			
M33-2			1	1					0.960
M33-1						0.826			
M33-4	**-0.556**		0.810			1			
M41-1	0.829		1	1		1	1		1
M41-2					0.429	**-0.190**	1	**-0.476**	1
M41-3						0.949	1		0.949
M41-4	0.515	1	1		0.879	1		0.939	
M50-1					**-0.200**	0.911		0.911	
M50-2						0.750			
M50-3	0.134		1	**-0.672**		1	**-0.940**		**-1**
M56-1					1	1		0.176	
M56-2						0.833			
M56-3					**-0.255**	1		0.961	
M59-1						1			
M59-2	1		1	1		0.964	1		0.893
M59-3						1			
M63-1	0.952		1			0.952			
M63-3						1			
M67-1	0.886	**-0.429**	0.829	0.943	**-0.886**	0.086	0.429	0.943	0.714
M67-2	**-0.484**	**-0.290**	0.032		0.419	0.613		0.226	
M67-3	**-1**		0.450	0.900		1	1		0.900
M67-4					0.286	0.952		0	
M68-1						0.947			
M68-2	**-0.125**	0.958	1		1	1		1	
M68-3						0.952	1		**-0.095**

Table 4.1: The Kendall's τ values for pairwise comparisons of the mean dynamic values that correspond to the dynamic markings for each Mazurka. The negative values are highlighted in bold and red coloured.

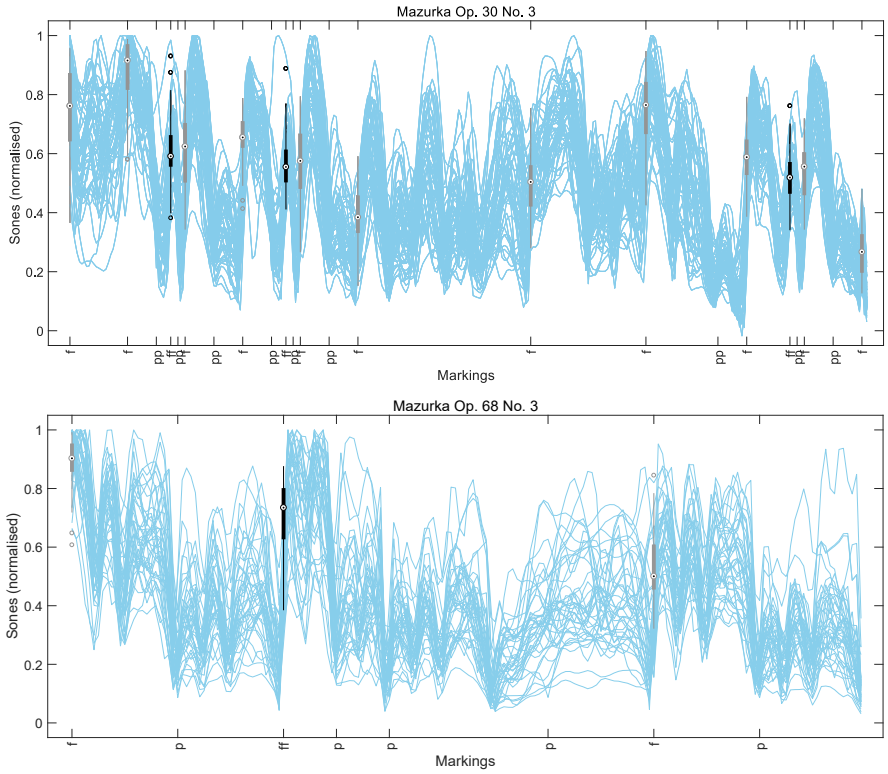


Figure 4.1: The Mazurka cases where only the ***f*-*ff*** pair has negative τ value. The dynamic values of the markings belonging to the pair in Mazurka Op. 30 No. 3 (top), and Mazurka Op. 68 No. 3 (bottom) are presented as box plots. Positions of other markings $s \in \mathrm{S}$ appear in x-ticks as in score sequence. At the background the curves represent the score—beat dynamic values per recording.

less than that for the eleven ***f***'s in a significant number of recordings. The position of the ***ff***'s in the score offer an explanation for this result. The three ***ff*** markings belong to the same repeated phrase which is shown in Figure 4.3 and they are located between two ***pp*** markings. Their duration is one score bar. In every recording, the average ***pp*** is significantly softer than the average

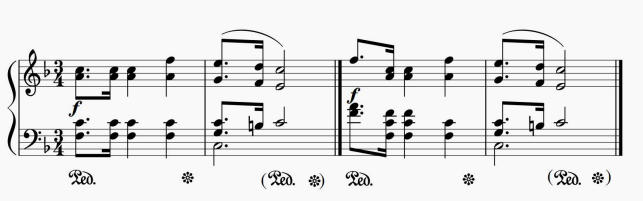


Figure 4.2: The repeated phrase where the ***f*** appears in Mazurka Op. 68 No. 3 (left part: score beats 1–2, right part: score beats 133–134.)

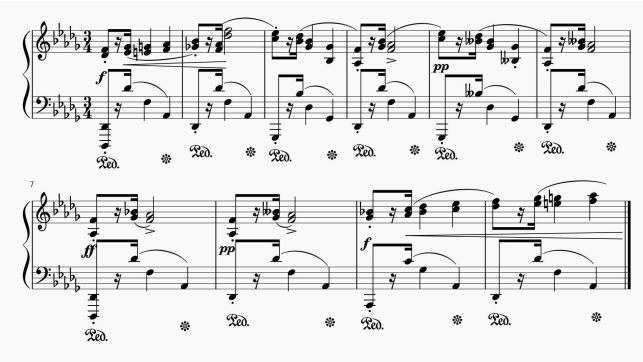


Figure 4.3: The repeated phrase where the ***ff*** appears in Mazurka Op. 30 No. 3 and its relation to the neighbouring markings ***pp***, and ***f*** (score beats 25–54, 73–102, and 283–312.)

ff ($\tau = 1$). This means that there is indeed a loudness change during the score sequence ***pp-ff-pp***. The average ***f***'s are louder, perhaps because the ***ff*** did not have to be so much louder than the ***pp***'s to make a large contrast.

The response to the ***f***'s nearest the ***pp-ff-pp*** trios are louder on average than the fs's, as shown in Figure 4.3. Reasons for this include the ***crescendo*** ($<$) marking right after each ***f***, an instruction to increase in loudness, and the fact that almost all pianists apply a crescendo right after the second ***pp*** in this excerpt, which amplifies the dynamic level of the ***f*** that follows.

Mazurka cases where E(*pp*) > E(*p*)—marking pair *pp-p* is only one with negative τ value.

There are three cases, Mazurka Op. 33 No. 4 in B minor, Mazurka Op. 67 No. 3 in C major, and Mazurka Op. 68 No. 2 in A minor, where there is only one negative τ value which is at the marking pair ***pp-p***, meaning that a significant number of individual recordings have E(***pp***) > E(***p***). Figure 4.4 shows the dynamic values for all recordings of these Mazurkas in score time, and the distribution of the pianists' responses to the marking pairs ***pp-p*** in particular.

In Mazurka Op. 33 No. 4, there is one ***pp*** marking played louder than the two ***p*** markings on average. One explanation is the loudness progression throughout the duration of the ***pp*** (in the gap between the ***pp*** and ***p*** markings), as it can be observed in the top plot of Figure 4.4. Many bars separate the ***pp*** and the ensuing ***p***, as shown in Figure 4.5, the first half of these intermediate bars are louder than the second half, resulting in a significant loudness drop that leads into the ***p*** marking.

In Mazurka Op. 67 No. 3, there are four ***pp***'s in a row which are louder on

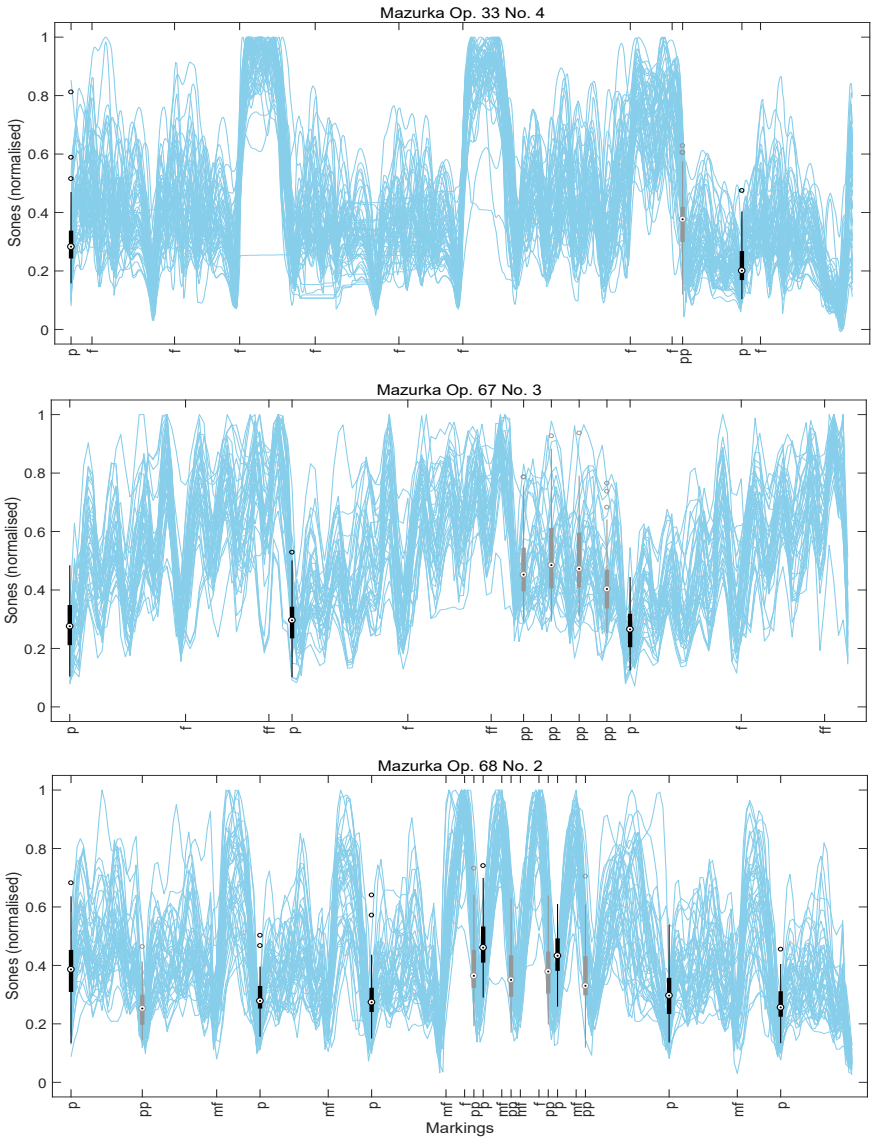


Figure 4.4: The Mazurka cases where only the ***pp***-***p*** pair has negative τ value. Box plots of the dynamic values of the markings belonging to the pair in Mazurkas Op. 33 No. 4, Op. 67 No. 3, and Op. 68 No. 2 are presented. Positions of other markings $s \in \mathrm{S}$ appear in x-ticks as in score sequence. At the background the curves represent the score–beat dynamic values per recording.

average than the three ***p***'s. More specifically, the three ***p***'s are located at the beginning of the same phrase which is repeated, a fact that likely explains the very similar loudness level at the location of all the ***p***'s. As Figure 4.6 shows, every ***pp*** is preceded by a sustained note in ***sf***, which might be the reason for the higher loudness level.

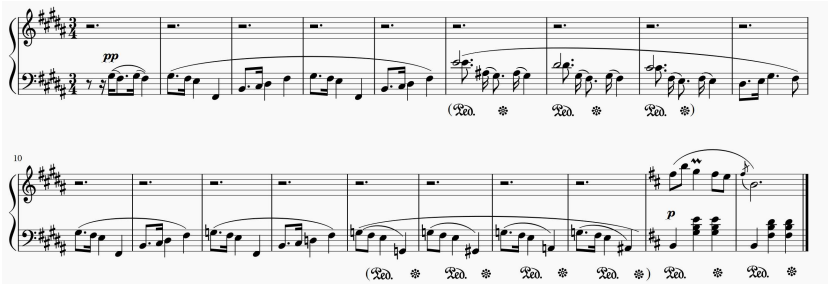


Figure 4.5: The position of the ***pp*** marking in Mazurka Op. 33 No. 4. The response of the recordings at this position is higher than the response at the position of the ***p*** marking on average.

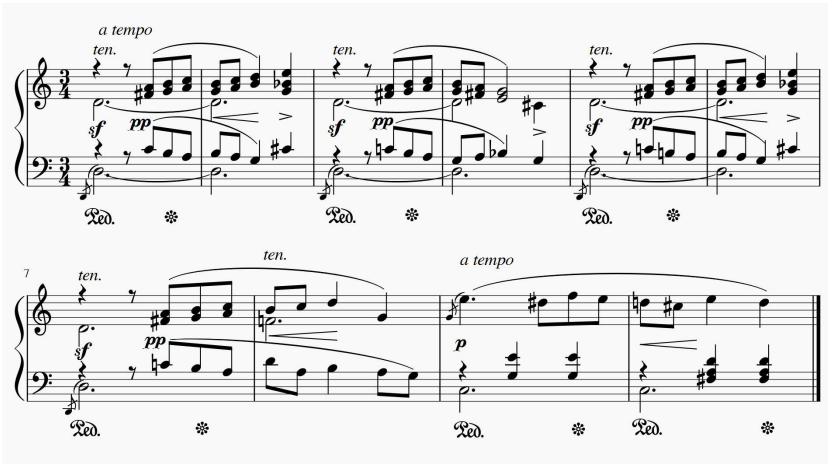


Figure 4.6: The ***pp*** markings in Mazurka Op. 67 No. 3. The response of the recordings at these positions is higher than the response at the positions of the ***p*** markings on average.

In Mazurka Op. 68 No. 2, there is a middle part, where there are four closely clustered humps in Figure 4.4 corresponding to a repeated section, where there are ***pp*** and ***p*** markings. In this middle part, the ***pp***'s are indeed softer than the ***p***'s on average. However, these ***pp***'s are on average louder than the average of the remaining ***p***'s.

Mazurka cases where E(*p*) > E(*mf*)—marking pair *p*-*mf* is only one with negative τ value.

There are two cases, Mazurka Op. 50 No. 1 in B minor, and Mazurka Op. 56 No. 3 in C major, where there is only one negative τ value which is at the marking pair ***p*-*mf***, meaning that a significant number of individual recordings have E(***p***) > E(***mf***). Figure 4.7 shows the dynamic values for all recordings of this Mazurka in score time, and the distribution of the pianists' responses to

the marking pairs ***p*-*mf*** in particular.

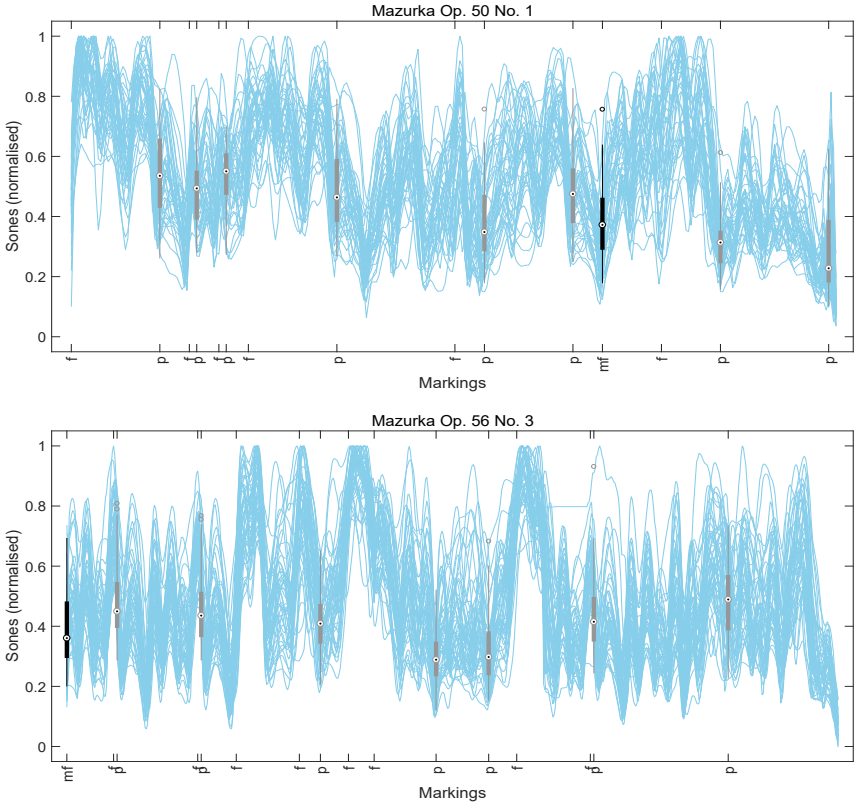


Figure 4.7: The Mazurka cases where only the ***p*-*mf*** pair has negative τ value. The dynamic values of the markings belonging to the pair in Mazurka Op. 50 No. 1 (top), and Mazurka Op. 56 No. 3 (bottom) are presented as box plots. Positions of other markings $s \in \mathrm{S}$ appear in x-ticks as in score sequence. At the background the curves represent the score–beat dynamic values per recording.

In the case of Mazurka Op. 50 No. 1 the loudness values at the single ***mf*** are lower than the ones at the ***p***'s on average. Observe Figure 4.7 that the ***p*** marking that precedes the ***mf*** is shown to be louder, but there is a loudness drop in the intervening measures, meaning that there is indeed an increase in loudness at the ***mf*** in most recordings. Figure 4.8 shows the score position for both markings. The loudness value decreases as the phrase closes before the ***mf***, and is locally ascending around the ***mf***, but this change is not captured when considering only loudness at the markings.

In the case of Mazurka Op. 56 No. 3, the single ***mf*** marking is located at the beginning of the piece and it is reported as being less loud than the average of the ***p*** markings. In all but one case, the marking preceding a ***p*** is ***f***, and in three of these cases, the ***f*** is in the bar immediately before the ***p***. When the

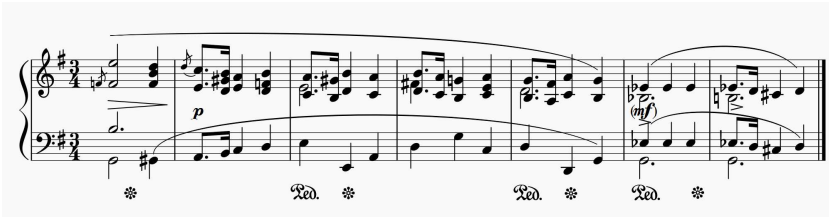


Figure 4.8: The ***mf*** marking in Mazurka Op. 50 No. 1 and its relation with the preceded ***p*** marking.

markings are very close, there is too little time to realise a ***p*** and the effect of the drop in loudness can only be detected in later bars.

The case of Mazurka Op. 41 No. 2: E(*p*) > E(*f*) and E(*mf*) > E(*f*).

In Mazurka Op. 41 No. 2, E(***p***) > E(***f***), and E(***mf***) > E(***f***). For the ***p***-***f***, and ***mf***-***f*** pairs in this Mazurka, the τ value is negative, meaning that a significant number of individual recordings have E(***p***) > E(***f***), and E(***mf***) > E(***f***). Figure 4.9 shows the dynamic values for all recordings of this Mazurka in score time, and the distribution of the loudness values at the markings ***p***, ***mf***, and ***f*** in particular.

Note that the first ***f*** marking is generally played softer than the second one, and this affects the average level of the response to this specific marking. In order to explore this behaviour, we present in Figure 4.10 the location of the two ***f***'s as they appear in the score. The first one is preceded by the text indicator ***(dim.)***, which lowers the threshold for change perception at the ***f***. This is followed by a short ***Crescendo***, which allows the ***f*** to expand in loudness after the marking; in this case, the pianist must allow room for this expansion. The section concludes with the text indicator ***(dim.)***. Next, we examine the next two observations following the ***f***. The ***p*** marking, which is relatively loud, is located at the beginning of the score, where pianists are more likely to place more emphasis on the opening notes. Also, the ***mf*** marking, which can be seen at the bottom part of Figure 4.10, is inside parentheses, which suggests freedom on how it could be interpreted.

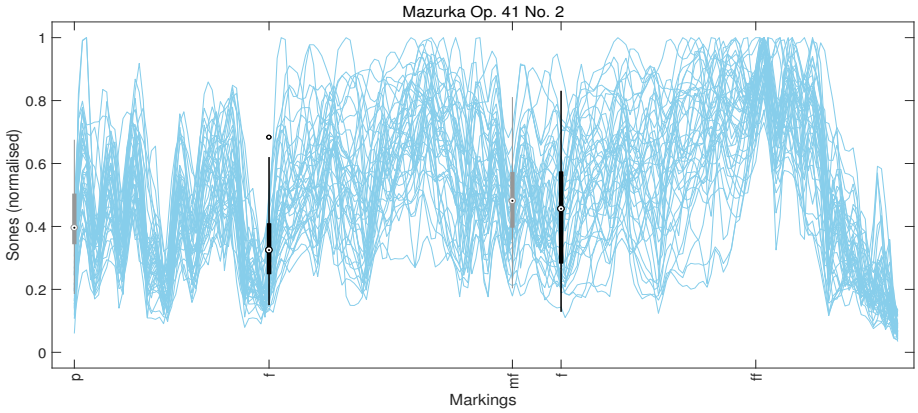


Figure 4.9: Box plots of the dynamic values of the markings belonging to the pairs ***p*-*f***, and ***mf*-*f*** in Mazurka Op. 41 No. 2. Positions of other markings $s \in$ S appear in x-ticks as in score sequence. At the background the curves represent the score–beat dynamic values per recording.

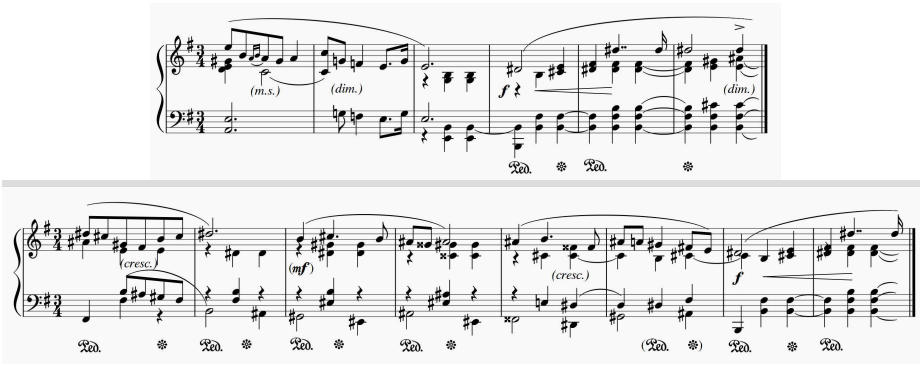


Figure 4.10: The location of the first ***f*** marking (top), and the location of the second ***f*** marking in Mazurka Op. 41 No. 2 (bottom), score—beats 40–57, and 103–126, respectively.

The case of Mazurka Op. 50 No. 3: E(*pp*) > E(*ff*), E(*p*) > E(*ff*), and E(*f*) > E(*ff*).

In Mazurka Op. 50 No. 3, an irregularity that is related to the interpretation of the ***ff*** marking is observed. More specifically, the single ***ff*** marking, located at the penultimate score measure, is reported as lower in loudness level than the average loudness of the ***pp***'s, ***p***'s, and ***f***'s, respectively. The counter-intuitive loudness transitions (negative τ values) for the pairs ***pp***–***ff***, ***p***–***ff***, and ***f***–***ff*** is caused by the fact that the single ff at the end is played extraordinarily soft. Figure 4.11 shows this finding, as well as the score progression for extremely well-behaved (τ equals 1) loudness pairs ***pp***–***f***, and ***p***–***f***, and the relatively small agreement (τ value equals to 0.134) for the pair ***pp***–***p***.

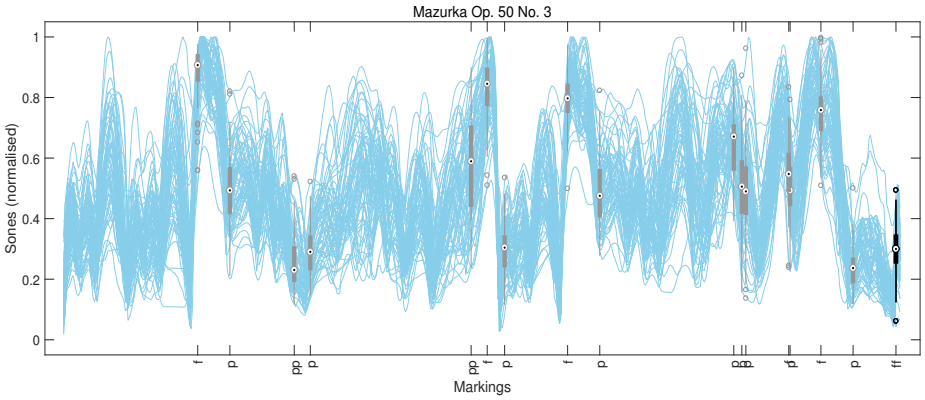


Figure 4.11: Box plots of the dynamic values of the markings belonging to the pairs ***pp***—***ff***, ***p***—***ff***, and ***f***—***ff*** in Mazurka Op. 50 No. 3. Positions of other markings $s \in$ S appear in x-ticks as in score sequence. At the background the curves represent the score–beat dynamic values per recording.

The case of Mazurka Op. 67 No. 1: E(*pp*) > E(*mf*), and E(*p*) > E(*mf*).

In Mazurka Op. 67 No. 1, the marking pairs ***pp***-***mf***, and ***p***-***mf*** have negative τ values, meaning that a significant number of individual recordings have E(***pp***) > E(***mf***), and E(***p***) > E(***mf***). Figure 4.12 shows the distribution of the loudness levels throughout the recordings at the positions where the specific markings appear in the score. As a side note we should highlight that this is the only Mazurka from the ones we test which includes all the markings $\in S$ at least once.

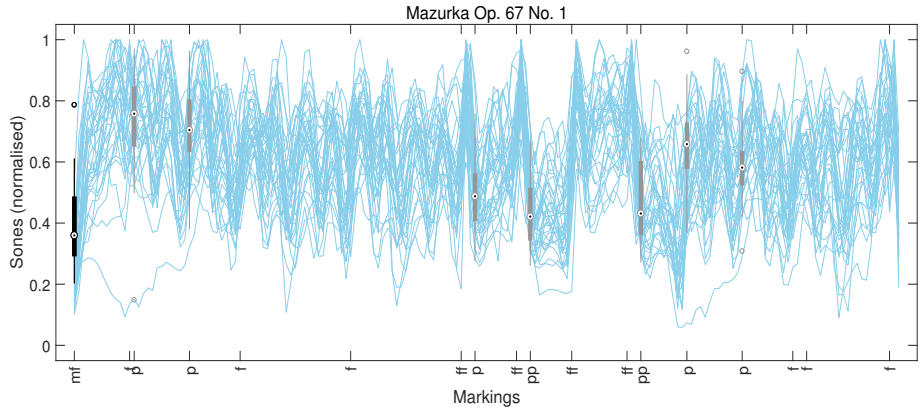


Figure 4.12: Box plots of the dynamic values of the markings belonging to the pairs ***pp***—***mf***, and ***p***—***mf*** in Mazurka Op. 67 No. 1. Positions of other markings $s \in$ S appear in x-ticks as in score sequence. At the background the curves represent the score–beat dynamic values per recording.

In order to explore the reason why the ***p*** markings seem louder in average than the ***mf*** marking, we could focus on the fact that all ***p***'s, except from the third one, are followed by a ***Crescendo*** marking directly after the beat in which they appear. Consider the first ***p*** marking, which is followed by a Crescendo, and preceded by a f marking one-score beat away. The ***pp*** markings are each preceded by a ***ff*** marking two score-beats away, which may affect the overall higher response to the ***pp***'s, although the balance of E(***pp***) < E(***p***) is kept in most of the recordings.

The case of Mazurka Op. 67 No. 2: E(*pp*) > E(*p*), and E(*pp*) > E(*mf*).

In Mazurka Op. 67 No. 2, the marking pairs ***pp*-*p***, and ***pp*-*mf*** have negative τ values, meaning that a significant number of individual recordings have E(***pp***) > E(***mf***), and E(***p***) > E(***mf***). Figure 4.13 shows that the average response to ***pp*** is louder than the response to ***mf***, and to ***p*** in most of the recordings. Both pp's location in a repeated phrase in the score is shown in Figure 4.14. Observing how the loudness changes have been reported throughout the recordings, we notice that the second ***p*** in the repeated phrase is softer than the ***pp*** although there is a ***crescendo*** marking in the previous measure. This may be related to the fact that in most recordings the pianists choose to emphasise the phrase closing, which is where the specific marking is located. Then the ***mf*** marking that follows is louder than the ***p***, but it fails to supersede the robust loudness values of the ***pp***'s.

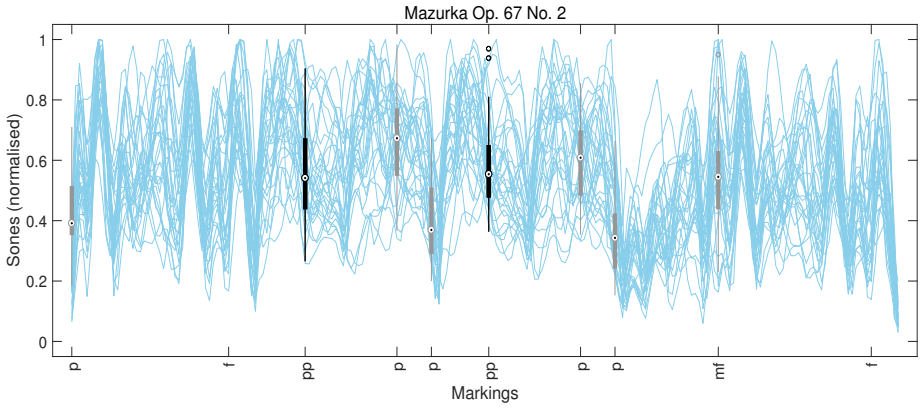


Figure 4.13: Box plots of the dynamic values of the markings belonging to the pair ***pp*–*p*** and the pair ***pp*–*mf*** in Mazurka Op. 67 No. 2. Positions of other markings $s \in$ S appear in x-ticks as in score sequence. At the background the curves represent the score–beat dynamic values per recording.

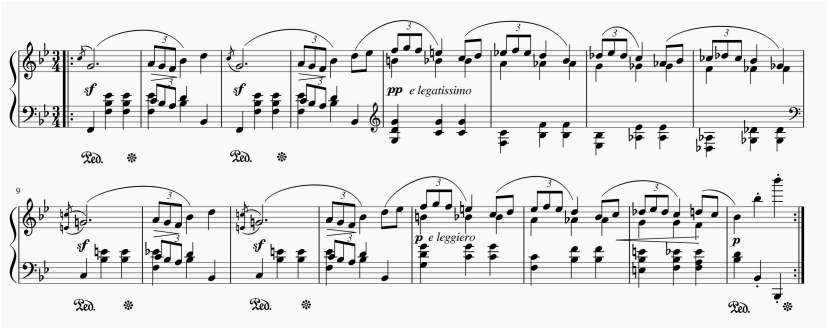


Figure 4.14: The repeated phrase in Mazurka Op. 67 No. 2 which includes the ***pp*** marking, and the two ***p*** markings that follow up.

In summary, the case studies in this section have shown that we cannot assign fixed thresholds for the different dynamic markings. Rather, the important factors are the relative loudness of neighbouring markings, the inter-relations of nearby markings and other score information, the structural location of the markings, the local shaping of loudness, and generally the creative license of the performer. The results also suggest that how well the use of loudness complies with the OLS is related not so much to the number of different markings present, but rather the local context and memory.

In order to deepen the understanding of the way in which markings are expressed, the next section focuses on pair-wise marking comparisons as represented by dynamic change values for each pair of consecutive markings.

4.1.2 Is the Ordinal Loudness Sequence Preserved In Pairwise Instances?

In this section we examine individual responses to consecutive pairs of distinct dynamic markings. We investigate consecutive responses to markings, for example, (***p***, ***f***) and (***f***, ***p***), that indicate an ascending or descending dynamic change, highlighting results contrary to the OLS, meaning that the recordings follow the loudness change from one dynamic marking to another in the order as it is defined in Equation 4.1. In the analysis of this section, we consider only pairs of distinct consecutive markings.

Figure 4.15 shows the log loudness ratios for pairs of distinct consecutive dynamic markings throughout the forty-four Mazurkas. For each pair of dynamic markings, the agreement with the OLS may vary depending on different expression strategies. Suppose the loudness of the $(k-1)$-th dynamic marking, say a ***p***, is expressed by a pianist as ℓ_{k-1}, and the loudness of the k-th dynamic marking, say a ***f***, is expressed as ℓ_k, then the log ratio (as summarized in Figure 4.15) is $\log(\frac{\ell_k}{\ell_{k-1}})$. The data is found to be more easily compared with the

log representation.

Observe that there are cases where the mean value of the ratios in the top (left) plot is lower than 0, meaning that a good number of recordings do not respect the OLS at those specific markings. Subsequently, there are cases where the mean value of the ratios in the bottom (right) plot is higher than 0, meaning that good number of recordings do not adhere to the OLS at those specific markings.

In Table 4.2 we show the Mazurkas that have marking pairs in which the average loudness change contradicts the OLS. These correspond to marking pairs with below 0 log ratios in the top (left) plot and those with above 0 log ratios in the bottom (right) plot of Figure 4.15. In the parentheses we present the proportion of the outlier pairs over all consecutive marking pairs present in that specific Mazurka. We do not consider the sequential pairs that consist of the same marking.

Ascending pair	Mazurka (proportion of outlier pairs)
(***pp***,***p***)	M33-4 (0.25) M41-1 (0.25) M67-3 (0.11) M68-2 (0.05)
(***p***,***mf***)	M50-1 (0.17)
(***mf***,***f***)	M41-2 (0.75)
(***p***,***f***)	M41-2 (0.75) M50-1 (0.17) M50-2 (0.14) M67-1 (0.09)
Descending pair	**Mazurka (proportion of outlier pairs)**
(***p***,***pp***)	M17-2 (0.25) M50-3 (0.08) M67-2 (0.14)
(***f***,***mf***)	M41-2 (0.75)

Table 4.2: List of marking pairs that contradict the OLS with information on Mazurkas where the pairs appear as well as their proportion with respect to the total number of pairs in that Mazurka.

Mazurka Op. 41. No. 2 has the largest proportion of outlier pairs. More specifically, the sequence of the markings in this Mazurka is {***p***, ***f***, ***mf***, ***f***, ***ff***}, and in all transitions except for the last one, the average loudness change ratio over all recordings contradicts the OLS. The unpredictability of the loudness progressions is evidenced in the drastic variations in the loudness curves of the recordings of this Mazurka as presented in Figure 4.9 in Section 4.1.1.

Moving on, Figure 4.16 shows the average standard deviation of the log loudness ratios for different marking pairs over all Mazurkas. One would expect that, as the distance between markings in the OLS increases, the ratio of the loudness change would also increase. But, this is not the case, as illustrated by the following observations.

In all cases except for the pairs ***f***–***ff*** and ***mf***–***f*** the average SD is larger in the ascending direction than the descending one. This means that, more often than not, the transition between the two markings is more consistent when

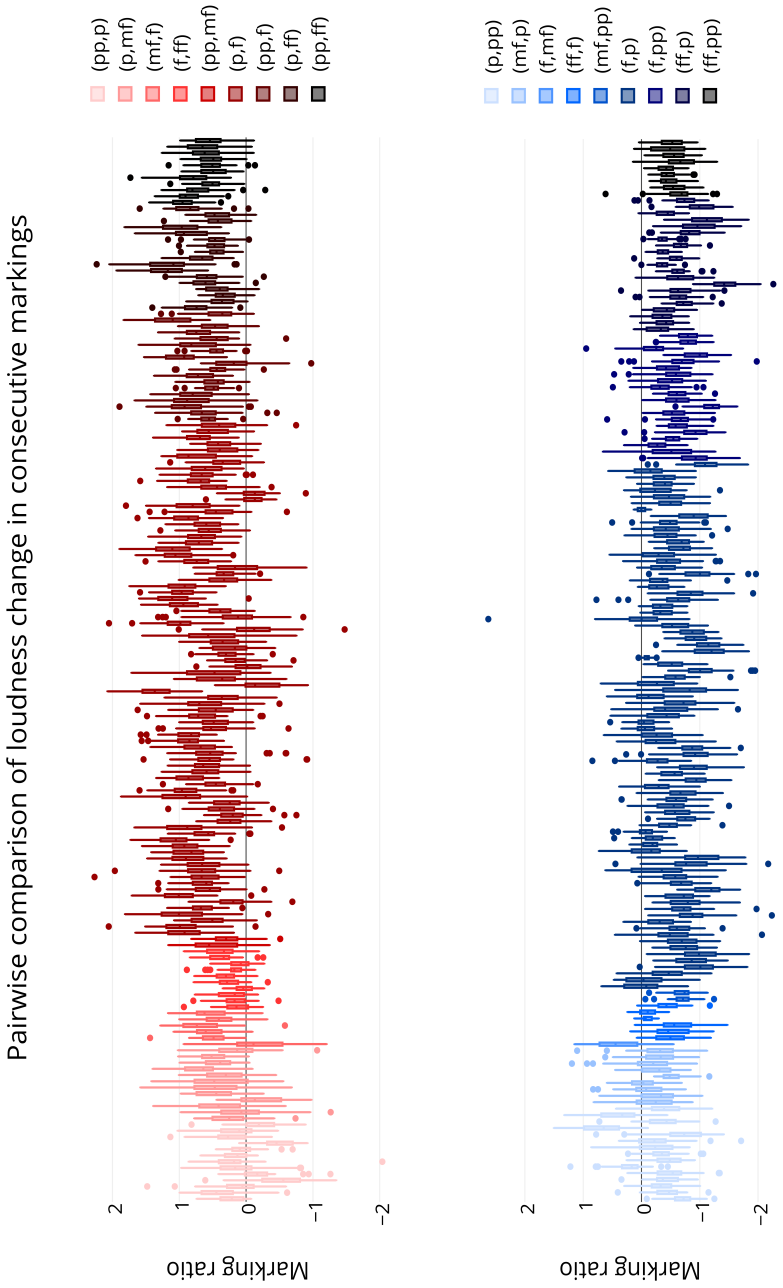


Figure 4.15: The log ratio of the loudness in the transition from a marking m_{k-1} to a marking m_k in the score sequence (m_{k-1}, m_k) where $\ell_{k-1} < \ell_k$ (top-left) and $\ell_k < \ell_{k-1}$ (bottom-right) following the OLS.

pianists move from a louder to a softer marking.

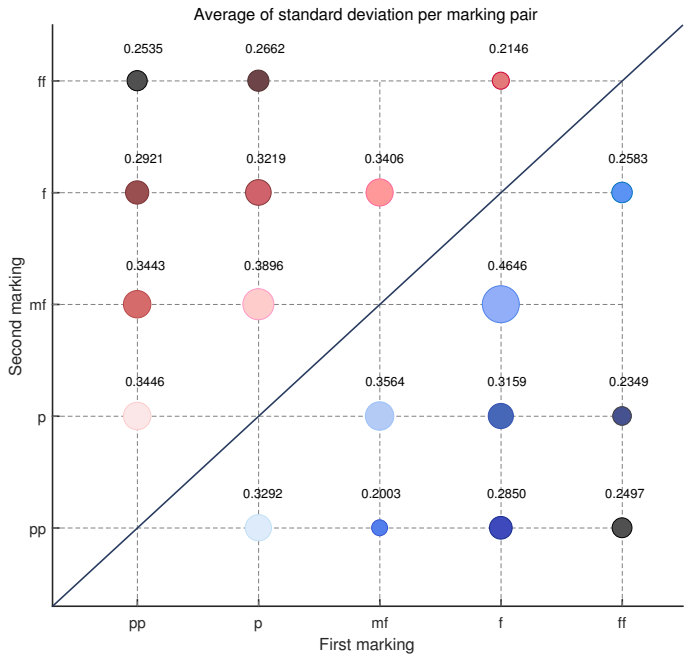


Figure 4.16: The average standard deviation values of the log loudness ratio for each marking—transitions are from the x-axis to the y-axis.

Note that the five smallest SD values can be found for the sequences (***mf***, ***pp***), (***f***, ***ff***), (***ff***, ***p***), (***ff***, ***pp***), and (***pp***, ***ff***) with corresponding values 0.2003, 0.2146, 0.2349, 0.2497, and 0.2535. This means that the dynamic change is more consistent for pairs of markings near the upper limit of the OLS upper limit and at the extremes of the OLS (those having the greatest distance on the OLS scale).

The above observations relate to loudness transitions between consecutive and distinct markings, thus dealing with local behaviour. Having considered the average behaviours of individual markings in the previous section, the next section will study the evolution of these individual markings over the course of a performance of a piece.

4.1.3 Analysis of different manifestations of the same dynamic markings throughout a piece

We have already observed from the previous findings that the dynamic markings are not performed in the same way when they appear in the score more than once. In this section, we answer the question: "How different is the performance of the same dynamic marking over the course of a piece?" In response to this

question, we compare the dynamic values corresponding to the same marking whenever it appears, using two kinds of Analysis of Variance (ANOVA). A one-way ANOVA is used to compare the means of the dynamic values whenever a specific dynamic marking appears in the score more than once, while a two-way ANOVA is used to compare the means of the dynamic values whenever a specific dynamic marking appears in the score more than once, and the means of each performer's response over the observations.

Both ANOVA types test the hypothesis that the data samples are drawn from populations that do not have the same mean against the null hypothesis that there is no difference between the groups. A result is statistically significant if it allows us to reject the null hypothesis. In order to measure statistical significance of the results, we use the P–value which evaluates how well the data supports that the null hypothesis is true. A low P–value suggests that your sample provides enough evidence that you can reject the null hypothesis for the entire population, meaning in our case that the population means are likely different.

It is worth noting that, in defense of our use of the P–value, we have merely used the P–value to determine if there is an effect, and not to quantify the effect (Nuzzo [2014]). In the future, the effect size and the confidence interval can be used to convey the magnitude and the relative importance of the effect; this is currently out of scope of this research.

The two sets of ANOVA results are presented in the "P–value 1" and the "P–value 2" columns respectively in Table 4.3. P–values less than the chosen significance level $\alpha = 0.05$ are highlighted in bold (and coloured red). $P > 0.05$ means that there is no significant difference between the dynamic values of the same markings.

Two cases exist where the P–value1 is in bold (and red) and P–value2 is not. The two cases are the two ***ff***'s in Mazurka Op. 6 No. 1, and the three ***f***'s in Mazurka Op. 67 No. 3. No significant difference was found according to the one-way ANOVA ($P > 0.05$), but the groups indeed differ according to the two-way ANOVA ($P < 0.05$). This means that the group means across recordings are similar, but the means diverge when considering the treatments of markings across recordings.

Figure 4.17 illustrates the responses to the ***ff***'s in Mazurka Op. 6 No. 1 and to the ***f***'s in Mazurka Op. 67 No. 3 in separate graphs. The x-axis indexes the recordings. The y-axis shows the loudness values. In each graph, horizontal lines show the mean loudness values for each instance of each marking. Different shapes indicate the dynamic at which each performer realises each marking. Note in both graphs that the means for each performer differ greatly across performers, which explains why P–value2 is less than 0.05 for these two cases.

Mazurka	#	P–value 1	P–value 2	Mazurka	#	P–value 1	P–value 2
	pp				***mf***		
M06-1	5	$< 10^{-10}$	$< 10^{-10}$	M17-3	3	$9.4285{\cdot}10^{-7}$	$< 10^{-10}$
M07-1	2	**0.3473**	**0.1574**	M67-4	2	$2.3081{\cdot}10^{-6}$	$< 10^{-10}$
M07-3	4	$< 10^{-10}$	$< 10^{-10}$	M68-2	7	$< 10^{-10}$	$< 10^{-10}$
M24-2	2	0.0134	0.0038		***f***		
M24-4	7	$< 10^{-10}$	$< 10^{-10}$	M06-1	3	**0.6376**	**0.5168**
M30-3	10	$< 10^{-10}$	$< 10^{-10}$	M06-2	6	$< 10^{-10}$	$< 10^{-10}$
M30-4	2	**0.2532**	**0.1297**	M06-3	5	$< 10^{-10}$	$< 10^{-10}$
M33-2	2	0.0015	$< 10^{-10}$	M07-1	7	$< 10^{-10}$	$< 10^{-10}$
M41-1	2	**0.5410**	**0.4526**	M07-2	5	$6.0140{\cdot}10^{-7}$	¡$\cdot 10^{-10}$
M50-3	2	$< 10^{-10}$	$< 10^{-10}$	M07-3	3	$< 10^{-10}$	$< 10^{-10}$
M67-1	2	**0.2452**	**0.0625**	M17-1	6	$< 10^{-10}$	$< 10^{-10}$
M67-2	2	**0.7054**	**0.4278**	M17-2	3	$2.3741{\cdot}10^{-5}$	$6.3259{\cdot}10^{-6}$
M67-3	4	0.0137	$< 10^{-10}$	M24-2	5	$8.7497{\cdot}10^{-7}$	$< 10^{-10}$
M68-2	5	$< 10^{-10}$	$< 10^{-10}$	M24-4	4	$< 10^{-10}$	$< 10^{-10}$
	p			M30-1	3	$2.3096{\cdot}10^{-9}$	$< 10^{-10}$
M06-1	8	$< 10^{-10}$	$< 10^{-10}$	M30-2	4	**0.2580**	**0.0762**
M06-2	7	$2.4228{\cdot}10^{-5}$	$< 10^{-10}$	M30-3	11	$< 10^{-10}$	$< 10^{-10}$
M06-3	12	$< 10^{-10}$	$< 10^{-10}$	M30-4	3	$< 10^{-10}$	$< 10^{-10}$
M07-1	3	$1.1657{\cdot}10^{-4}$	$3.3299{\cdot}10^{-8}$	M33-1	2	$6.1271{\cdot}10^{-9}$	$< 10^{-10}$
M07-2	8	$2.0958{\cdot}10^{-7}$	$< 10^{-10}$	M33-2	7	$< 10^{-10}$	$< 10^{-10}$
M07-3	9	$< 10^{-10}$	$< 10^{-10}$	M33-4	9	$< 10^{-10}$	$< 10^{-10}$
M17-2	2	$< 10^{-10}$	$< 10^{-10}$	M41-1	4	$< 10^{-10}$	$< 10^{-10}$
M17-3	6	0.0026	$7.3828{\cdot}10^{-6}$	M41-2	2	0.0114	$< 10^{-10}$
M17-4	5	$< 10^{-10}$	$< 10^{-10}$	M41-4	3	$< 10^{-10}$	$< 10^{-10}$
M24-1	3	$3.1148{\cdot}10^{-9}$	$1.0286{\cdot}10^{-8}$	M50-1	6	$< 10^{-10}$	$< 10^{-10}$
M24-2	5	$7.0718{\cdot}10^{-6}$	$1.9338{\cdot}10^{-9}$	M50-2	4	$3.3993{\cdot}10^{-6}$	$< 10^{-10}$
M24-3	6	$3.4907{\cdot}10^{-6}$	$< 10^{-10}$	M50-3	5	$< 10^{-10}$	$< 10^{-10}$
M24-4	11	$< 10^{-10}$	$< 10^{-10}$	M56-1	5	$< 10^{-10}$	$< 10^{-10}$
M30-1	5	$< 10^{-10}$	$< 10^{-10}$	M56-2	3	$< 10^{-10}$	$< 10^{-10}$
M30-2	10	$3.3254{\cdot}10^{-7}$	10^{-10}	M56-3	8	$< 10^{-10}$	$< 10^{-10}$
M30-4	11	$< 10^{-10}$	$< 10^{-10}$	M59-1	3	$< 10^{-10}$	$< 10^{-10}$
M33-1	3	0.0014	$5.9449{\cdot}10^{-5}$	M59-2	3	$< 10^{-10}$	$< 10^{-10}$
M33-3	3	**0.3044**	**0.1860**	M59-3	5	$< 10^{-10}$	$< 10^{-10}$
M33-4	2	$4.3868{\cdot}10^{-7}$	$< 10^{-10}$	M63-1	4	$< 10^{-10}$	$< 10^{-10}$
M41-1	5	$< 10^{-10}$	$< 10^{-10}$	M63-3	2	**0.2103**	**0.1473**
M41-3	4	$< 10^{-10}$	$< 10^{-10}$	M67-1	5	$3.7128{\cdot}10^{-7}$	$2.7101 \cdot 10^{-9}$
M41-4	2	$< 10^{-10}$	$< 10^{-10}$	M67-2	2	0.0030	$5.3387{\cdot}10^{-5}$
M50-1	8	$< 10^{-10}$	$< 10^{-10}$	M67-3	3	**0.1370**	$< 10^{-10}$
M50-2	10	$< 10^{-10}$	$< 10^{-10}$	M67-4	4	$2.4282{\cdot}10^{-6}$	$< 10^{-10}$
M50-3	9	$< 10^{-10}$	$< 10^{-10}$	M68-1	6	$< 10^{-10}$	$< 10^{-10}$
M56-1	8	$< 10^{-10}$	$< 10^{-10}$	M68-2	2	**0.4895**	**0.2835**
M56-2	4	$< 10^{-10}$	$< 10^{-10}$	M68-3	2	$< 10^{-10}$	$< 10^{-10}$
M56-3	7	$< 10^{-10}$	$< 10^{-10}$		***ff***		
M59-1	5	$< 10^{-10}$	$< 10^{-10}$	M06-1	2	**0.0937**	0.0044
M59-2	2	$< 10^{-10}$	$< 10^{-10}$	M06-3	4	0.0510	$8.3895{\cdot}10^{-5}$
M59-3	6	$< 10^{-10}$	$< 10^{-10}$	M07-3	2	**0.5328**	**0.3838**
M63-1	4	$< 10^{-10}$	$< 10^{-10}$	M24-4	11	$< 10^{-10}$	$< 10^{-10}$
M63-3	2	**0.1518**	**0.1277**	M30-3	3	$5.2395{\cdot}10^{-5}$	$< 10^{-10}$
M67-1	5	$< 10^{-10}$	$< 10^{-10}$	M30-4	2	**0.2045**	**0.0601**
M67-2	5	$< 10^{-10}$	$< 10^{-10}$	M33-2	3	$1.9163{\cdot}10^{-8}$	$< 10^{-10}$
M67-3	3	**0.4995**	**0.2927**	M59-2	2	0.0096	$1.1610{\cdot}10^{-4}$
M67-4	5	$< 10^{-10}$	$< 10^{-10}$	M67-1	4	**0.9314**	**0.8267**
M68-1	6	$< 10^{-10}$	$< 10^{-10}$	M67-3	3	**0.7174**	**0.6024**
M68-2	7	$< 10^{-10}$	$< 10^{-10}$				
M68-3	5	$9.8827{\cdot}10^{-6}$	$< 10^{-10}$				

Table 4.3: Results for one-way (P–value 1) and two–way (P–value 2) ANOVA tests for marking groups per Mazurka. $P < 0.05$ indicates the conclusion that the data means of the groups of values that correspond to a specific dynamic marking differ. Symbol "#" indicates the number of the same markings that appear in the specific Mazurka.

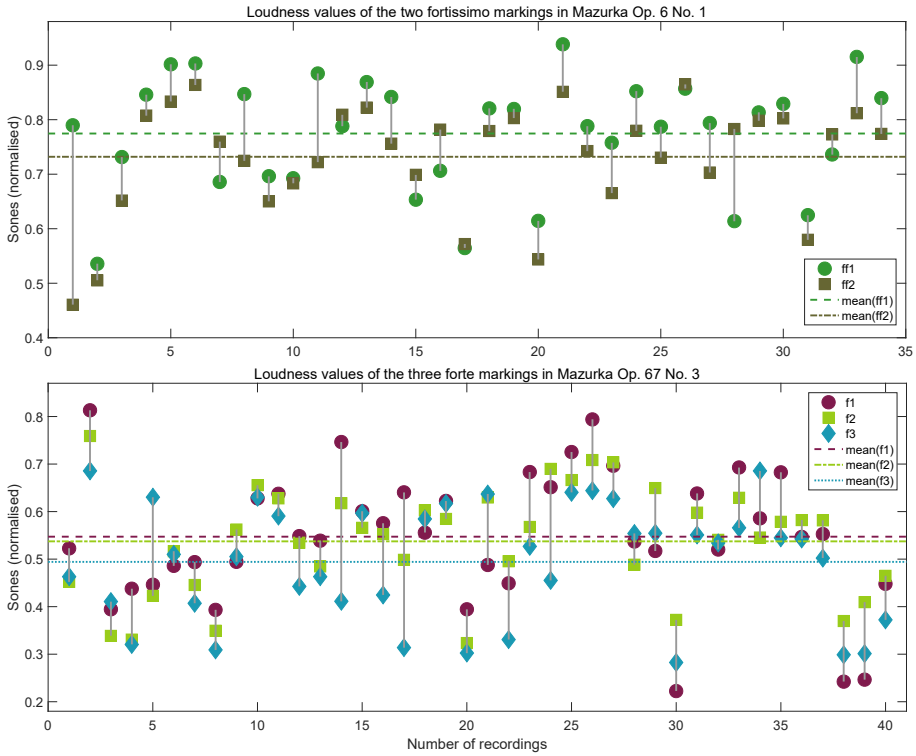


Figure 4.17: The case of the loudness values of groups of dynamic markings (***ff***'s in Mazurka Op. 6 No. 1—top, and ***f***'s in Mazurka Op. 67 No. 3—bottom) that the one-way ANOVA shows no significant difference in the marking group means (dotted lines), but the two-way ANOVA which adds the diversity in the marking group means per pianist, shows the opposite.

In order to further analyse the cases where the markings appeared in the score of the same Mazurka more than once and their treatments are found to be different, we implement the multiple comparison test of means. Specific cases are highlighted in Figure 4.18. In each plot in Figure 4.18, the confidence interval of an instance of a marking is shown as a horizontal line. The length of each line for each figure is the same because the confidence interval is computed for the marking in that Mazurka. The vertical lines delineate author-selected highlight regions. We consider each plot in turn:

Mazurka Op. 6 No. 3, ***p***'s: the markings between the vertical bars—$\boldsymbol{p}_4$, $\boldsymbol{p}_5$, $\boldsymbol{p}_6$, and $\boldsymbol{p}_7$—are part of a repeated phrase with alternating ***ff***'s and ***p***'s, as shown in Figure 4.19; as a result, their treatment is starkly different than that for the other ***p***'s. Mazurka Op. 24 No. 4, ***p***'s: the first and last marking are almost perfectly aligned, giving the impression of an overall loudness stability, but the dynamics of this marking varies greatly in between.

Mazurka Op. 63 No. 1, ***f***'s: the first ***f*** is significantly louder than the ones

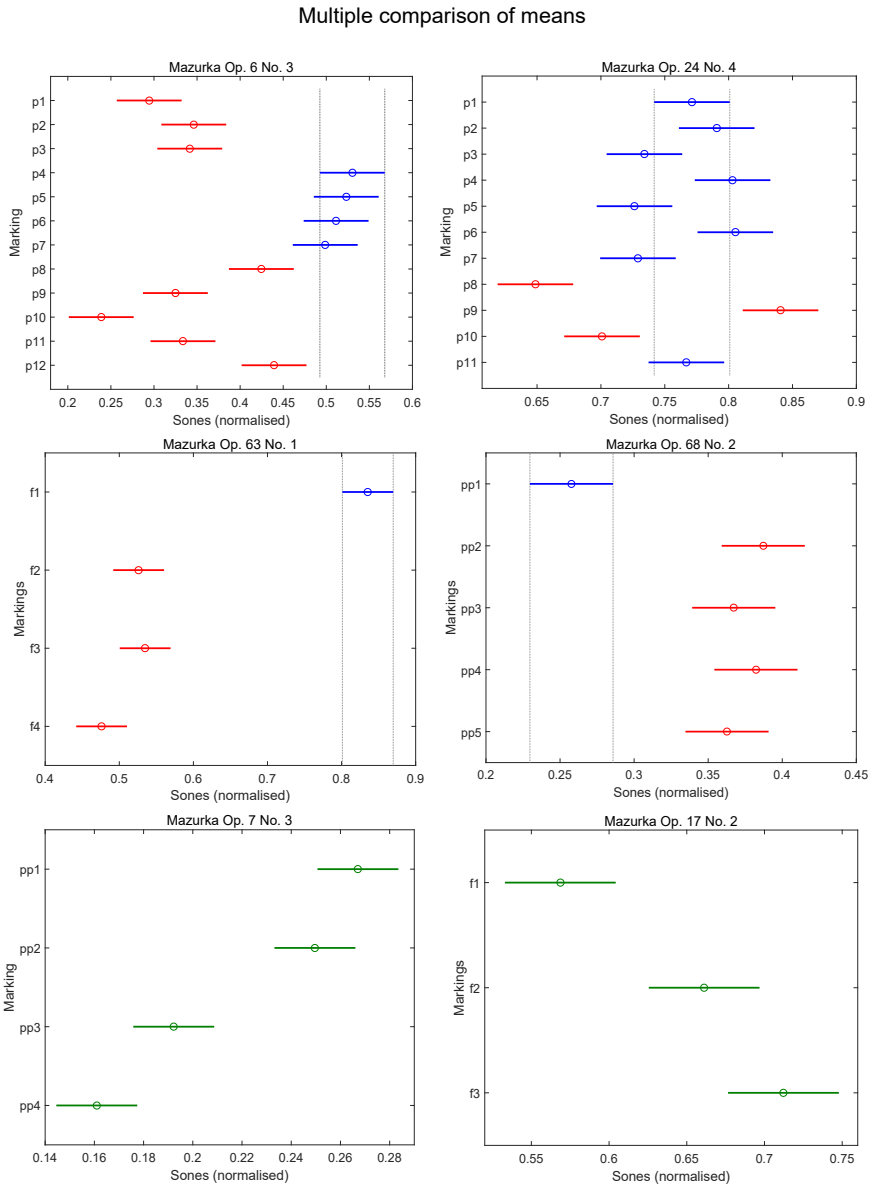


Figure 4.18: Multiple comparison of the means for the groups of the same markings: twelve ***p***'s in Mazurka Op. 6 No 3, eleven ***p***'s in Mazurka Op. 24 No. 4, four ***f***'s in Mazurka Op. 63 No. 1, five ***pp***'s in Mazurka Op. 68 No. 2, four ***pp***'s in Mazurka Op. 7 No. 3, and three ***f***'s in Mazurka Op. 17 No. 2 .

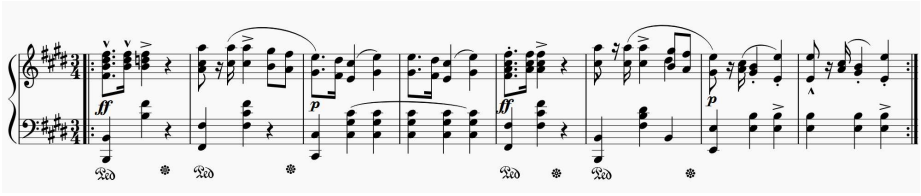


Figure 4.19: The repeated phrase in Mazurka Op. 6 No. 3 where the markings $\boldsymbol{p}_4$, $\boldsymbol{p}_5$, $\boldsymbol{p}_6$ (repetition of $\boldsymbol{p}_4$), and $\boldsymbol{p}_7$ (repetition of $\boldsymbol{p}_5$) appear.

that follow. Mazurka Op. 68 No. 2, ***pp***'s: the first ***pp*** is significantly softer than the ones that follow. In both cases, except for the outlying starting response, the means of the responses to other instances of the marking are similar.

Mazurka Op. 7 No. 3, ***pp***: the responses to the ***pp*** markings are decreasing over time. Mazurka Op. 17 No. 2, ***f***: the responses to the ***f*** markings are increasing over time. These two cases demonstrate two possible responses to sequences of the same markings.

In summary, we have studied the ways in which dynamics evolve over time using one- and two-way ANOVA tests and multiple comparison tests of means. We have shown that, most of the time, there exists a significant difference between dynamic values for the same markings that appear in the score more than once. We rejected the hypothesis that the means are drawn from the same population when analysing mean responses to the same marking, as well as variations in individual recordings. Also, we have shown the kinds of patterns that emerge in responses to sequences of the same dynamic markings across scores.

4.2 Conclusions - Discussion

In this chapter we have investigated the relationship between loudness levels and dynamic markings in the score. The statistical analysis presented rejects hypotheses such as "a dynamic marking is performed in the same way when it appears in the score more than once" and "each performance follows the dynamic structure of the piece as indicated by the score markings". Our findings underpin the statement that while modern interpretations are more faithful to the score [Rink, 2002, p.4],the manifested sound for the same effect, such as a ***crescendo***, can be distinct and vary greatly between performers, between performances, and within the same performance.

In addition, our findings reveal that one simple rule for each dynamic marking cannot suffice to define its possible dynamic levels; in fact, the realised dynamics are constrained by many other factors, the most important ones of which are: the current, previous, and next dynamic markings; the distance from the current marking to the previous and next markings; the nearest non-dynamic marking annotated between the previous and current or next dynamic marking, such as ***crescendo***; and, any qualifying annotation appearing simultaneously with the current dynamic marking, such as ***dolcissimo***.

The factors above have played an important role in the study that is presented in Chapter 7 (Kosta et al. [2016]); they constitute the list of features that have been extracted in order to implement machine-learning approaches firstly to predict loudness values corresponding to different dynamic markings and musical contexts and secondly to predict dynamic markings corresponding to different loudness levels and musical contexts.

Two important points considering our analysis are the following. Firstly, possible deviations from the score occur, mostly playing notes in different octave as written or applying rubato in position that is not indicated. Secondly, it is not clear what an analysis of the dynamics of a performance represents when based on an audio recording. Are dynamics either the performer's intention or the audio engineer's perception? The mastering of the sound during the production process can indeed affect the result of what we actually perceive while listening to the recording, which could be different from what was actually played. In our case, the data is taken from studio or professional live recordings and the musician's consent to the final result is assumed. The above point to the inherent complexity of the subject matter under investigation; as such, the results reported should be seen as providing a sound basis for a more sophisticated approach to understanding the relativity of expression of dynamic markings.

Chapter 5

Interlude: An analysis of outliers in performed loudness transitions

In this chapter, we investigate the relationship between dynamic markings in the score, such as ***p*** (piano), and performed loudness. In particular, we examine the performed loudness for consecutive pairs of distinct dynamic markings. The main aim of this study is to understand expressive variations in a musician's playing style relative to other musicians by analysing outliers in the ways performers navigate transitions between dynamic markings.

More specifically, we have conducted two experiments, the first one explores unusual (sometimes contrarian) interpretations of dynamic changes from one marking to the next, and the second investigates unusual interpretations of dynamic changes through entire performances. This study is not the first to explore extreme performances. An example in case is a detailed explanation of Cortot's unique playing style in his recording of Chopin's Berceuse, Op. 57 interpretation (Leech-Wilkinson [2015]).

For this study, the whole dataset as described in Chapter 3 has been used. By the time we have the dynamic value that corresponds to each marking, we create our data in a pairwise manner by subtracting from the dynamic value l_b of marking b the dynamic value $l_{(b-1)}$ of the previous marking. The result is discretised as follows:

$$f(x) = \begin{cases} 2, & x > 0.5 \\ 1, & 0 < x \leq 0.5 \\ 0, & x = 0 \\ -1, & -0.5 \leq x < 0 \\ -2, & x < -0.5 \end{cases} \tag{5.1}$$

where the input $x = l_b - l_{(b-1)}$ for every marking position b in each Mazurka. The methodology relies on the linearity of the sone values so that we are able to create levels of a same linear distance. This discretisation steps allows us to handle the noisy loudness data.

This chapter describes two experiments: the first deals with outliers in interpretations of dynamic changes from one marking to the next; the second investigates unusual interpretations of dynamic changes for entire recordings. A marking pair is considered an outlier if the value of its corresponding dynamic change from one marking to the next lies in the top or bottom quartile of the distribution of the dynamic change values. A recording is considered an outlier if it is grouped into a cluster having the least number of recordings for that piece.

The remainder of this chapter is organized as follows: In Sections 5.1, and 5.2 we present the description and the results from the two experiments, respectively; finally in Section 5.3 we present the conclusions and further discussion.

5.1 Outliers in individual marking pair transition

This experiment explores unusual interpretations of loudness changes across marking pairs. We compare dynamic levels at different pairs of markings that appear in the scores. We consider consecutive pairs $(l_{(b-1)}, l_b)$, which represent the transition from the marking at position $b1$ to that at position b. A marking pair is an outlier if its corresponding dynamic change lies in the upper or lower quartile of such changes. Thus, outliers represent unusual changes in dynamics from a particular marking to the following one, for example getting softer in a transition where in most recordings the music gets louder.

The results indicate that the highest proportion of outliers appear at the transitions (***p***, ***f***), (***f***, ***p***), (***pp***, ***f***), and (***p***, ***pp***), with proportions over the total number of outliers equal to 0.1969, 0.1710, 0.0774, and 0.0651, respectively. We observe two ways for outliers to occur: the first is to overshoot or undershoot the change from one marking to the next in relation with the most common

trends; the second is to contradict the change from one marking to the next in opposition to the most common behaviour. Figure 5.1 shows how the outliers are distributed, giving their proportions for every marking-pair over all Mazurkas where the specific transition occurs. Note that no data points exist for the pairs (***mf***, ***ff***), (***mf***, ***pp***), and (***ff***, ***mf***) as these marking pairs do not appear in the Mazurkas we analyse.

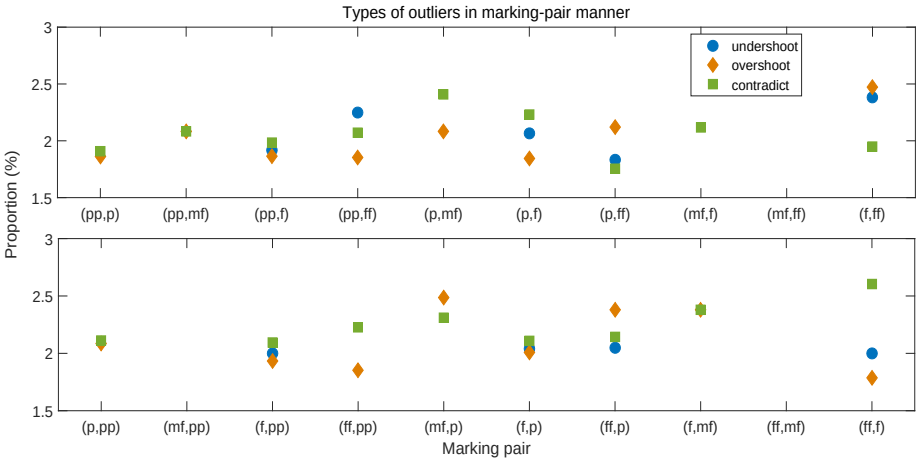


Figure 5.1: Distribution of outlier transitions between each dynamic marking pair (circle=undershoot, rhombus=overshoot, square=contradictory).

A transition that includes many outliers is the (***f***, ***p***) at the beginning of Mazurka Op. 59 No. 3. This transition had a mean value of -1, but over fifty-six recordings, eleven recordings overshot to -2, ten recordings made a contradictory $+1$ transition, and one recording even had an extreme contradictory transition of $+2$. The score where the specific markings are located is presented in Figure 5.2. It is worth mentioning that the specific pair has the biggest portion of outliers.

If we consider only pianists that have recorded more than thirty (out of the forty-four) Mazurkas, Cortot is the pianist with the highest number of outliers, as shown in Table 5.1.

Pianist	# Outliers	# Recordings	# Pairs	Ratio
Cortot	74	42	472	0.1568
Magin	43	35	379	0.1135
Rangell	47	37	416	0.1130

Table 5.1: Pianists whose recordings had the highest proportion of outlier transitions.

Table 5.2 presents the recordings with the highest rate of outliers. Rubinstein's recording of Mazurka Op. 30 No. 1, Ashkenazy's Op. 59 No. 2, and

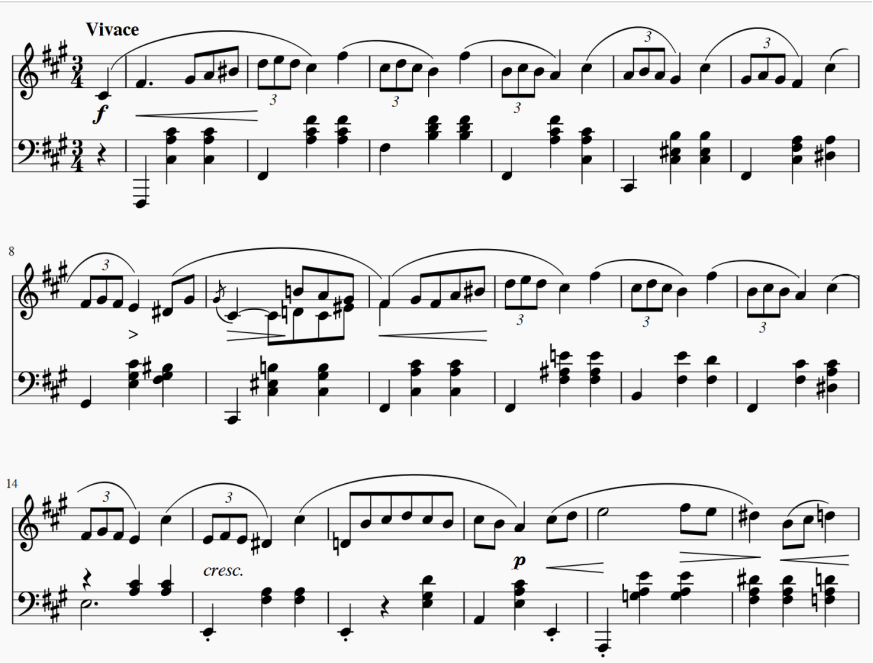


Figure 5.2: Score excerpt from Mazurka Op. 59 No. 3 where the markings of the first pair (***f***, ***p***) appears. This transition has the highest proportion of outliers.

Block's Op. 68 No. 3 appear on the list.

Mazurka	Pianist (year of recording)	# Outliers	# Pairs	Ratio
M30-1	Rubinstein (1952)	5	7	0.714
M59-2	Ashkenazy (1999)	5	7	0.714
M68-3	Block (1995)	5	7	0.714

Table 5.2: Recordings having the highest proportion of outlier transitions.

Mazurka Op. 24 No. 2 has the highest proportion of outliers. Figure 5.3 displays the dynamic values for all recordings of Mazurka Op. 24 No. 2; outlying transitions are highlighted with solid lines. In some cases, like the transition from the second to the third marking, there are no outliers despite the diversity of the dynamic behaviour. The reason is that two distinct sub-populations of the values have been created at the specific cases.

5.2 Recordings in outlier dynamic clusters

One question that follows naturally from the experiment that is described in Section 5.1 is how varied the loudness behaviour across single recordings is. In order to investigate this question, we map each recording to a time series

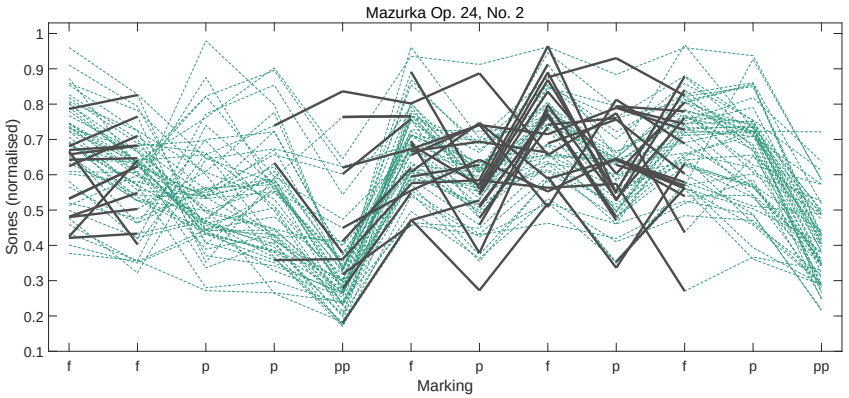


Figure 5.3: Loudness graphs for performances of Mazurka Op. 24 No. 2 (the one having the highest proportion of outliers); highlighted in bold are outlier transitions.

of discretised values of the function f defined in Section 5.1. Each recording then becomes a curve with discretised points in score time. The purpose of the procedure above is to create meaningful clusters of curves so as to further analyse the resulting shapes, and distinguish unusual behaviours.

To cluster the curves obtained, we implemented the k-means machine learning method. The number of clusters, k, is defined using the "gap statistic" by Tibshirani et al. [2001]; we further limit k to the range $[3, 8]$ to ensure a reasonable number of recordings appear in each cluster, but also provide the flexibility of detecting clusters that include curves of an unusual behaviour at the same time.

The result of this process is a number of loudness behaviour clusters for each Mazurka. Clusters with the least number of recordings are labelled as outliers; then, we create a list of pianists that appear at these outlier clusters. Table 4 presents the top three outlier cluster pianists whose recordings appear in more than thirty different Mazurkas.

Pianist (year of recording)	# Outlier clusters	# Recordings	Ratio
Cortot (1951)	12	42	0.2727
Poblocka (1999)	9	33	0.2727
Sztompka (1959)	8	38	0.2105

Table 5.3: Pianists having the highest proportion of recordings in outlier clusters.

Magin is the only pianist whose recordings were the single element which was contained in a cluster at the cases of Mazurka Op. 7 No. 1 and Mazurka Op. 33 No. 3. Figure 5.4 highlights the loudness behaviour of Magin's recordings of both Mazurkas, and how they differ from the other recordings. Figure 5.4

shows the centroid for each cluster, thus the discretised dynamic value may not be equal to an outcome of the function f.

Magin appears to differ from the other recordings in interpreting the two last transitions in Mazurka Op. 7 No. 1, these are from ***f*** to ***pp*** and ***pp*** to ***f*** of the last three markings over thirteen markings in total. This observation essentially shows that however the values of Magin's cluster are relatively close to the corresponding centroids of other clusters, the miss-interpretation of the last ***pp*** marking was crucial for separating the whole recording from the others. In Mazurka Op. 33 No. 3 Magin interprets all four markings in a contrary way of the centroid points that most of the other recordings form and this observation explains the reason why a separate cluster forthe unique recording has been created.

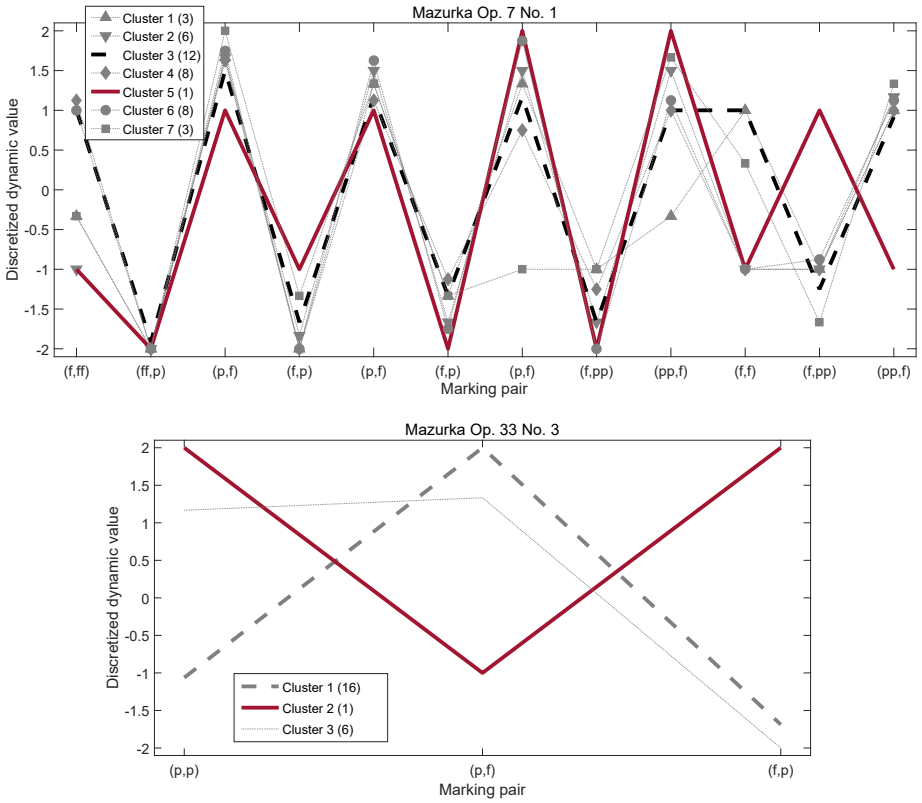


Figure 5.4: Loudness transition clusters found for Mazurka Op. 7 No. 1 (top) and Op. 33 No. 3 (bottom); in each case, the cluster comprising of only Magin's recording is shown in solid bold lines, the cluster having the highest number of recordings is shown as dotted bold lines.

Next, we focus on when there are commonalities in unusual interpretations. For this, we consider pianists whose recordings are in outlier clusters for every Mazurka, and we compute the number of times two pianists' recordings are

classified into the same outlier cluster for all Mazurkas. The results are shown in Table , where we give the number of outlier clusters containing the specific pianist pairs. We focus on pairs that co-occur in outlier clusters for more than twenty Mazurkas.

Pianist	Fliere	Milkina	Ezaki	Barbosa
Chiu	0	21	0	21
Smith	27	0	0	0
Rubinstein	0	26	0	0
Kushner	0	0	21	0
Czerny-Stefanska	0	0	0	21

Table 5.4: Pianists whose recordings co-occur in more than twenty outlier clusters.

Table 5.4 shows higher degrees of similarity in performed loudness between pianists Chiu–Milkina, Chiu–Barbosa, Smith–Fliere, Rubinstein–Milkina, Kushner–Ezaki, and Czerny-Stefanska–Barbosa. Thus, although certain pianists are more often in outlier clusters, the ways in which they differ in their loudness interpretations often follow shared patterns.

5.3 Discussion

In this study we have analyzed outliers in loudness interpretations in piano recordings. The results show that we are able to distinguish specific dynamic marking pairs which have a high proportion of outliers. In addition, we are able to detect recordings for which the interpretations do not follow the emerging patterns. There is considerable agreement between an interpretation and information derived from a score, nevertheless it is a matter of the pianists' musical choices to control a sense of longer-term intensity modulation. In any case, the different choices followed by the pianists hold the key to a performance stimulated by creativity and imagination.

Chapter 6

A change-point approach towards representing musical dynamics

Summary: This chapter proposes a novel application of change-point techniques to the question of how dynamic markings in a score correspond to performed loudness. Firstly, an exploratory study is presented: we apply and compare two change-point algorithms −Killick, Fearnhead, and Eckley's Pruned Exact Linear Time (PELT) method, and Scott and Knott's Binary Segmentation (BS) approach−to detecting changes in dynamics in the Mazurka recordings. Dynamic markings in the score, assumed to correspond to change points, serve as ground truth. The PELT algorithm has a higher average best F-measure compared to the BS algorithm, it also results in a smaller average Hausdorff distance. Secondly, a study is presented for a thorough analysis on the change-point positions derived using PELT and their relation to the corresponding score position. The results show that significant dynamic score markings do indeed correspond to change points, position of a text indication for change in expression or tempo can be a change point, and change points at score positions without dynamic markings demarcate salient structural events.

The aim of the studies presented in this chapter are to understand the connection between absolute loudness values and intended dynamics, which includes but are not only the notated markings.

In the first exploratory study, we evaluate the following steps: we analyze the dynamic values time series extracted from music audio. Using a change-point detection method, we estimate the position where the statistical properties of the sequence change. We compare and contrast the output of two change-point algorithms: the Pruned Exact Linear Time (PELT) algorithm by Killick et al. [2012] and the Binary Segmentation (BS) algorithm by Scott and Knott [1974]. We evaluate the algorithms by comparing the change points detected with the location of the dynamic markings obtained from the score as ground truth.

In the second study, we do not use the position of the dynamic markings as ground truth data; instead, the evaluation of the results focuses on the analysis of the meaning of the change point positions found, without using the dynamic marking locations as ground truth.

The chapter is organized as follows: Section 6.1 explains the change-point techniques that have been explored; Section 6.2 presents the methodology followed for extracting the change points; Section 6.3 provides the results of the comparison between the change-point methods. Section 6.4 presents the analysis of the change-points positions derived from PELT. The latter sections include discussion subsections (Section 6.3.1 and Section 6.4.1, respectively).

6.1 Change-points techniques

In this section, we present the techniques for detecting changes in our data that have been explored in our studies.

Binary Segmentation (BS) versus Pruned Exact Linear Time (PELT) Using loudness time series, our aim is to detect points where changes to the underlying statistical properties of the data occur. The null hypothesis H_0 corresponds to the non-existence of a change point. Detection of a change point negates the null hypothesis H_0. The general likelihood ratio-based approach to change-point detection provides the asymptotic distribution of the likelihood ratio test statistic for a change in the mean and in the variance of normally distributed observations (Killick et al. [2010]).

We use the R package "changepoint" by Killick and Eckley [2014] to investigate the application of two multiple change-point search methods, BS and PELT, to the loudness time series. For both algorithms, the "meanvar" function was implemented, meaning that we detect the changes for both mean and variance for the datasets. Essentially, through variance, we capture the horizontal

density of the data, an example being the audio excerpt presented in Figure 6.1): The same raw sone values have been segmented by levels from changes in mean (a), levels from changes in variance (b), and levels from changes in both mean and variance combined. The highlighted box in all sub-figures show that there is a possibility an extra change point to been detected in (c), following the information from the variance shown in (b).

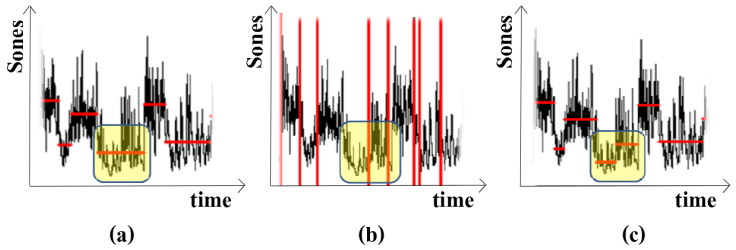


Figure 6.1: An example change-point detection result from the same audio excerpt (raw data in sones). (a): levels from changes in mean; (b): levels from changes in variance; (c): levels from changes in both mean and variance combined.

In both the BS and PELT algorithms, the aim is to minimize

$$\sum_{i=1}^{m+1} [C(y_{(\tau_{i-1}+1):\tau_i})] + \beta f(m), \tag{6.1}$$

where τ is a time point which separates two groups of data $\{y_1, ...y_\tau\}$ and $\{y_{\tau+1}, ..., y_n\}$ having different statistical properties, C is the log-likelihood cost function for a segment, and $\beta f(m)$ is a penalty to avoid over fitting, where $f(m) = m$.

The BS algorithm applies a binary divide and conquer approach. It divides the data each time a change point is found, and applies the same process to the two new subsets. It is computational efficient; however, because the location of a change point depends on that of other change points, the BS algorithm gives an approximate solution as the whole data set is not scanned each time.

The PELT algorithm, on the other hand, provides an exact solution as it is based on dynamic programming. The dynamic programming approach is further helped by a pruning step which makes its computational cost linear in the data Killick et al. [2012]. As the data set grows, the number of change points increases, meaning that the change points thus obtained also tend not to be closely bunched.

6.2 Change-points extraction

For the purpose of the experiment, we follow the set up as described in Section 6.1 for both BS and PELT methods: in order to extract the change points, we use the R package "changepoint" by Killick and Eckley [2014], choosing the options $BinSeg$ (for BS) and $PELT$ (for PELT) and we implement the "meanvar" function. We use the raw sone values for the whole dataset of audio recordings as it is presented in Chapter 3.

In the case of the PELT method, we set the penalty values to alter in $[90, 1200]$ for every time series. In order to accommodate an adaptive estimation of the number of change points for each time series, we have implemented the heuristic described by Lavielle [2005]: the method introduces a contrast function J and the number of change points can be estimated by minimizing a penalized version of J; the minimum penalized contrast (MPC) estimate of the number of change points is defined as the greatest value of this number such that the second derivative of J is greater than a given threshold S. Intuitively when a true change-point is added to a segmentation then the fit of the model to the data is much improved. As more and more true change-points are added the fit continues to improve. In contrast, when a false change-point is added the improvement to the fit is small. Thus if we look at the rate of change of the fit this can help indicate what the true number of changes for a series is. This is what we do using the threshold S to determine the cut-off between true and false changes.

For an n^{th} recording of an m^{th} Mazurka having a frame sequence (A_{m_n}) and length $|(A_{m_n})|$, the threshold value is calculated by the following formula:

$$S = k \cdot |(A_{m_n})| \cdot \log |(A_{m_n})|. \tag{6.2}$$

The selection of the k values is defined for each Mazurka as shown in Table 6.1.

Mazurka index	M06-1	M06-2	M06-3	M07-1	M07-2	M07-3	M17-1	M17-2	M17-3	M17-4	M24-1
k	$7 \cdot 10^{-7}$	$9 \cdot 10^{-7}$	$1 \cdot 10^{-7}$	$3 \cdot 10^{-7}$	$5 \cdot 10^{-6}$	$1 \cdot 10^{-7}$	$2 \cdot 10^{-7}$	$2 \cdot 10^{-7}$	$9 \cdot 10^{-7}$	$4 \cdot 10^{-6}$	$8 \cdot 10^{-7}$
Mazurka index	M24-2	M24-3	M24-4	M30-1	M30-2	M30-3	M30-4	M33-1	M33-2	M33-3	M33-4
k	$8 \cdot 10^{-6}$	$4 \cdot 10^{-6}$	$3 \cdot 10^{-6}$	$1 \cdot 10^{-7}$	$9 \cdot 10^{-7}$	$2 \cdot 10^{-7}$	$7 \cdot 10^{-7}$	$1 \cdot 10^{-7}$	$3 \cdot 10^{-7}$	$1 \cdot 10^{-8}$	$3 \cdot 10^{-6}$
Mazurka index	M41-1	M41-2	M41-3	M41-4	M50-1	M50-2	M50-3	M56-1	M56-2	M56-3	M59-1
k	$1 \cdot 10^{-7}$	$7 \cdot 10^{-7}$	$1 \cdot 10^{-7}$	$2 \cdot 10^{-7}$	$4 \cdot 10^{-7}$	$7 \cdot 10^{-7}$	$9 \cdot 10^{-7}$	$6 \cdot 10^{-7}$	$1 \cdot 10^{-7}$	$9 \cdot 10^{-7}$	$1 \cdot 10^{-7}$
Mazurka index	M59-2	M59-3	M63-1	M63-3	M67-1	M67-2	M67-3	M67-4	M68-1	M68-2	M68-3
k	$9 \cdot 10^{-7}$	$9 \cdot 10^{-7}$	$6 \cdot 10^{-7}$	$9 \cdot 10^{-6}$	$1 \cdot 10^{-6}$	$4 \cdot 10^{-6}$	$2 \cdot 10^{-6}$	$2 \cdot 10^{-7}$	$1 \cdot 10^{-6}$	$9 \cdot 10^{-6}$	$2 \cdot 10^{-7}$

Table 6.1: The value of the parameter k in the threshold formula that has been used for all the recordings per Mazurka. Mazurkas are indexed as "M<*opus*>-<*number*>."

Essentially, we want to constrain the number of detected change points in all

the corresponding recordings by adjusting a single k value per Mazurka. The k values above provide a small standard deviation of the total number of change points detected in all the respective recordings. The average of these standard deviation values is 9.94.

In the case of the BS method, we compute the change points using the selected penalty value derived from the result of the PELT method. We also consider as maximum number of change points detected the number of change points selected from the PELT method. This is due to the fact that we want to extract results that are comparable.

6.3 Comparison results

In this section, we describe and report the results of the tests we conducted using the PELT and BS algorithms in answering the question of whether change points are located on the same score beat locations as dynamic markings. Assuming that change points correspond to the dynamic markings in the score and the dynamic markings are change points, the positions of the score markings serve as ground truth for the change points that have been derived for each recording as described in 6.2. Comparing the change points detected and the positions of the score dynamic markings, we identify the test that returns the highest F-measure:

$$\mathrm{F} = \frac{2\mathrm{PR}}{\mathrm{P} + \mathrm{R}}, \tag{6.3}$$

where precision value P, is the number of true positives (i.e. the number of change points that are in locations of markings) divided by the total number of change points, and recall value R, is the number of true positives divided by the total number of markings. A value of 1 for F-measure indicates that all change points are markings, a value of 1 for the precision value P indicates that all the change points correspond to dynamic markings and a value of 1 for the recall value R indicates that all markings are captured by change points.

Table 6.2 reports the F-measure, the precision value P and the recall value R as an average of all recordings per Mazurka. Also it reports the maximum values that have been captured per Mazurka.

In all pieces, the average F-measure value for the PELT algorithm (10.8%) is better than that of the BS algorithm (8.3%). Bold numbers indicate which method has the highest average F-measure per Mazurka.

To determine how closely the change points estimate the dynamic marking positions, we compute the Hausdorff distance between the change points and

	BS							PELT						
Mazurka	F		P		R		H_d	F		P		R		H_d
index	avg	max	avg	max	avg	max	avg	avg	max	avg	max	avg	max	avg
M06-1	0.09	0.22	0.08	0.19	0.11	0.35	32.94	**0.11**	0.31	0.08	0.27	0.17	0.41	29.03
M06-2	**0.14**	0.33	0.12	0.50	0.19	0.50	35.71	0.13	0.32	0.09	0.26	0.22	0.50	22.52
M06-3	0.26	0.45	0.25	0.43	0.29	0.52	22.38	**0.32**	0.55	0.28	0.64	0.41	0.67	16.55
M07-1	0.11	0.28	0.08	0.25	0.17	0.50	40.63	0.11	0.18	0.06	0.11	0.36	0.67	35
M07-2	0.14	0.32	0.12	0.26	0.19	0.50	60.11	**0.15**	0.30	0.10	0.19	0.31	0.75	33.40
M07-3	0.14	0.30	0.14	0.36	0.16	0.41	29.31	**0.19**	0.34	0.16	0.31	0.26	0.41	23.67
M17-1	0.10	0.24	0.07	0.25	0.22	0.67	36.09	**0.13**	0.30	0.08	0.18	0.40	1	35.84
M17-2	0.03	0.15	0.02	0.10	0.08	0.40	33.32	**0.06**	0.21	0.04	0.14	0.19	0.60	29.80
M17-3	0.02	0.14	0.01	0.10	0.03	0.22	83.72	0.02	0.10	0.02	0.08	0.06	0.22	50.44
M17-4	0.01	0.10	0.01	0.07	0.03	0.17	60.82	**0.03**	0.14	0.02	0.08	0.12	0.50	51.57
M24-1	**0.06**	0.14	0.03	0.08	0.20	0.75	176.57	0.05	0.12	0.03	0.07	0.26	0.50	187.63
M24-2	0.05	0.19	0.04	0.22	0.07	0.25	60.88	**0.06**	0.18	0.04	0.11	0.10	0.42	55.34
M24-3	0.03	0.18	0.02	0.13	0.06	0.33	47.87	**0.12**	0.32	0.07	0.19	0.39	1	38.21
M24-4	0.08	0.20	0.12	0.33	0.07	0.19	96.02	**0.13**	0.24	0.11	0.21	0.17	0.31	24.26
M30-1	0.16	0.52	0.12	0.38	0.29	0.86	31.98	**0.21**	0.36	0.14	0.23	0.49	0.86	26.18
M30-2	0.07	0.28	0.07	0.25	0.08	0.31	35.74	**0.15**	0.35	0.12	0.29	0.20	0.46	25.78
M30-3	0.22	0.39	0.21	0.35	0.23	0.50	35.35	**0.30**	0.52	0.25	0.47	0.39	0.67	31.20
M30-4	0.04	0.17	0.03	0.20	0.04	0.24	42.76	**0.08**	0.22	0.06	0.18	0.11	0.35	30
M33-1	**0.12**	0.44	0.07	0.31	0.28	0.80	52.06	0.09	0.28	0.05	0.17	0.27	0.80	56.19
M33-2	0.10	0.30	0.09	0.28	0.13	0.40	71.36	**0.12**	0.31	0.08	0.20	0.27	0.73	68
M33-3	0.03	0.18	0.02	0.11	0.10	0.67	46.30	**0.05**	0.16	0.03	0.09	0.29	1	46.17
M33-4	0.12	0.50	0.10	0.46	0.18	0.64	100.87	0.12	0.38	0.09	0.33	0.22	0.55	81.29
M41-1	0.06	0.23	0.04	0.20	0.09	0.27	45.69	**0.08**	0.24	0.06	0.16	0.15	0.45	42.89
M41-2	0.06	0.20	0.04	0.12	0.17	1	38.02	**0.07**	0.19	0.04	0.11	0.24	1	41.21
M41-3	0.07	0.28	0.05	0.17	0.18	0.80	31.54	**0.09**	0.24	0.05	0.14	0.28	0.80	27.92
M41-4	0.04	0.42	0.03	0.33	0.06	0.57	66.73	0.04	0.24	0.03	0.17	0.08	0.43	66.73
M50-1	0.10	0.28	0.09	0.25	0.13	0.36	33.89	**0.14**	0.31	0.11	0.28	0.21	0.43	23.36
M50-2	0.04	0.19	0.03	0.17	0.06	0.36	57.55	**0.07**	0.23	0.05	0.17	0.12	0.36	42.75
M50-3	0.04	0.12	0.03	0.09	0.06	0.24	77.18	**0.05**	0.15	0.03	0.11	0.08	0.24	83.87
M56-1	0.04	0.13	0.02	0.09	0.07	0.23	38.44	**0.07**	0.16	0.05	0.10	0.17	0.38	39.21
M56-2	0.07	0.30	0.05	0.21	0.15	0.67	95	0.07	0.29	0.04	0.18	0.21	0.83	95
M56-3	0.02	0.10	0.01	0.08	0.02	0.13	99.75	0.02	0.11	0.01	0.08	0.03	0.20	93.98
M59-1	0.02	0.09	0.01	0.06	0.05	0.29	159.34	**0.03**	0.11	0.02	0.07	0.13	0.43	155
M59-2	0.05	0.16	0.03	0.11	0.10	0.38	63.95	**0.06**	0.19	0.04	0.13	0.15	0.38	64.04
M59-3	0.03	0.20	0.02	0.13	0.06	0.40	65.05	**0.06**	0.24	0.04	0.17	0.14	0.40	64.04
M63-1	0.19	0.37	0.14	0.28	0.29	0.56	53.79	0.19	0.37	0.13	0.28	0.33	0.67	53.88
M63-3	**0.03**	0.20	0.02	0.14	0.07	0.33	63.29	0.02	0.17	0.01	0.10	0.11	0.67	63.58
M67-1	0.10	0.33	0.17	0.67	0.08	0.22	49.34	**0.18**	0.35	0.15	0.38	0.23	0.50	14.80
M67-2	0.06	0.26	0.05	0.22	0.09	0.56	40.68	0.06	0.20	0.04	0.13	0.12	0.44	30.97
M67-3	0.11	0.30	0.08	0.27	0.16	0.58	22.70	**0.18**	0.37	0.13	0.33	0.35	0.92	19.68
M67-4	0.04	0.19	0.03	0.15	0.07	0.27	51	**0.07**	0.18	0.05	0.11	0.17	0.45	33.6
M68-1	0.11	0.53	0.10	0.63	0.13	0.73	45.29	**0.12**	0.45	0.08	0.35	0.27	0.73	26.89
M68-2	0.09	0.24	0.08	0.24	0.11	0.35	29.02	**0.17**	0.40	0.12	0.29	0.29	0.65	23.06
M68-3	0.17	0.40	0.12	0.38	0.33	0.71	23	**0.20**	0.45	0.13	0.33	0.52	0.86	23

Table 6.2: Average and maximum F-measure, average and maximum precision value P, average and maximum recall value R and average Hausdorff distance values for the change points detected to all recordings per Mazurka. Mazurkas are indexed as "M<*opus*>-<*number*>.

the marking positions:

$$\mathrm{d}_H(M, C) = \left\{ \sup_{m \in M} \inf_{c \in C} \mathrm{d}(m, c), \sup_{c \in C} \inf_{m \in M} \mathrm{d}(m, c) \right\}, \tag{6.4}$$

where M and C are the sets of score-beat positions of the markings and the change points, respectively. The distance is equal to zero when there is a perfect match, and it is small when the maximum distance from $m \in M$ to the closest point $c \in C$ is small.

For each threshold in BS and PELT, we report the minimum Hausdorff distance in Table 6.2. For the PELT algorithm, the average Hausdorff distance is 48.2 and for BS is 56.4.

6.3.1 Conclusions and discussion

In this study we have introduced the use of change-point detection methods to determining changes in performed dynamic levels. We have shown that the PELT change-point algorithm performs better on average for identifying the positions of dynamic markings than the BS method. The Hausdorff distance shows that the PELT algorithm results in change points that are closer to the dynamic marking positions.

It is noticeable that the F-measure values are low in any case, however this is a first attempt to explore such music boundaries. One direction to interpret the results is the following observation. A common phenomenon was the presence of change points that appear in many recordings however they are not near to the markings that were tested. We found that a number of these popular change points corresponded to events such as the start of a ***crescendo***, a ***calando*** or a ***poco rit.*** In the following section we present a more systematic analysis of such cases.

6.4 Analysis of change-points locations

In this section we provide a thorough analysis of the meaning of change points extracted by PELT in terms of what they represent in the score. Each plot in Appendix B shows the bar chart of the change points across score-time in beats for all recordings per Mazurka, as well as the local maxima that are established, represented as peaks. The peaks that amass a number of change points being equal or more than the $\frac{2}{5}$ of the total number of recordings are highlighted by inserting the expressive score marking appearing in the specific beats. In the cases of Mazurka Op. 24 No. 2 and Mazurka Op. 63 No. 3 there was not any peak amassing such number of change points. In the cases where there is no

marking in that beat, we present the closest marking appearing in the range of a bar. In the cases where there is no marking in the range of a bar, a dash line is inserted. In Table 6.3 we have gathered all the score markings and expressions that have been present in change-point peak locations.

Dynamic markings	***pp***, crescendo,	***p***, diminuendo	***mf***,	***f***,	***ff***
Expressions	calando, sotto voce,	con forza, poco piu vivo	dolce,	perdendosi,	espressivo
Tempo markings	ralletando, rubato,	ritenuto, slentando	tenuto,	a tempo,	Tempo I
Accents	sf,	fermata			

Table 6.3: Markings that have been present in the score in positions of change-point peaks; by peaks we define the positions that amass a number of change points being equal or more than the $\frac{2}{5}$ of the total number of recordings in a specific Mazurka, meaning that more than this amount of recordings had a change point at that specific position.

Figure 6.2 presents the distribution of the meaning of all peaks of the change-points bar charts according to what appears in the score at the given position. A dynamic marking is more likely to appear, followed by a crescendo or diminuendo boundary position. The distribution of an accent follows up, the options of which are presented in Table 6.3, then the indication "other" which refers to unlabeled score positions, mostly a phrase or a motif boundary. These are followed by the distribution of a change point being a tempo marking or an expression, again the options of which are presented in Table 6.3.

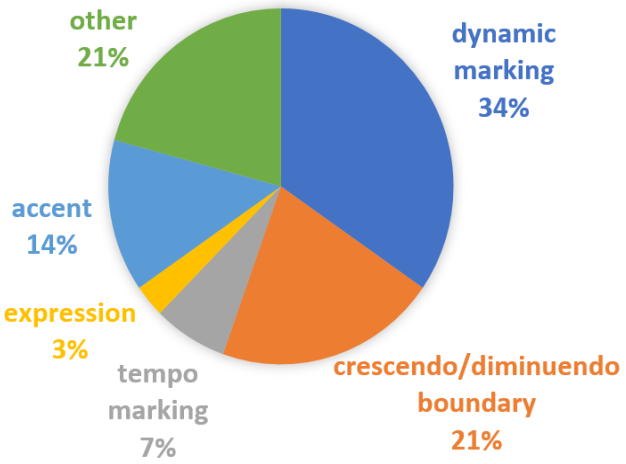


Figure 6.2: The distribution of what each of the peaks at change-point positions represent in the score.

Next we further investigate the locations where there is a peak however there is not any specific marking that indicate a change in the expressive style at the specific position (option "other" in Figure 6.2). More specifically, we count sixty-two such situations in our results. If we look through the score, we notice that most of the cases these positions indicate either a boundary of a phrase or a boundary of a motif in the melody. For each of the two cases, we introduce an example. In Figure 6.3 a phrase boundary from Mazurka Op. 33 No. 2 is presented; the position of the beat which is identified by a box was a change point for thirty-one out of fifty recordings. This example of a phrase boundary represents most of such events as it highlights the end of the melody, followed by score beats that include only the accompaniment. In Figure 6.4 a motif boundary from Mazurka Op. 30 No. 1 is presented where the box identifies the position of the score beat that was a change point for thirty-three out of fourty-five recordings.

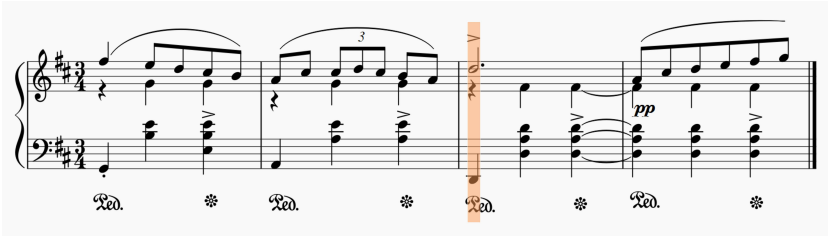


Figure 6.3: One phrase boundary detected in Mazurka Op. 33 No. 2 in beat 23–highlighted in box.

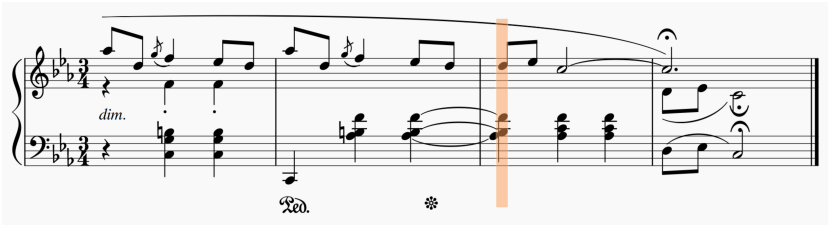


Figure 6.4: One motif boundary detected in Mazurka Op. 30 No. 1 in beat 155–highlighted in box.

6.4.1 Discussion

In the expanded study we introduce a more systematic way of detecting the change points in the loudness time series of our data. We then analyse the most popular locations of the change points across the recordings. The results show that positions having a dynamic marking in the score are captured. The

positions where a motif or a phrase boundary appear mostly come along with a change in tempo as well. This is a phenomenon that can be further analysed, especially by investigating an analogy to a tipping point as described in Chew [2013].

Chapter 7

Mapping between dynamic markings and performed loudness: A machine learning approach

Summary: In this chapter, we investigate the relationship between dynamic markings in the score and performed loudness by applying machine-learning techniques—decision trees, support vector machines, artificial neural networks, and a k-nearest neighbor method—to the prediction of loudness levels corresponding to dynamic markings, and to the classification of dynamic markings given loudness values. The methods are applied to forty-four recordings of performances of Chopin's Mazurkas each by eight pianists. The results show that loudness values and markings can be predicted relatively well when trained across recordings of the same piece, but fail dismally when trained across the pianist's recordings of other pieces, demonstrating that score features may trump individual style when modeling loudness choices. Evidence suggests that all the features chosen for the task are relevant, and analysis of the results reveal the forms (such as the return of the theme) and structures (such as dynamic marking repetitions) that influence predictability of loudness and markings.

This chapter is organized as follows: In Section 7.1 we describe the data set used for this study as well as the features extracted and in Section 7.2 the learning task and the algorithms employed. Section 7.3 presents and discusses the results obtained when predicting loudness levels at dynamic markings; and, Section 7.4 does the same for results obtained when classifying loudness levels into dynamic markings. Finally, Section 7.5 summarizes the conclusions with some general discussions.

7.1 Material

For the purpose of this study, we wanted to get the biggest possible number of Mazurka pieces played by the biggest possible number of same pianists, therefore we examine recordings of eight pianists' performances of forty-four Mazurkas. The Mazurkas, the number of dynamic markings, as well as the method to extract the loudness information for each marking are detailed in Chapter 3. The eight pianists together with the recording's year and index are identified in Table 7.1.

Pianist	Chiu	Smith	Ashkenazy	Fliere	Shebanova	Kushner	Barbosa	Czerny
Year	1999	1975	1981	1977	2002	1990	1983	1989
ID	P1	P2	P3	P4	P5	P6	P7	P8

Table 7.1: Pianist's name, year of the recording, and pianist ID.

For each dynamic marking in the dataset, we have extracted the following associated features:

1. label of the current dynamic marking (M);
2. label of the previous dynamic marking (MPR);
3. label of the next dynamic marking (MN);
4. distance from the previous dynamic marking (PRD);
5. distance to the next dynamic marking (ND);
6. nearest non-dynamic marking annotation between the previous and current dynamic marking, e.g. ***crescendo*** (PRA);
7. nearest non-dynamic marking annotation between current and next dynamic marking, e.g. ***crescendo*** (NA); and,
8. any qualifying annotation appearing simultaneously with the current dynamic marking, e.g. ***dolcissimo*** (MA).

In addition to the feature set described above, we also have an associated loudness value, L, which is the ℓ_b value on the beat of the dynamic marking.

We consider at most one value each for the features "PRA," "NA," and "MA." If two different annotations occur on the same beat, we choose the one related to dynamic changes, an exception being the case where the annotations ***sf*** and ***a tempo*** appear simultaneously. In this case, to be consistent with the time range of other non-dynamic marking annotations, we choose ***a tempo*** over ***sf***, as it applies to more than one score beat. In the case where there was an annotation related to change in dynamics and a qualifying term such as ***poco*** preceding it, we use the annotation without the qualifier, to limit the number of dynamic terms.

7.2 Learning task and algorithms

In this chapter we explore different machine learning techniques to induce a model for predicting the loudness level at particular points in the performance. Concretely, our objective is to induce a regression model M of the following form:

$$M(FeatureSet) \rightarrow Loudness$$

Where M is a function which takes as input the set of features (*FeatureSet*) described in the previous section, and *Loudness* is the predicted loudness value. In order to train M we have explored the following machine learning algorithms (as implemented in Hall et al. [2009]):

Decision Trees (DT). Decision trees by Quinlan [1986] use a tree structure to represent possible branching on selected attributes so as to predict an outcome given some observed features. The decision tree algorithm recursively constructs a tree by selecting at each node the most relevant attribute. The selection of the most relevant attribute, at each node of the tree is based on the *information gain* associated with each attribute and the instances at each node of the tree. For a collection of loudness values, suppose there are b instances of class B and c instances of class C. An arbitrary object will belong to class B with probability $b/(b+c)$ and to class C with probability $c/(b+c)$. The expected information needed to generate the classification for the instance is given by

$$I(b,c) = -\left(\frac{b}{b+c}\log_2\frac{b}{b+c} + \frac{c}{b+c}\log_2\frac{c}{b+c}\right). \tag{7.1}$$

Suppose attribute A can take on values $\{a_1, a_2, ..., a_v\}$ and is used for the root of the decision tree, partitioning the original dataset into v subsets, with the i-th subset containing b_i objects of class B and c_i of class C. The expected in-

formation required for the i-th subtree is $I(b_i, c_i)$, and the expected information required for the tree with A as root is the weighted average

$$E(A) = \sum_{i=1}^{v} \frac{b_i + c_i}{b + c} I(b_i, c_i), \tag{7.2}$$

where the weight for the i-th subtree is the proportion of the objects in the i-th subtree. The information gained by branching on A is therefore gain$(A) = I(b, c) -$ E(A), and the attribute chosen on which to branch at the next node is thus $\arg\max_A gain(A)$. Decision trees can also be used for predicting numeric quantities. In this case, the leaf nodes of the tree contain a numeric value that is the average of all the training set values. Decision trees with averaged numeric values at the leaves are called *regression trees*.

Support Vector Machines (SVM). Support vector machines aim to find the hyperplane that maximizes the distance from the nearest members of each class, called support vectors. Cristianini and Shawe-Taylor [2000] use a nonlinear function $\phi(\cdot)$ to map attributes to a sufficiently high dimension, so that the surface separating the data points into categories becomes a linear hyperplane. This allows the model to predict nonlinear models using linear methods. The data can be separated by a hyperplane and the support vectors are the critical boundary instances from each class.

The process of finding a maximum margin hyperplane only applies to classification. However, support vector machine algorithms have been developed also for regression problems, i.e. numeric prediction, that share many of the properties encountered in the classification case. SVM for regression problems also produce a model that can usually be expressed in terms of a few support vectors and can be applied to nonlinear problems using kernel functions.

Artificial Neural Networks (ANN). Artificial neural networks are generally presented as systems of interconnected "neurons" which exchange messages between each other. They are based on the idea of the perceptron which includes an input, a hidden and an output layer of data points connected with nodes having numeric weights. The nodes of the input layer are passive, meaning they do not modify the data. The two aspects of the problem is to learn the structure of the network, and to learn the connection weights. A multilayer perceptron (MLP)–or ANN– has a linear activation function in all neurons, that is, a linear function that maps the weighted inputs to the output of each neuron. More formally, we assume that in one network there are d inputs, M hidden units and c output units. As it is described by Bishop [1995], the output of the jth hidden unit is obtained by first forming a weighted linear combination of the d

input values, to give

$$a_j = \sum_{i=1}^{d} w_{ji}^{(1)} x_i, \tag{7.3}$$

where $w_{ji}^{(1)}$ denotes a weight in the first layer, going from input i to hidden unit j. The activation of hidden unit j is then obtained by transforming the linear sum above using an activation function $g(\cdot)$ to give

$$z_j = g(a_j). \tag{7.4}$$

The outputs of the network are obtained by transforming the activations of the hidden units using a second layer of processing elements. For each output unit k, we construct a linear combination of the outputs of the hidden units of the form

$$a_k = \sum_{j=1}^{M} w_{kj}^{(2)} z_j. \tag{7.5}$$

We then obtain the activation of the kth output unit by transforming this linear combination using a non-linear activation function $\tilde{g}(\cdot)$, which returns

$$y_k = \tilde{g}(a_k). \tag{7.6}$$

We next consider how such a network can learn a suitable mapping from a given dataset. Learning is based on the definition of a suitable error function, which is minimized with respect to the weights and biases in the network. "If we define a network function, such as the sum-of-squares error [...] which is a differentiable function of the network outputs, then this error is itself a differentiable function of the weights. We can therefore evaluate the derivatives of the error with respect to the weights, and these derivatives can then be used to find weight values which minimize the error function." (Bishop [1995])

In this chapter, in order to determine the weights, that is, to tune the neural network parameters to best fit the training data, we apply the gradient descent back propagation algorithm of Chauvin and Rumelhart [1995]. The back propagation algorithm is used for evaluating the derivatives of the error function and learns the weights for a multi-layer perceptron, given a network with a fixed set of units and interconnections. The idea behind this algorithm is that the output corresponds to a propagation of errors backwards through the network. We empirically set the momentum applied to the weights during updating to 0.2 and the learning rate, that is the amount of the weights that are updated, to 0.3. We use a fully-connected multi-layer neural network with one hidden layer meaning that we have one input and one output neuron for each attribute.

k-Nearest Neighborhood (k-NN). k-NN is a type of instance-based learning which instead of performing explicit generalization, compares the new data instances with instances in the traning set previously stored in memory. In k-NN generalization beyond the training data is delayed until a query is made to the system. The algorithm's main parameter is k, the number of considered closest training vectors in the feature space. Given a query, the output's value is the average of the values of its k nearest neighbors. More formally, as it is described by Hastie et al. [2001], given a set of measurements (x_i, y_i), the k-nearest neighbor fit for the prediction $\hat{Y}$ of the output Y is

$$\hat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i, \tag{7.7}$$

where $N_k(x)$ is the neighborhood of the input instance x defined by the closest points x_i in the training sample.

In this chapter we report the results of the k-NN algorithm with $k = 1$ which finds the training instance closest in Euclidean distance to the given test instance. If several instances are qualified as the closest, the first one found is used. We tested the values of k equal to 1,2, and 3 and there was not a significant difference on the regression results.

7.3 Performed loudness-level modeling

This section assesses the fit of the machine-learned models as they predict loudness values in a pianist's recording of a Mazurka, first given other recordings of that Mazurka, and second given other recordings of that pianist. Two experiments have been conducted for this purpose. In the first one, each prediction model has been trained for each Mazurka separately. Then, each model has been evaluated by performing a 8-fold training-test validation in which instances of one pianist of the training set are held out in turn as test data while the instances of the remaining seven pianists are used as training data. In the second one, each prediction model has been trained for each pianist separately. Then each model has been evaluated by performing a 44-fold training-testing validation in which instances of one Mazurka of the training set are held out in turn as test data while the instances of the remaining forty-three Mazurkas are used as training data.

Section 7.3.1 presents the machine-learning algorithms' results of the first experiment, when predicting loudness given other pianists' recordings of the target piece; Section 7.3.2 presents the machine-learning algorithms' results of the second experiment, when predicting loudness given the target pianist's other

recordings; Section 7.3.3 considers the extremes in prediction results for the second experiment; Section 7.3.4 considers the degree of similarity in approach to loudness between pianists; and, Section 7.3.5 considers the relevance of the features selected.

7.3.1 Predicting loudness given other pianists' recordings of target piece

In the first experiment, we use the machine-learning methods described above to predict the loudness values at the dynamic markings of one *Mazurka*, given the loudness levels at the markings of the same *Mazurka* recorded by other pianists. We test the machine-learning models by assessing their predictions at points where the composer (or editor) has placed a dynamic marking because this is our closest source of ground truth. The evaluations focus on how well a pianist's loudness choices can be predicted given those of other seven pianists for the same Mazurka.

As a measure of accuracy, we compute the Pearson correlation coefficient between predicted (X) and actual (Y) loudness values using the formula

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{Y})^2}}, \tag{7.8}$$

where $x_i \in X$, $y_i \in Y$, and the size of X and of Y, n, varies from one Mazurka to the next, and is given in Table 3.1.

For each Mazurka, we compute the mean Pearson correlation coefficients over all recordings of that Mazurka to produce the graph in Figure 7.1. The results of each machine-learning algorithm—Decision Trees (DT), Support Vector Machines (SVM), Artificial Neural Networks (ANN), and k-Nearest Neighbor (k-NN)—is denoted on the graph using a different symbol.

Observe in Figure 7.1 that, with few exceptions, the mean Pearson correlation coefficient is fairly high. When disregarding the two most obvious outliers, Mazurka Op. 17 No. 3 and Mazurka Op. 24 No. 3, the mean Pearson correlation value ranged from 0.5192 to 0.9667 over all machine learning methods. Furthermore, the mean Pearson correlation coefficient, averaged over the four machine-learning techniques, and over all Mazurkas is equal to 0.8083. This demonstrates that, for most Mazurkas, the machine-learning methods, when trained on data from other pianists' recordings of the Mazurka, can reasonably predict the loudness choices of a pianist for that Mazurka.

In the next sections, we inspect the anomalous situation of Mazurka Op. 17 No. 3 and Mazurka Op. 24 No. 3, and the special cases when one particular

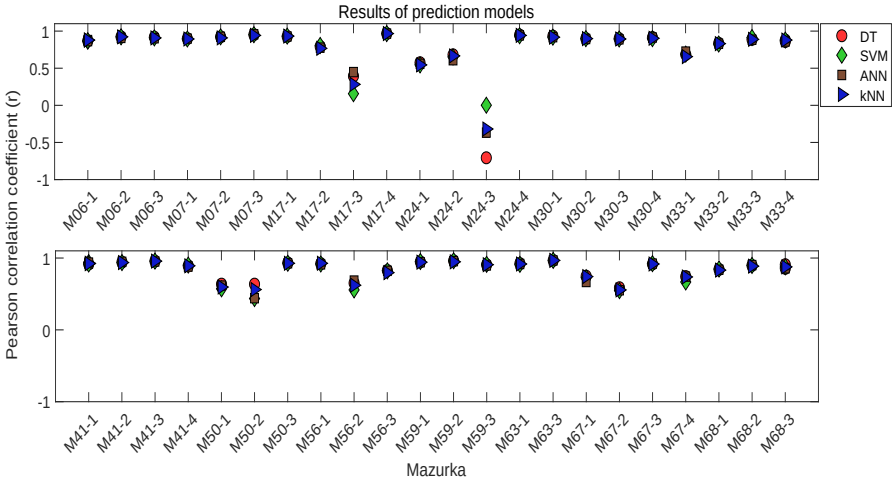


Figure 7.1: Pearson correlation coefficient between predicted and actual loudness values for each Mazurka, averaged over all recordings of the Mazurka, for each machine-learning method—Decision Trees (DT), Support Vector Machines (SVM), Artificial Neural Networks (ANN), and k-Nearest Neighbor (k-NN).

pianist deviates from the behavior of other pianists for specific Mazurkas.

Cases of low correlation between predicted and actual loudness values

While the overall Pearson correlation measure of success is good for the prediction of loudness values in a Mazurka recording when a machine-learning model is trained on other pianists' recordings of the Mazurka, two Mazurkas were found to be outliers to this positive result: Mazurka Op. 24 No. 3 and Mazurka Op. 17 No. 3.

For Mazurka Op. 24 No. 3, the mean (over all recordings) Pearson correlation value, when averaged over the four machine-learning techniques, is -0.347, meaning that the predictions are weakly negatively correlated from the actual loudness values. The mean Pearson correlation value for Mazurka Op. 17 No. 3, when averaged over the four machine-learning methods, while positive, is low at 0.320, meaning that the predictions are only weakly correlated with the actual loudness values.

Looking deeper into the case of these two Mazurkas, apart from the common key of A♭ major, they also share the property of having only the dynamic markings ***p*** and ***mf*** in the score, with extended mono-symbol sequences of ***p***'s. The other Mazurkas do not have the property of including only ***mf*** and ***markings***. In the case of Mazurka Op. 24 No. 3, the existing score markings are {***mf***, ***p***, ***p***, ***p***, ***p***, ***p***, ***p***}; for Mazurka Op. 17 No. 3, the score markings are {***mf***, ***p***, ***mf***, ***p***, ***p***, ***p***, ***p***, ***mf***, ***p***}. The narrow dynamic range of the notated symbols and the consecutive strings of the same symbols both will almost certainly lead to a

wide range of interpretations in order to create dynamic contrast and narrative interest.

Consider the case of Mazurka Op. 24 No. 3: Figure 7.2 shows the actual loudness values for the eight recordings at the points of the dynamic markings. Note that, in this Mazurka, apart from the initial ***mf***, the remaining dynamic markings are uniformly ***p***. The x-axis marks the sequence of dynamic markings in the score, and the eight distinct symbols on the graphs mark the loudness values (in sones) at these points in the recording. Note the wide range of interpretation of the loudness level for ***mf***; the loudness value ranges for the ***p***'s are often as wide as that for the ***mf***, with the recordings exhibiting many contradictory directions of change from one dynamic marking to the next. In particular, note that in three out of the eight recordings, the ***p*** immediately following the ***mf*** is actually louder than the ***mf***.

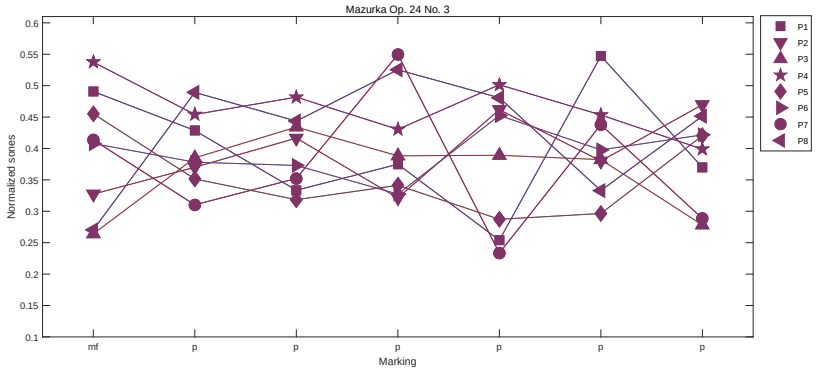


Figure 7.2: Representation of the loudness levels on the marking positions in score time for Mazurka M24-3 for the eight pianists.

For the case of Mazurka Op. 24 No. 3, the Pearson correlation coefficient between the predicted and actual loudness values, for each of the four machine-learning methods, are uniformly negative. Next, we consider the cases when the predictions are negatively correlated for only one recording of a Mazurka while the remaining seven recordings of the same Mazurka had predictions positively correlated with the actual loudness values.

Cases when one recording is negatively correlated while others are not

The tests in the previous section showed that there can be a high degree of divergence in loudness interpretation amongst recordings of a Mazurka. In this section, we examine the special cases of solitary deviant behavior, when one pianist chose starkly different loudness strategies than the others when recording a

Mazurka. For this, we consider the cases when the predictions for one recording has a negative average (over all four machine-learning techniques) Pearson correlation value while those for the other seven have positive average correlation coefficients between predicted and actual loudness values.

We identified four Mazurkas for which the predictions for one recording was negatively correlated on average (over all the machine-learning algorithms) and the other seven were positively correlated. The average Pearson correlation values for each pianist's recordings of these four Mazurkas are plotted in Figure 7.3. The average correlation value of the solitary deviant recording for each Mazurka is highlighted with a red dot. For each red dot marking, the correlation value of the worst-predicted pianist, the correlation values of the other pianists who recorded the Mazurka are shown as grey squares in the same vertical column.

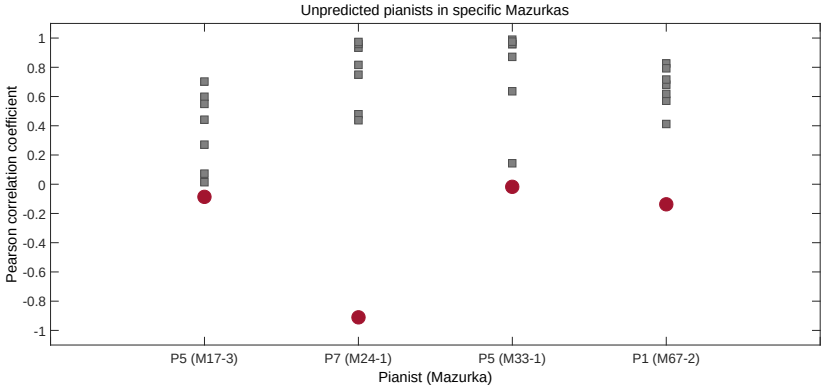


Figure 7.3: Pearson correlation coefficients of Mazurkas for which the worst predicted pianist's recording scored, averaging over all machine-learning methods, a negative r value (red dots): pianist P5 for M17-3, P7 for M24-1, P5 for M33-1, and P1 for M67-2. The average coefficients for the remaining pianists are shown as grey squares.

We can see from the display that even when the machine-learning algorithms did poorly in the case of a particular pianist, they often did fairly well for other pianists who recorded the same Mazurka. Figure 7.3 demonstrates why the loudness values of certain Mazurka recordings cannot be predicted well when machine-learning algorithms are trained on other pianists' recordings of the same Mazurka.

We next turn our attention to the extreme case of Mazurka Op. 24 No. 1, in which the loudness value predictions for one recording, that of Barbosa (P7), are almost perfectly negatively correlated with the actual values. Figure 7.4 shows the loudness time series, in score time, for all eight recordings of Mazurka Op. 24 No. 1. The loudness time series for Barbosa, the worst-predicted pianist, is highlighted in bold. The score dynamic markings for this Mazurka are labeled

on the x-axis. The actual loudness values recorded for pianist P7 at these points are marked by red dots. The actual loudness values of other pianists——Chiu (P1), Smith (P2), Ashkenazy (P3), Fliere (P4), Shebanova (P5), Kushner (P6), Czerny (P8)—at this point are marked by black symbols on the same vertical line.

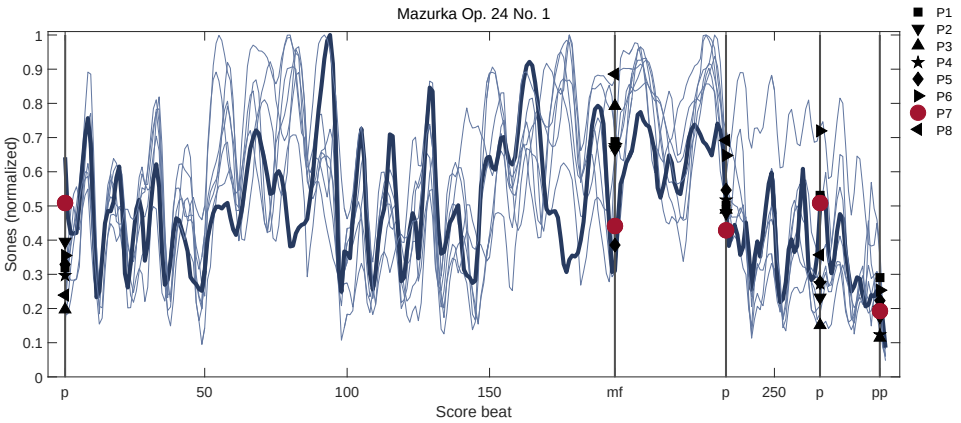


Figure 7.4: Loudness time series in score-beat time for recordings of Mazurka Op. 24 No. 1. Loudness time series for Barbosa's recording (P7) is shown in bold; loudness value for Barbosa's recording at dynamic markings are indicated by red dots, those of other pianists——Chiu (P1), Smith (P2), Ashkenazy (P3), Fliere (P4), Shebanova (P5), Kushner (P6), Czerny (P8)—are shown as black symbols.

As illustrated by the graph, in Barbosa's recording, he employs a performance strategy contrary to that of all or almost all of the other pianists. Furthermore, the loudness level of Barbosa's recording sampled at the points of the first four dynamic markings are relatively level, in contrast to the strategies exhibited by the other pianists.

Having considered the prediction of loudness values in a Mazurka recording by training machine-learning algorithms on recordings of the same Mazurka by other pianists, we next consider predictions of models trained on recordings of other Mazurkas by the same pianist.

7.3.2 Predicting loudness given target pianist's recordings of other pieces

In the second experiment, we use the machine-learning methods to predict the loudness values at the dynamic markings of one Mazurka, given the loudness levels at the markings of other Mazurkas recorded by the same pianist. The evaluations focus on how well a pianist's loudness choices can be predicted given those of them made in the other forty-three Mazurkas. For this purpose,

a 44-fold training-testing validation has been implemented. As before, we use as measure of prediction accuracy the Pearson correlation coefficient between actual and predicted values.

The results are displayed in Figure 7.5, which shows the mean, minimum, maximum, and median Pearson correlation coefficient values over all Mazurka recordings by each of the pianists listed in Table 7.1; the results are broken out into the four machine-learning methods employed—Decision Trees (DT), Support Vector Machines (SVM), Artificial Neural Networks (ANN), and k-Nearest Neighbor (k-NN).

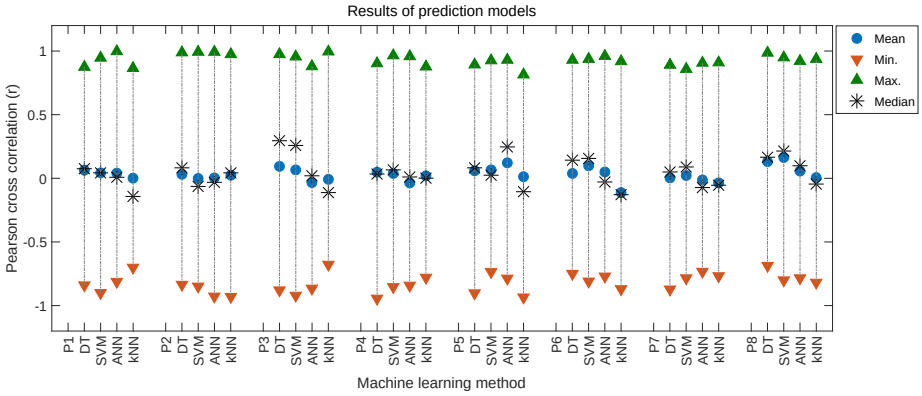


Figure 7.5: Pearson correlation coefficient mean, min, max, and median values for each method—Decision Trees (DT), Support Vector Machines (SVM), Artificial Neural Networks (ANN), and k-Nearest Neighbor (k-NN)—for each pianist—Chiu (P1), Smith (P2), Ashkenazy (P3), Fliere (P4), Shebanova (P5), Kushner (P6), Barbosa (P7), Czerny (P8).

Contrary to the previous experiment, where the machine-learning models were trained on other pianists' recordings of the same Mazurka and did fairly well, when training the models on the same pianist's other Mazurka recordings, the average cross-validation results for each pianist are close to zero. The minimum is close to -1, implying that sometimes the loudness values of a recording can be directly contrary to the predictions, and the maximum is close to 1, implying that sometimes the loudness values behave as predicted. The results thus demonstrate that it can be extremely difficult to predict loudness values in a recording given the pianist's past behavior in recordings of other pieces. The results fare far better when training on other pianists' recordings of the same piece.

In Section 7.3.3, we seek to gain some insights into why the Pearson correlation values were so highly variable between the predicted and actual values for this experiment. In particular, we examine in detail the Mazurkas for which predicted and actual loudness values were most strongly and positively correlated

and most strongly and negative correlated across all machine-learning methods.

The previous results have shown that while there may be some small variability in the prediction quality of the four machine-learning methods, they agree on the prediction difficulty amongst the recordings. In the next section, we perform a second check on the prediction quality using a Euclidean measure.

Comparing machine-learning methods

To check for variability in the prediction quality of the four machine-learning algorithms, we compute the accuracy for each algorithm using the Euclidean distance. The Euclidean distance between the predicted (X) and actual (Y) loudness values (in sones) in the held out data is given by the formula

$$d(X, Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}, \tag{7.9}$$

where $x_i \in X$, $y_i \in Y$, and n is the size of X, and Y, which varies from one Mazurka to the next.

We average these Euclidean distances over all Mazurkas for a given pianist to produce the results shown in Figure 7.6. The results are grouped by pianist: Chiu (P1), Smith (P2), Ashkenazy (P3), Fliere (P4), Shebanova (P5), Kushner (P6), Barbosa (P7), Czerny (P8). For each pianist, the graph shows accuracy results for each of the four machine-learning methods—Decision Trees (DT), Support Vector Machines (SVM), Artificial Neural Networks (ANN), and k-Nearest Neighbor (k-NN)—averaged over all Mazurka recordings by that pianist.

The results show that the algorithms perform consistently one relative to another. The span of average Euclidean distance between predicted and actual values is relatively small. In comparison, the DT algorithm produced the best results, followed closely by the SVM algorithm then the k-NN algorithm; the ANN algorithm is a more distant fourth. While the ANN algorithm fared worse in this exercise, we shall see in Section 7.4.2 that the ANN gives better results when the machine-learning algorithms are applied to the problem of predicting dynamic marking labels.

7.3.3 A pianist's interpretation may not be predictable based on his/her approach to other pieces

In this section, we dig deeper into the Pearson correlation results of the previous section to consider the extreme cases of when the predicted and actual loudness values are most consistently positively and consistently negatively correlated. We consider the Mazurkas for which all methods produced loudness predictions

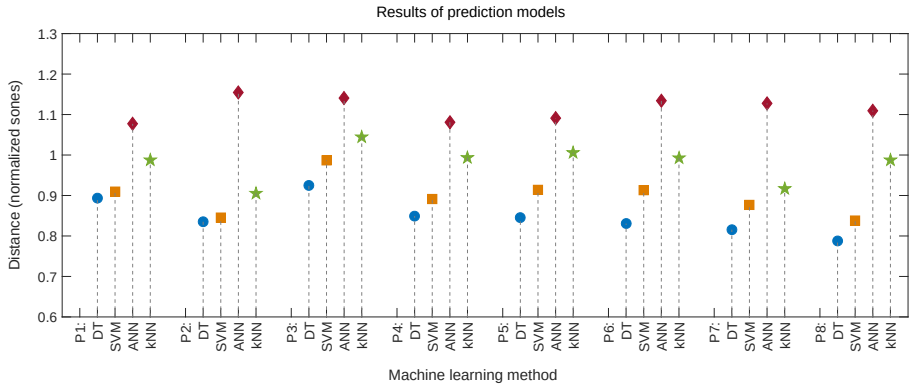


Figure 7.6: Euclidean distance between predicted and actual loudness values for each pianist——Chiu (P1), Smith (P2), Ashkenazy (P3), Fliere (P4), Shebanova (P5), Kushner (P6), Barbosa (P7), Czerny (P8)—for each method—Decision Trees (DT), Support Vector Machines (SVM), Artificial Neural Networks (ANN), and k-Nearest Neighbor (k-NN)—averaged over all Mazurka recordings by the pianist.

that were negatively correlated with the actual loudness values for all pianists; we also single out the Mazurkas for which all methods produced predictions that were positively correlated with the actual values for all pianists.

Most negatively and most positively correlated results

In the cross-validation, the highest Pearson correlation coefficient between predicted and actual loudness, 0.9981, is encountered in Chiu (P1)'s recording of Mazurka Op. 63 No. 3 (M63-3), and the lowest correlation value, −0.9444, in Fliere (P4)'s recording of Mazurka Op. 17 No. 2 (M17-2).

The four Mazurkas for which the Pearson correlation is negative for all the machine-learning methods and for all eight pianists are Mazurkas Op. 7 No. 1 (M07-1), Op. 24 No. 2 (M24-2), Op. 24 No. 4 (M24-4), and Op. 50 No. 3 (M50-3). This means that, for these Mazurkas, the pianists' recorded loudness strategies are contrary to those gleaned from their recordings of the other Mazurkas. The three Mazurkas for which the Pearson correlation coefficient over all pianists and for all machine-learning methods was positive are Mazurkas Op. 30 No. 4 (M30-4), Op. 41 No. 2 (M41-2), and Op. 68 No. 2 (M68-2). For these Mazurkas, the pianists' recorded loudness strategies are in accordance to those gleaned from their recordings of the other Mazurkas.

The results are summarized in Figure 7.7, which presents the Pearson correlation result for each of the eight pianists for the Mazurkas mentioned; each pianist's data point for a Mazurka shows the average over the four machine-learning methods. Note that, for each Mazurka, the correlation coefficients are

relatively closely grouped for all pianists. In the next section, we shall examine more closely the four Mazurkas having all-negative correlation values, i.e. the ones to the left of the dividing line.

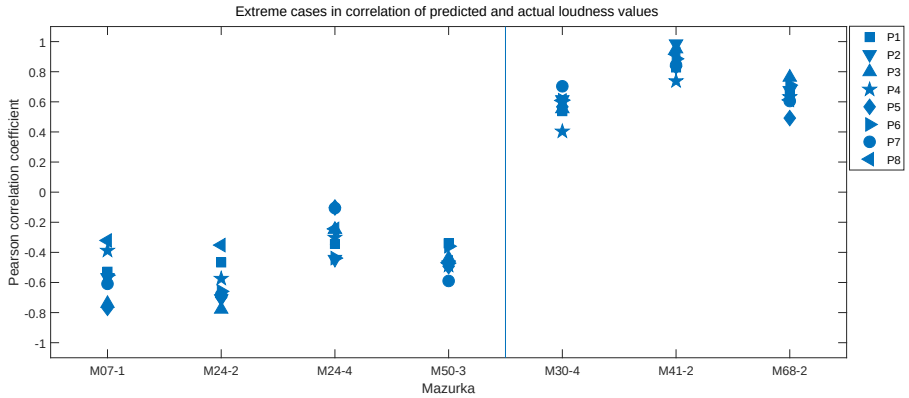


Figure 7.7: Pearson correlation coefficient values for each pianist, averaged over all machine-learning methods, for the Mazurkas having all negative (left: M07-1, M24-2, M24-4, M50-3) and all positive (right: M30-4, M41-2, M68-2) correlation values.

Cases of negative correlation

Here, we focus on the four Mazurkas with negative Pearson correlation coefficients for all pianists and all machine-learning methods. Figure 7.8 shows the predicted and actual loudness values at each dynamic marking in the four Mazurkas, concatenated in sequence. The values shown are the average over all pianists and all machine-learning methods for each marking and Mazurka.

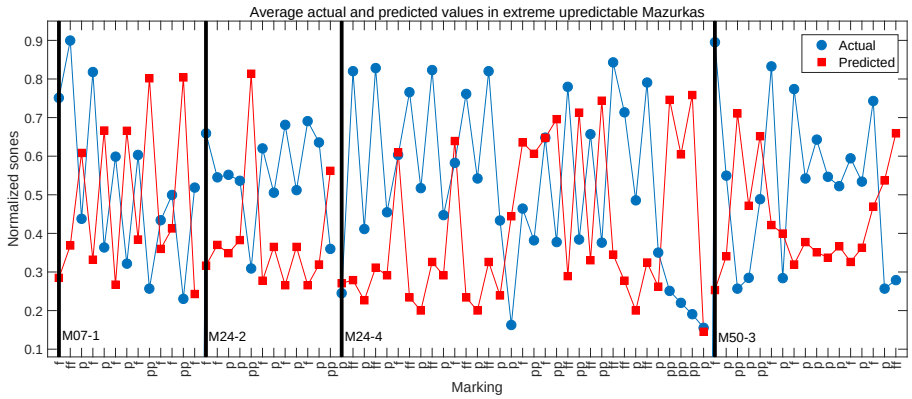


Figure 7.8: Average actual (circles) and predicted (squares) dynamic values at dynamic marking positions for the Mazurkas M07-1, M24-2, M24-4, and M50-3 where the average of the Pearson correlation coefficient over pianists and methods was negative.

As can be seen, for most dynamic markings in these four Mazurkas, the actual loudness levels are the opposite of the predicted loudness values. The difference in loudness strategy may be due to deliberate contrarian behavior, or to aspects of the piece relevant to dynamic choices not being captured in the selected features. A discussion on future directions for feature analysis will be presented in Section 7.3.5.

There may also be score-inherent factors in these four Mazurkas that contributed to the negative correlation results. These include the oscillatory nature—and hence longer range dependencies—of some of the dynamic marking sequences, and the presence of sequences of identical markings—which lead to greater variability in interpretation. Instances of oscillatory sequences include the sequence (***p***, ***f***, ***p***, ***f***, ***p***, ***f***) in Mazurka Op. 7 No. 1, the sequence (***f***, ***p***, ***f***, ***p***, ***f***, ***p***) in Mazurka Op.24 No. 2, and the sequence (***pp***, ***ff***, ***pp***, ***ff***, ***pp***, ***ff***) in Mazurka 24 No. 4. Examples of monosymbol sequences include the sequence of three ***pp***'s in Mazurka 24 No. 4, and the sequence of four ***p***'s in Mazurka 50 No. 3.

The analysis of the results of the loudness value prediction experiments leads us to suspect that the poor predictability of some of the Mazurka recordings may be due to high variability in performed loudness strategies among the pianists for the range of Mazurkas represented. Hence, in the next section, we describe and report on an experiment to test the degree of similarity in the loudness strategies employed by one pianist vs. that used by another.

7.3.4 Inter pianist similarity

To determine the degree of similarity in the loudness strategies employed by different pianists, machine-learning models were trained on all Mazurka recordings by one pianist and used to predict the loudness values in all Mazurka recordings by another pianist. The goodness of fit is measured using the mean (over all Mazurka recordings) of the average (over all machine-learning methods) Pearson correlation coefficient.

The results are shown in Figure 7.9 in the form of a matrix, where the (i,j)-th element in the matrix represents the percentage of the mean averaged Pearson correlation coefficient between the loudness value predictions of the model trained on data of pianist i and the corresponding actual loudness data of pianist j. The correlation values range from 68.36% to 78.43%. This shows the high level of similarity between pianists for interpreting dynamic markings in the pieces investigated.

Higher correlation values are observed in columns P1 and P5. We can deduce that Chiu (P1)'s interpretation model, followed closely by Shebanova P5's

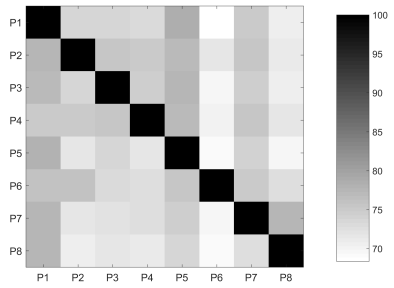


Figure 7.9: Matrix displaying Pearson correlation coefficient values (in %), averaged over all machine-learning methods, when applying a model trained on recordings by one pianist to predict the loudness values of another pianist.

model, best predicts the loudness values of recordings by other pianists. Note that Chiu's recordings also achieved some of the highest correlation values, meaning that they were most predictable. Furthermore, the loudness strategies of pianists P1 and P5 best fit each other, i.e. the highest non-diagonal Pearson correlation value is found when P1's model is used to predict the loudness levels in P5's recording, and vice versa. On the other hand, the loudness strategies of pianist P6 is the one most dissimilar to that in other recordings, as shown by the almost white non-diagonal squares in column P6.

7.3.5 Discussion on feature analysis

The k-NN method is known to be highly sensitive to irrelevant features, i.e. it performs considerably less well than other algorithms in the presence of irrelevant features. As the results show no demonstrable trend in this respect, this leads us to think that all the extracted features in our feature set are indeed relevant.

A natural question that follows is: which of the extracted features are more salient for predicting performed loudness? To investigate this question, we apply the RELIEF algorithm from Robnik-Sikonja and Kononenko [1997] to rank the features according to relevance for predicting loudness. The RELIEF algorithm evaluates the worth of an attribute by repeatedly sampling an instance and considering the value of the given attribute for the nearest instance of the same and different class.

The results are shown in Table 7.2, with features being ranked from most important (1) to least important (8). The feature-relevance analysis indicates that the current dynamic marking proved to be the most informative feature for the task, followed by the text qualifiers, and the preceding dynamic marking. Interestingly, the ranking of features according to relevance for predicting loudness is the same for all pianists, which may indicate that all the considered

pianists give the same relative importance to the investigated features when deciding loudness levels in their performances.

Ranking	Feature	Ranking	Feature	Ranking	Feature
1	M	4	MN	7	ND
2	MA	5	PRD	8	NA
3	PRA	6	MPR		

Table 7.2: Ranking of importance of the features for the loudness prediction task.

In order to evaluate the incremental contribution of each of the features studied, we have implemented the machine learning models using different subsets of features. More concretely, we have considered feature subsets by incrementally adding features to the training set one at a time, starting with the most important feature, i.e. the highest ranked, and continuing to add features according to their rank order. In Figure 7.10 we present the results for each pianist's recording, averaged over all machine learning methods.

As can be seen in the figure, for all pianists, the loudness prediction accuracy for their recordings, with few exceptions, increases monotonically with the number of features. This confirms our belief that all the features studied are indeed relevant and contribute to the accuracy of the loudness level predictions for all recordings. It is worth noticing that the highest ranked feature, i.e. the dynamic marking, contains on its own substantial predictive power: the correlation coefficient of the loudness models induced with only dynamic marking information alone ranges from 0.64 (for pianist P6's recordings) to 0.75 (for pianist P1's recordings).

7.4 Dynamic marking prediction

Classification remains one of the staple tasks of machine-learning algorithms. This section assesses the fit of the machine-learned models as they are applied to the classification of a Mazurka recording's loudness values into loudness categories as indicated by dynamic markings. In particular, we examine the following problem: given a loudness level in a recorded Mazurka, what is the dynamic marking that best represents conceptually the loudness value at a particular instance in time.

As in Section 7.3, we have conducted two experiments. In the first experiment each classification model has been trained for each Mazurka separately and a 8-fold cross-validation implemented for predicting the dynamic markings given other pianists' recordings of the target piece. In the second experiment

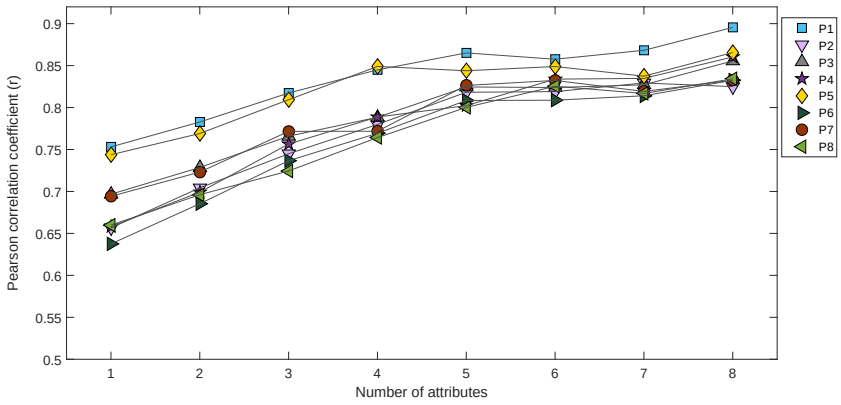


Figure 7.10: Upper bound for average (over all machine learning methods) Pearson correlation coefficient as number of features considered increases; features are added in the order given in Table 7.2. Numbers result from the optimal case where the testing data is equal to the training data.

each classification model has been trained for each pianist separately and a 44-fold cross-validation has been implemented for predicting the dynamic markings of one Mazurka given the other Mazurka recordings of the same pianist. For evaluation, we compute the percentage of correctly classified instances.

In the remaining part of this section we present the analysis of the results of the two experiments. More specifically, Section 7.4.1 presents and discusses the results for predicting dynamic markings given other pianists' recordings of the target piece; Section 7.4.2 does the same for predicting markings given the target pianist's recordings of other pieces; Section 7.4.3 looks at which of the dynamic markings are more easily classified than others; and, Section 7.4.4 evaluates the ease or difficulty of predicting the markings of a Mazurka using the ratio of correctly classified markings.

7.4.1 Predicting dynamic markings given other pianists' recordings of target piece

In the first experiment, we use the four machine-learning methods to predict the dynamic-marking labels of loudness values given the loudness-marking data of other pianists' recordings of the target piece. The 8-fold cross-validation test results of this task give us an average of 99.22% of Correctly Classified Instances (CCI) over all methods. The result above means that it is highly plausible to train a machine-learning algorithm on other pianists' performance of a Mazurka to predict the dynamic marking labels of the target recording. The constraints imposed by the features available when making a prediction, such as the exact labels in their values enhance this direction.

We next move on to the complementary experiment in which the machine-learning algorithms predict dynamic-marking labels based on loudness-marking data of other Mazurka recordings by the target pianist.

7.4.2 Predicting dynamic markings given target pianist's recordings of other pieces

In this experiment, we train classifiers—using DT, ANN, SVM, and k-NN algorithms—on the loudness-label mappings of other Mazurka recordings by the same pianist in order to identify the dynamic marking corresponding to a given loudness value in the target recording. Table 7.3 shows the results obtained by the classifiers in terms of the mean percentage of Correctly Classified Instances (CCI) over all Mazurkas.The highest mean CCI values for each pianist are highlighted in bold.

Method \ Pianist	P1	P2	P3	P4	P5	P6	P7	P8	AVG
DT	27.822	26.387	28.012	24.542	24.798	27.858	29.230	26.437	26.8858
SVM	27.735	25.283	23.962	22.830	25.283	24.717	26.227	25.670	25.2134
ANN	**30.755**	27.925	27.736	**26.981**	**26.604**	**31.132**	30.189	**30.460**	**28.9727**
k-NN	28.301	**28.491**	**28.113**	25.660	25.660	26.981	**30.377**	27.547	27.6412

Table 7.3: Percentage of Correctly Classified Instances (CCI). Maximum values per pianist are highlighted in bold. The last column contains the average values per row and the highest average of CCI is highlighted in bold.

As can be seen by the preponderance of numbers in bold in the ANN row, the ANN algorithm gives slightly better results in terms of mean CCI than other methods. In particular, it yields the best results for pianists Chiu (P1), Fliere (P4), Shebanova (P5), Kushner (P6), and Czerny (P8), followed by the k-NN algorithm that gives slightly better results for the pianists Smith (P2), Ashkenazy (P3), and Barbosa (P7). The highest average value of the percentage of correctly classified instances over all pianists is given by the ANN algorithm. Recall that in Section 7.3.2, ANN had the lowest prediction correlation to actual figures. That was for the case of loudness prediction; here, it performed best for dynamic-marking label prediction.

The CCI's reported in Table 4 are not high, given that a classifier that always predicts $\boldsymbol{p}$ would achieve 42%. This suggests that more data is needed for this task. One avenue for further analysis is to identify the markings that are more easily classified by considering the ones that have been predicted correctly by all recordings; this study is reported in Section 7.4.3. Another direction is to identify the Mazurka which has the highest number of markings that are more easily predicted, by observing for every Mazurka the ratio of the markings

correctly classified for all recordings to the total number of markings in the piece; this is reported in Section 7.4.4. We use the ANN algorithm as a basis for both studies due to its better performance found here.

7.4.3 Easily predicted markings

In this section, we seek to determine which of the dynamic-marking labels are more easily classified by the machine-learning algorithms when trained on other Mazurka recordings by the target pianist. A specific marking in a Mazurka is considered to be easily predicted if it has been classified correctly for all recordings. In Table 7.4 we present the markings across all Mazurkas that have been identified as being easily predicted according to this criterion.

Mazurka	Marking (position)	Mazurka	Marking (position)	Mazurka	Marking (position)
M06-3	***p*** (1), ***p*** (16)	M33-2	***ff*** (5)	M63-1	***p*** (3), ***p*** (5)
M07-3	***f*** (5)	M50-2	***p*** (2), ***p*** (4), ***p*** (6), ***p*** (7)	M67-1	***p*** (4), ***p*** (15)
M17-3	***p*** (5), ***p*** (9)	M50-3	***p*** (12)	M67-3	***p*** (11)
M17-4	***ff*** (5)	M56-1	***mf*** (12)	M68-2	***pp*** (2)
M24-4	***f*** (6), ***f*** (11)	M56-3	***p*** (12)	M68-3	***p*** (2)
M30-3	***pp*** (14)	M59-3	***p*** (11)		

Table 7.4: Markings that have been predicted correctly for all recordings of the Mazurka containing that marking; the numbers in parentheses indicate the position of that marking in the sequence of dynamic markings in that Mazurka.

Two patterns emerge from Table 7.4: the first has to do with the range of the dynamic markings in the Mazurka; the second has to do with the existence of important structural boundaries such as key modulations or cadences near the marking that impact the dynamics. We shall describe each case in greater detail providing specific examples of each case.

In the first case, when considering the range of markings present in a score, the marking at the edges of this range tend to correspond to extreme loudness values in the recording. Thus, these dynamics at the extremes would be the ones most easily categorized. For example, the ***ff*** marking in Mazurka Op. 17 No. 4 (M17-4) is a case in point. The markings that appear in this Mazurka are: {***pp***, ***p***, ***ff***}. Clearly, ***ff*** is at the extreme of this marking set, and in the loudness spectrum. Even in its position in the score, it is placed uniquely in such a way as to highlight the extreme nature of its dynamic level. Figure 7.11 shows the score position of the marking: it is preceded by a ***crescendo*** and followed by a ***p***. It is no wonder that this marking is correctly classified in all recordings of this Mazurka.

In the second case, structural boundaries are often inflected in performance

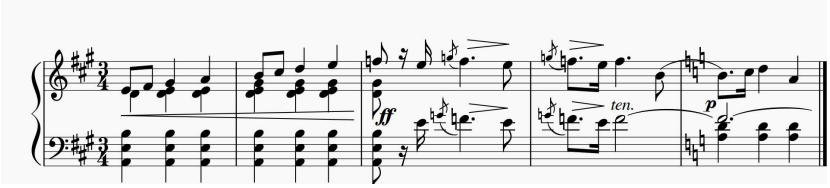


Figure 7.11: Case of the ***ff*** marking in Mazurka Op. 17 No. 4 (M17-4), correctly classified in all recordings of the Mazurka when the machine-learning algorithms are trained on the markings in the remaining Mazurkas by the target pianist.

in such a way as to reinforce their existence: the dynamic may drop immediately preceding or immediately after the boundary. As an example of a dynamic marking following a key change, consider the second ***f*** marking of Mazurka Op. 24 No. 4 (M24-4), which is in the key of B♭ minor. The score segment for this example is shown in Figure 7.12. Just before this easily predicted marking a phrase is repeated ending in the key of F major, and the repeat is marked ***sotto voce*** (in an undertone). The ***f*** appears at the return of B♭ minor; the performer is thus encouraged to make a sharp distinction between the ***f*** and the preceding ***sotto voce***. Four bars after the ***f*** is a ***pp*** marking, which would bring into sharper relief the ***f***. These factors all conspire to make this ***f*** more easily detectable.

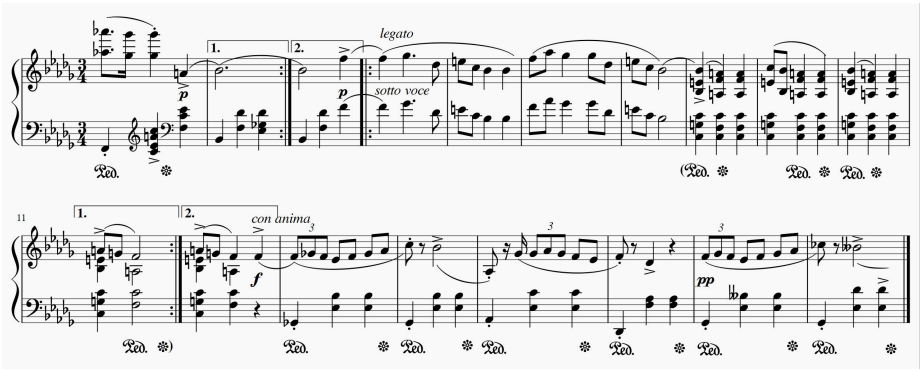


Figure 7.12: Case of the ***f*** marking in Mazurka Op. 24 No. 4 (M24-4), correctly classified in all recordings of the Mazurka when the machine-learning algorithms are trained on the markings in the remaining Mazurkas by the target pianist.

An example of an extreme dynamic marking near a structural boundary is the case of the marking ***pp*** in Mazurka Op. 30 No. 3 (M30-3), which is located at a cadence prior to a return to the main theme of the piece. The score segment for this example is shown in Figure 7.13. As can be seen in the score, the ***pp*** is proceeded by the text indicator ***dim.***, for diminuendo; furthermore, the text indicator ***slentando***, meaning to become slower, is paired with the marking; and, the marking is followed by a ***f*** marking paired with the text indicator

risoluto, meaning bold, at the return of the main theme. The extra meaning imputed to this ***pp*** as a signifier of the impending return of the ***f*** main theme makes it again a more extreme kind of ***pp***, and thus easier to classify.

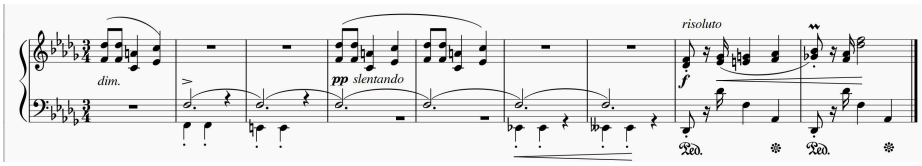


Figure 7.13: Case of the ***pp*** marking in Mazurka Op. 30 No. 3 (M30-3), correctly classified in all recordings of the Mazurka when the machine-learning algorithms are trained on the markings in the remaining Mazurkas by the target pianist.

In the next section, we consider the ease or difficulty of classifying dynamic markings through a different study, this time on the ratio of correctly classified markings.

7.4.4 Easy/hard to predict Mazurkas: ratio of correctly-classified markings

We have observed over the course of the experiments that there were Mazurkas for which the ratio of the markings that have been correctly classified is high for each recording, while for others that same ratio is low. To study this ratio across all Mazurkas, we have run a set of cross-validation experiments for which the markings of a specific Mazurka constituted the testing set and the markings of the remaining ones constituted the training set. The resulting ratio, averaged over all recordings of the Mazurka, of correctly classified markings to total markings is laid out in Figure 7.14 in a monotonically increasing order.

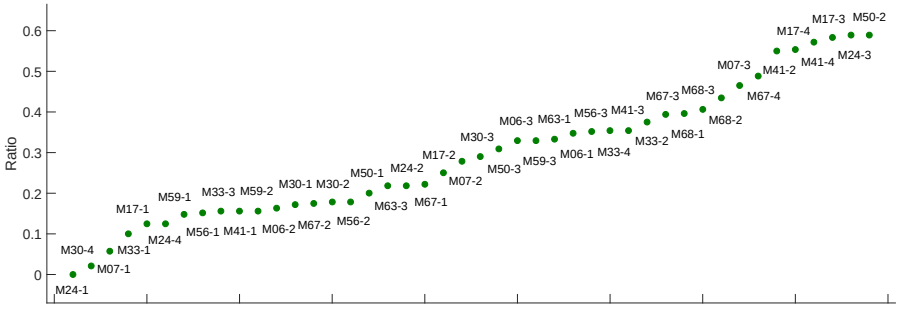


Figure 7.14: Average ratio, over all machine-learning methods, of correctly classified markings and number of markings per Mazurka over all pianists.

Note that Mazurka Op. 24 No. 1 has the lowest average ratio. Recall that Mazurka Op. 24 No. 1 was also the one in which the loudness value predictions

for one recording was almost perfectly and negatively correlated with the actual values; this was described in Section 7.3.1.

In contrast, Mazurka Op. 50 No. 2 has the highest ratio, 0.5893, meaning that this Mazurka has the highest number of correctly classified markings for every recording. We then consider in detail the loudness values at dynamic markings in Mazurka Op. 50 No. 2 for all eight recordings of that Mazurka. In Figure 7.15 the markings that are correctly classified for all recordings are highlighted with a solid vertical line, while the ones that are correctly classified for seven out of eight recordings are highlighted with a dashed vertical line. The loudness levels at these easily-classified markings follow the patterns established by other recordings.

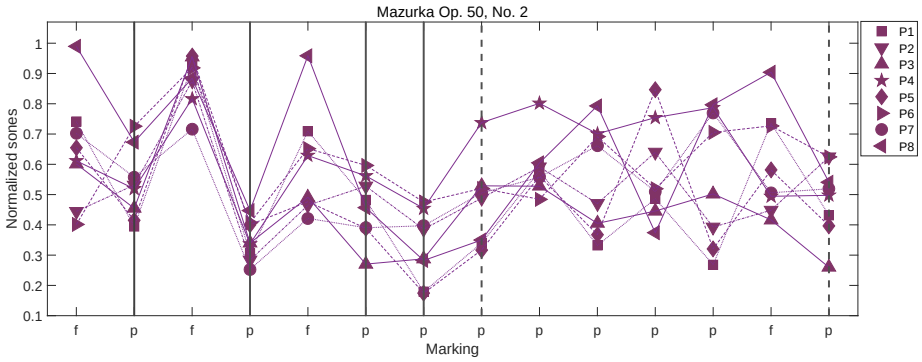


Figure 7.15: Loudness values at dynamic marking positions for Mazurka Op. 50 No. 2. Solid vertical lines indicate markings that are correctly classified for all eight recordings; dashed vertical lines indicate markings that are correctly classified for seven out of eight recordings.

Recall that consecutive strings of the same symbol posed significant challenges to the prediction of loudness values described in earlier sections. Note that Mazurka Op. 50 No. 2 again features a series of seven ***p***'s, with high degrees of variation in the loudness levels at these markings and in the ***f*** immediately following the sequence.

7.5 Conclusions

In this chapter we have investigated the relationship between loudness levels and dynamic markings in the score. To this end, we have implemented machine-learning approaches for the prediction of loudness values corresponding to different dynamic markings and musical contexts, as well as for the prediction of dynamic markings corresponding to different loudness levels and musical contexts. The methods—Decision Trees, Support Vector Machines, Artificial Neu-

ral Networks, and a k-Nearest Neighbor algorithm—are applied to forty-four recordings of performances of Chopins Mazurkas each by eight pianists.

The results in Section 7.3 highlight that, using any of the methods, loudness values and markings can be predicted fairly well when training across recordings of the same piece, but fail dismally when training across recordings of other pieces by the same pianist. This happens possibly because the score is a greater influence on the performance choices than the performers individual style for this test set. More specifically, insights from the data structure come forward, including Mazurkas in which loudness values can be predicted easier than others. This finding is related to the range of the different markings that appear in a piece as well as their position and function in the score in relation to structurally important elements.

The results in Section 7.4 highlight the good performance of the methods on predicting the dynamic markings given the loudness-marking data of the pianists' recordings of a target piece. When training across recordings of other pieces by the same pianist, the results, while not exceptional with respect to the prediction of dynamic markings, show notable trends on markings that are classified correctly, and on pieces that have higher ratios of classified markings over all markings. These trends are based mostly on the relationship between the position of the markings and the structure of the piece.

Different tuning of the parameters in the existing machine learning techniques for the prediction of loudness levels as well as for the prediction of dynamic markings may give better results. The scope of this study, however, is not to find the optimal solution to the tuning issue, but to point out universal characteristics of this kind of complex data. Improvements of the results may occur by considering possible alterations in the feature set or the data set. From the features set point of view, it would be especially interesting if the results were to improve with more score-informed features, and features that correspond to changes of other expressive parameters, such as tempo variations. From the data set point of view, the poor results of the 44-fold-validation experiments suggest that there may be insufficient data or features to train a model for such a complicated task. The results may be a consequence of the choice of training data, and perhaps training on a selection of pieces that include the desired or similar sequence of markings could enhance the results. Alternatively, it may be that musicians approach each new piece differently, and that there is little benefit from training a model on recordings of other pieces to predict performance features in a target piece.

Modeling expressive performance trends using artificial intelligence appears to be a rather delicate matter. The models are limited by the particular representations of changes that happen throughout a music piece with respect to

the expression rendering. The results that are drawn from the current study are likely to appear in similar studies for other instruments and data sets. Future work in this direction should include the comparison of performances from different eras, or performances of musicians from particular schools. At the same time it should be kept in mind that the freedom of employing a particular dynamic range in one interpretation is a matter of the performer's idiosyncrasy.

Chapter 8

Conclusions and further work

8.1 Conclusions

The output of the thesis serves as the first novel steps of transcribing the dynamic markings ***pp, p, mf, f, ff*** from music audio to score by detecting salient locations in loudness time series and labelling the corresponding position with one of the dynamic markings. Also the output serves for investigating salient performance parts, extracted patterns and similarities among different performances.

This research is a step forward for solving problems of assigning salience to musical features, the indicating of movements and arrivals, and the marking of boundaries and change. This direction includes the automatic extraction of signifiers having to do, apart from loudness, with tonality (Herremans and Chew [2016]), timing (Li et al. [2016]), as well as vibrato and portamento (Yang et al. [2015]).

The thesis explores two main research directions: a) From dynamic marking to performed loudness. Observations are derived from the comparison of the interpretation of same dynamic score markings, such as ***p*** (piano) and ***f*** (forte), from different interpretations of the same piano pieces. b) From performed loudness to score features. Implementation of Change-point statistical methods from time series analysis have been applied and change points serve as dynamic boundaries in the music audio.

We have created a list of factors that indicate salient characteristics for identifying the dynamic level of a dynamic marking at a certain score location. The factors serve as features for our bidirectional study for predicting loudness

levels and classifying dynamic markings. Also we have shown that significant dynamic score markings do indeed correspond to change points. Change points in score positions without dynamic markings also serve to make salient and focus the listener's attention on certain changes.

8.2 Future Directions

This work can be used as inspiration for other projects, such as a visualisation tool for reflecting change points derived from famous pianists playing on the corresponding score. Feature of the tool can involve the visualisation of change points from an individual recording on the top of others, for comparing performance styles. This type of tool can serve educational or musicological purposes.

Another project can be the systematic exploration of the mapping between timing and loudness changes. Our data has been created in such a way that a direct mapping to the corresponding score beat position is feasible. Finally, a data-driven machine learning approach for composing music can be benefited by incorporating the expressive features provided by our data as well as the outcome of our analysis. This approach can give more humanised characteristics to the output music.

Future steps include the expansion of the current list of factors, by incorporating new parameters related to additional score notation. One parameter could be the number of notes that are played simultaneously at a marking position as this may affect the final loudness result. Also the extraction of structural based parameters, such as the marking position related to the position where a phrase starts could be of further investigation as there are many ways of interpreting a music piece utilizing both compositional and expressive parameters, with respect to structural aspects.

Other future directions include the exploration of new ways in order to capture the loudness level that corresponds to one marking by expanding the window size of the three consecutive beats at a marking position or by using different loudness detection techniques. Also other change-point methods can be considered and compared with the ones used.

While this research focussed on step changes in dynamics, more gradual changes, such as the analysis of a ***crescendo*** or ***diminuendo*** could be a step-forward for future contributions. Also the first steps towards understanding, predicting and classifying tempo-related markings such as ***ritenuto*** or ***fermata***, as well as analysing their connection with dynamic markings have been done as an expansion of our work (Vaquero et al. [2017 (to appear]).

We can consider as future work the same analysis in music audio recordings

from a different composer or different period, as well as different instrumentation.

Appendix A

Glossary of music terms

This appendix includes a glossary of music terms used in this thesis. Big part of the definitions are taken from Rutherford-Johnson et al. [2012].

Agitato Agitated, restless.

Allegro ma non troppo Quick but not too quick.

Allegro A composition in allegro style or tempo.

Animato Animated, lively.

A tempo Return to the previous speed, after a slowing down or speeding up.

Con anima In a spirited manner.

Con forza With force.

calando Getting softer, dying away.

Dolce Soft, smooth.

Dolcissimo As soft and sweet as possible.

Espressivo Expressive.

Fermata A pause.

Gajo Joyful, merry.

Legato Bound together. Performance of music do that there is no perceptible pause between notes, i.e. in a smooth manner, the opposite of *staccato*. Indicated by slur or curved line. On string instruments, legato passages are played with one stroke of the bow; in vocal music, the legato passage is sung in one breath. *Legato touch* in piano playing requires holding down one key until the finger is on another.

Legato assai Superlative of *legato*.

Legatissimo Superlative of *legato*.

Leggiero Light, nimble.

Lento At a slow tempo.

Maestoso Majestic and stately.

Moderato Moderate speed.

Motif The shortest intelligible and self-existent melodic or rhythmic figure. Every 'theme' or 'subject' perhaps has several *motifs*, and almost every musical passage will be found to be a development of some *motif*.

Perdendosi Dying away.

Phrase Short section of a composition into which the music, whether vocal or instrumental, seems naturally to fall. Sometimes this is four measures, but shorter and longer phrases occur. It is an exact term: sometimes a phrase may be contained within one breath, and sometimes sub-divisions may be marked. In notation, phrase-marks are the slurs placed over or under the notes as a hint of their proper punctuation in a performance.

Portamento The currying of the sound from note to note smoothly and without any break, hence very *legato* and momentarily sounding the pitches in between any two indicated by the notation.

Rallentando Slowing down, gradually.

Risoluto Bold strong.

Ritenuto Held back, slower.

Rubato Robbed time. A feature in performance in which strict time is for a while disregarded—what is 'robbed' from some note or notes being "paid back' later. When this is done with genuine artistry and instinctive music sensibility, the effect is to impart an admirable sense of freedom and spontaneity. Done badly, rubato merely becomes mechanical. The question of rubato in Chopin is particularly contentious, since its use in his music may be dangerously open to abuse. Accounts of his playing (and of Mozart's) suggest that he kept the left-hand in strict time, and added rubato with the right.

Scherzando Playful, joking.

Sforzando Forced, accented.

Smorzando Extinguishing, gradually dying away.

Sotto voce Very softly, in an undertone.

Stretto Quicker tempo. In fugue: when entry of the answer occurs before subject is completed, overlapping with it. Often a way of increasing excitement.

Tempo I. A directive to perform the indicated passage of a composition in the original tempo of the composition, usually after a diversion from that original tempo.

Tenuto Held.

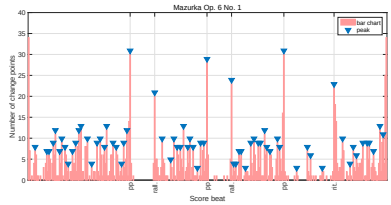
Vibrato A tremulous or pulsating effect produced in an instrumental or vocal tone by minute and rapid variations in pitch.

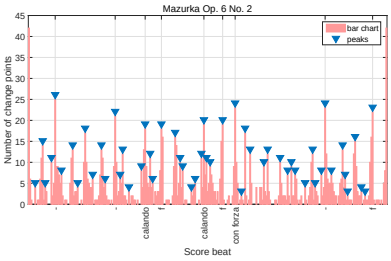
Vivo Lively.

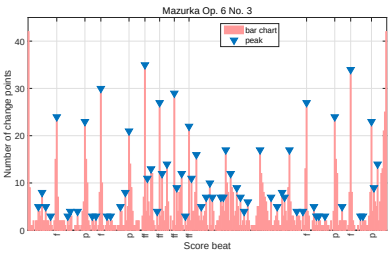
Appendix B

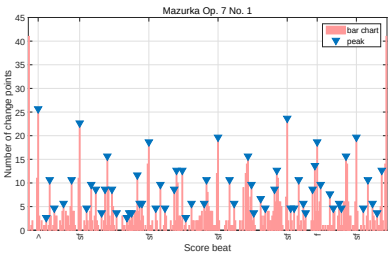
Change-points plots across the Mazurkas' recordings

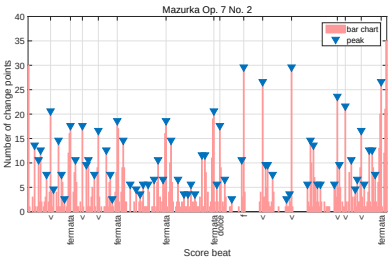
In this appendix we include the plots that are derived from the experiment described in Section 6.4 of Chapter 6. They show the locations of the change points across recordings as bar charts for each Mazurka separately. Also the local maxima have been marked.

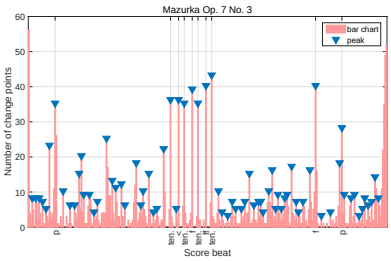

Mazurka Op. 6 No. 1
bar chart
peak
Number of change points
Score beat

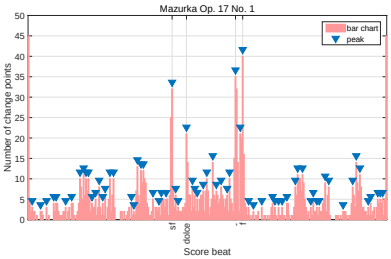

Mazurka Op. 6 No. 2
bar chart
peaks
Number of change points
calando
calando
con forza
Score beat

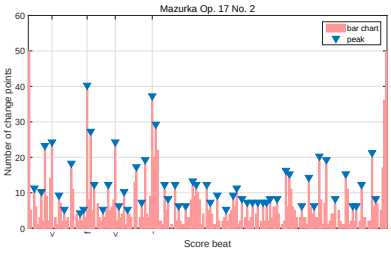

Mazurka Op. 6 No. 3
bar chart
peak
Number of change points
Score beat

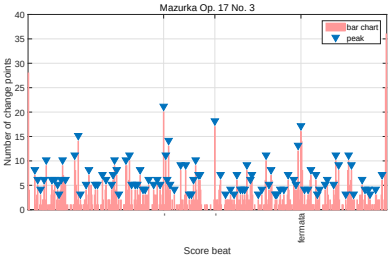

Mazurka Op. 7 No. 1
bar chart
peak
Number of change points
Score beat

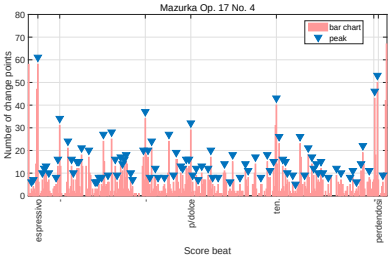

Mazurka Op. 7 No. 2
bar chart
peak
Number of change points
fermata
fermata
fermata
fermata
dolce
fermata
Score beat

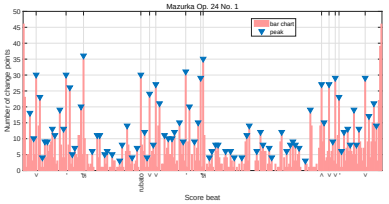

Mazurka Op. 7 No. 3
bar chart
peak
Number of change points
Score beat

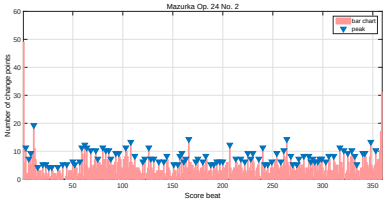

Mazurka Op. 17 No. 1
bar chart
peak
Number of change points
sf
dolce
Score beat

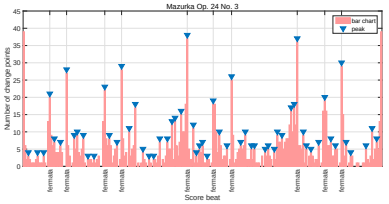

Mazurka Op. 17 No. 2
bar chart
peak
Number of change points
Score beat

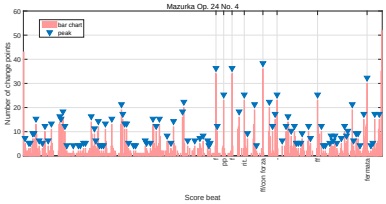

Mazurka Op. 17 No. 3
bar chart
peak
Number of change points
fermata
Score beat

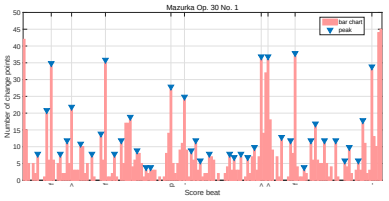

Mazurka Op. 17 No. 4
bar chart
peak
Number of change points
espressivo
più dolce
ten.
perdendosi
Score beat

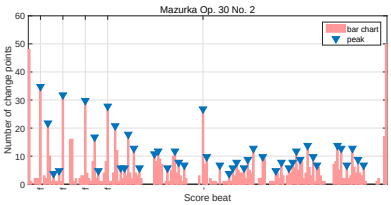

Mazurka Op. 24 No. 1
bar chart
peak
Number of change points
Score beat

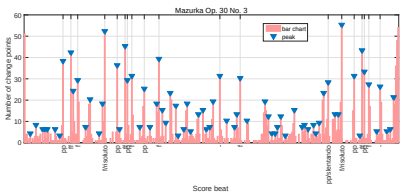

Mazurka Op. 24 No. 2
bar chart
peak
Number of change points
Score beat

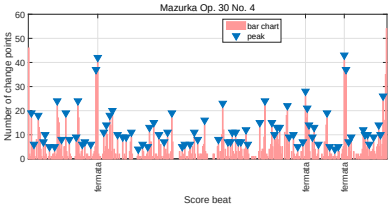

Mazurka Op. 24 No. 3
bar chart
peak
Number of change points
Score beat

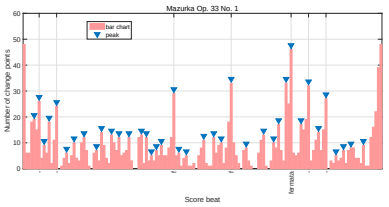

Mazurka Op. 24 No. 4
bar chart
peak
Number of change points
Score beat

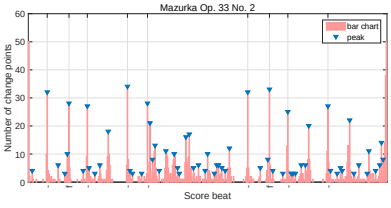

Mazurka Op. 30 No. 1
bar chart
peak
Number of change points
Score beat

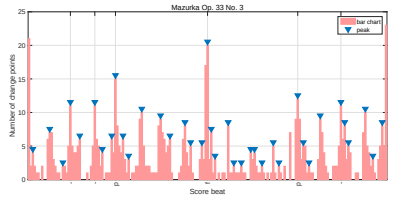

Mazurka Op. 30 No. 2
bar chart
peak
Number of change points
Score beat

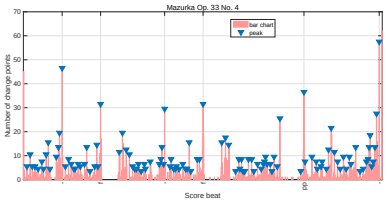

Mazurka Op. 30 No. 3
bar chart
peak
Number of change points
Score beat

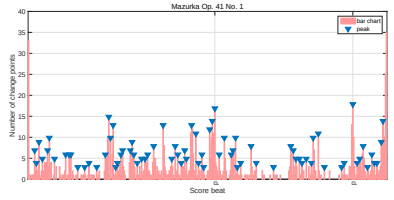

Mazurka Op. 30 No. 4
bar chart
peak
Number of change points
fermata
Score beat

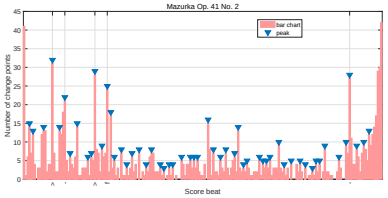

Mazurka Op. 33 No. 1
bar chart
peak
Number of change points
Score beat

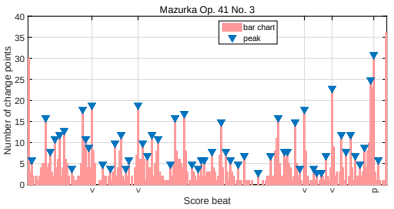

Mazurka Op. 33 No. 2
bar chart
peak
Number of change points
Score beat

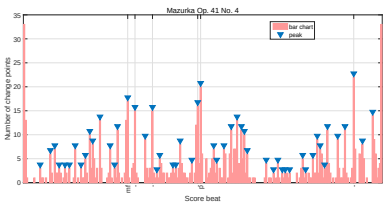

Mazurka Op. 33 No. 3
bar chart
peak
Number of change points
Score beat

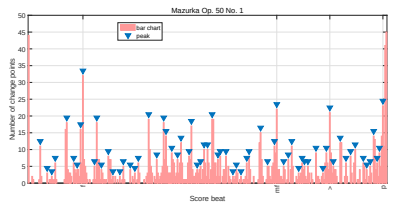

Mazurka Op. 33 No. 4
bar chart
peak
Number of change points
Score beat

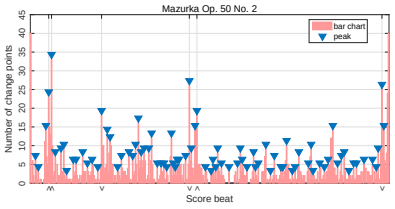

Mazurka Op. 41 No. 1
bar chart
peak
Number of change points
Score beat

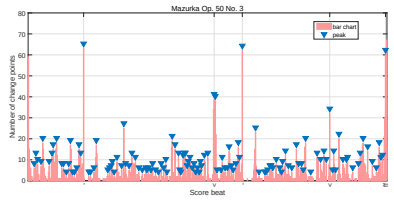

Mazurka Op. 41 No. 2
bar chart
peak
Number of change points
Score beat

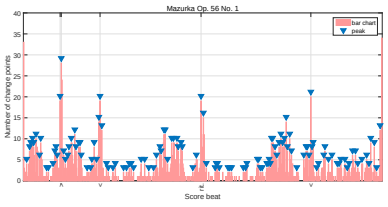

Mazurka Op. 41 No. 3
bar chart
peak
Number of change points
Score beat

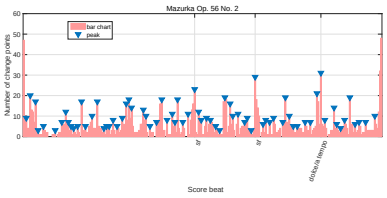

Mazurka Op. 41 No. 4
bar chart
peak
Number of change points
Score beat

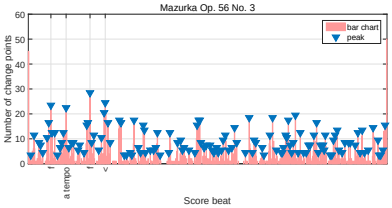

Mazurka Op. 50 No. 1
bar chart
peak
Number of change points
Score beat

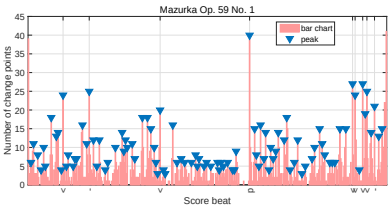

Mazurka Op. 50 No. 2
bar chart
peak
Number of change points
Score beat

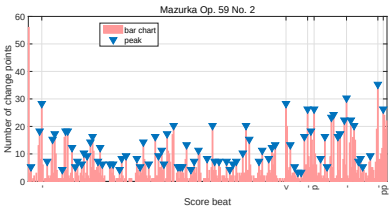

Mazurka Op. 50 No. 3
bar chart
peak
Number of change points
Score beat

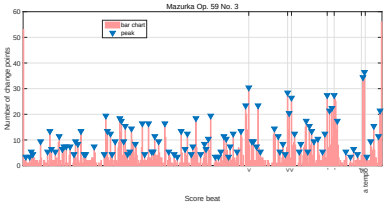

Mazurka Op. 56 No. 1
bar chart
peak
Number of change points
Score beat

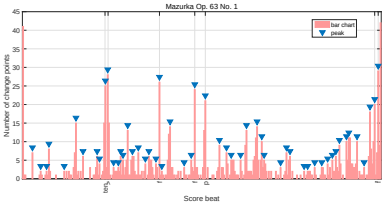

Mazurka Op. 56 No. 2
bar chart
peak
Number of change points
Score beat

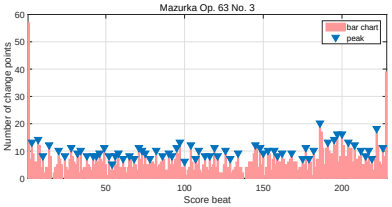

Mazurka Op. 56 No. 3
bar chart
peak
Number of change points
Score beat

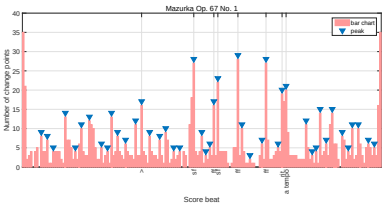

Mazurka Op. 59 No. 1
bar chart
peak
Number of change points
Score beat

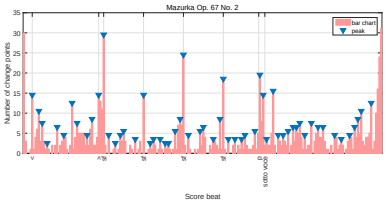

Mazurka Op. 59 No. 2
bar chart
peak
Number of change points
Score beat

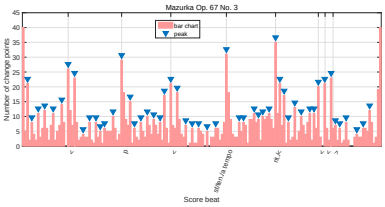

Mazurka Op. 59 No. 3
bar chart
peak
Number of change points
Score beat

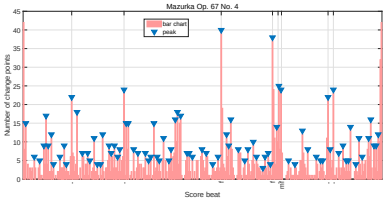

Mazurka Op. 63 No. 1
bar chart
peak
Number of change points
Score beat

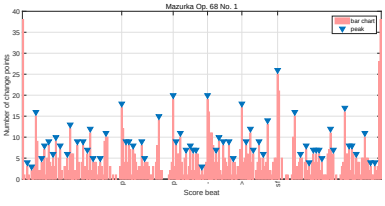

Mazurka Op. 63 No. 3
bar chart
peak
Number of change points
Score beat

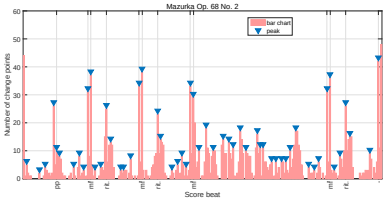

Mazurka Op. 67 No. 1
bar chart
peak
Number of change points
Score beat

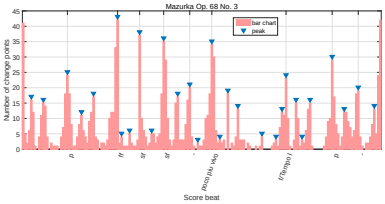

Mazurka Op. 67 No. 2
bar chart
peak
Number of change points
Score beat


Mazurka Op. 67 No. 3
bar chart
peak
Number of change points
Score beat


Mazurka Op. 67 No. 4
bar chart
peak
Number of change points
Score beat


Mazurka Op. 68 No. 1
bar chart
peak
Number of change points
Score beat


Mazurka Op. 68 No. 2
bar chart
peak
Number of change points
Score beat


Mazurka Op. 68 No. 3
bar chart
peak
Number of change points
Score beat

Bibliography

Eleanor Bailie. *Chopin: A Graded Practical Guide*. Pianist's repertoire. Kahn & Averill, 1998. ISBN 9781871082678. URL `https://books.google.co.uk/books?id=eoOBQgAACAAJ`.

Sam Barrett. Reflections on music writing : Coming to terms with gain and loss in early medieval latin song. In Andreas Haug and Andreas Dorschel, editors, *Vom Preis des Fortschritts: Gewinn Und Verlust in der Musikgeschichte*. Universal Edition, 2008.

Jacob Beck and William A. Shaw. Ratio-estimations of loudness intervals. *Journal of the acoustical society of America*, 80:59–65, 1967.

Alfonso Benetti Jr. Expressivity and musical performance: practice strategies for pianists. In *Performance Studies Network International Conference, Cambridge*, 2013. URL `http://www.cmpcp.ac.uk/wp-content/uploads/2015/11/PSN2013_Benetti.pdf`.

Gerald Bennett. The sense of the phrase. compositional grouping in music. In Johan Sundberg, editor, *Gluing tones: grouping in music composition, performance and listening*. Royal Swedish Academy, 1993.

Stefan Benus, Agustín Gravano, and Julia Hirschberg. Prosody, emotions, and...'whatever'. In *Interspeech, Antwerp*, 2007. URL `http://www1.cs.columbia.edu/nlp/papers/2007/benus_al_07a.pdf`.

Anjali Bhatara, Anna K Tirovolas, Lilu Marie Duan, Bianca Levy, and Daniel J Levitin. Perception of emotional expression in musical performance. *Journal of Experimental Psychology: Human Perception and Performance*, 37(3):921, 2011.

Erica Bisesi and Richard Parncutt. Second vienna talk. In *Proceedings of the Second Vienna Talk, University of Performing Arts Vienna*, pages 26–30, 2010.

Christopher M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

R. A. W. Bladon and Björn Lindblom. Modeling the judgment of vowel quality differences. *The Journal of the Acoustical Society of America*, 69(5):1414–1422, 1981.

Yves Chauvin and David E. Rumelhart. *Backpropagation: theory, architectures, and applications*. Psychology Press, 1995.

Eric Cheng and Elaine Chew. Quantitative analysis of phrasing strategies in expressive performance: Computational methods and analysis of performances of unaccompanied bach for solo violin. *Journal of New Music Research*, 37 (4):325–338, 2008. URL `http://dx.doi.org/10.1080/09298210802711660`.

Elaine Chew. The tipping point analogy for musical timing. In *Performance Studies Network International Conference, Cambridge*, 2013.

Nicholas Cook. *Music: A Very Short Introduction*. Very Short Introductions. OUP Oxford, 2000. ISBN 9780192853820. URL `https://books.google.com.au/books?id=Dl2gvBIKvpEC`.

Nicholas Cook, Eric Clarke, Leech-Wilkinson Daniel, and John Rink. *The Cambridge companion to recorded music*. Cambridge University Press, 2009.

Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines: And Other Kernel-based Learning Methods*. Cambridge University Press, New York, NY, USA, 2000. ISBN 0-521-78019-5.

Michael Scott Cuthbert and Christopher Ariza. music21: A toolkit for computer-aided musicology and symbolic music data. In *Proceedings of the 9th International Conference on Music Information Retrieval*, pages 637–642, 2010.

Simon Dixon, Werner Goebl, and Gerhard Widmer. The performance worm: Real time visualisation of expression based on langner's tempo-loudness animation. In *International computer music conference (ICMC)*, 2002.

Jean-Jacques Eigeldinger. Placing chopin: reflections on a compositional aesthetic. In John Rink and Jim Samson, editors, *Chopin Studies 2*, volume 2. Cambridge University Press, 1994.

Alfred Einstein. *Music in the romantic era: a history of musical thought in the 19th century*. New York : W.W. Norton, 1975.

Sebastian Ewert, Meinard Müller, and Peter Grosche. High resolution audio synchronization using chroma onset features. In *thirty-fourth IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1869–1872, Taipei, Taiwan, 2009.

Dorottya Fabian. Commercial sound recordings and trends in expressive music performance: why should experimental researchers pay attention? In Dorottya Fabian, Renee Timmers, and Emery Schubert, editors, *Expressiveness in music performance: Empirical approaches across styles and cultures*. Oxford University Press, 2014.

Dorottya Fabian, Renee Timmers, and Emery Schubert. *Expressiveness in music performance: Empirical approaches across styles and cultures*. Oxford University Press, 2014.

Harvey Fletcher and Wilden A. Munson. Loudness. *Journal of the acoustical society of America*, 5:82–108, 1933.

Sebastian Flossmann, Werner Goebl, Maarten Grachten, Bernhard Niedermayer, and Gerhard Widmer. The magaloff project: An interim report. *Journal of New Music Research*, 39(4):363–377, 2010.

John M Geringer. Loudness estimations of noise, synthesizer, and music excerpts by musicians and nonmusicians. *Psychomusicology: A Journal of Research in Music Cognition*, 12(1):22, 1993.

John M Geringer. Continuous loudness judgments of dynamics in recorded music excerpts. *Journal of Research in Music Education*, 43(1):22–35, 1995.

Werner Goebl, Simon Dixon, and Emery Schubert. Quantitative methods: motion analysis, audio analysis, and continuous response techniques. In Dorottya Fabian, Renee Timmers, and Emery Schubert, editors, *Expressiveness in music performance: Empirical approaches across styles and cultures*. Oxford University Press, 2014.

Maarten Grachten and Florian Krebs. An assessment of learned score features for modeling expressive dynamics in music. *IEEE Transactions on Multimedia*, 16(5):1211–1218, Aug 2014. ISSN 1520-9210.

Maarten Grachten and Gerhard Widmer. Linear basis models for prediction and analysis of musical expression. *Journal of New Music Research*, 41(4): 311–322, 2012. URL `http://dx.doi.org/10.1080/09298215.2012.731071`.

Agustín Gravano, Stefan Benus, Ramiro H. Chávez, Julia Hirschberg, and Lauren Wilcox. On the role of context and prosody in the interpretation of 'okay'. In *Proceedings of ACL*, pages 800–807, 2007.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November 2009. ISSN 1931-0145.

W.M. Hartmann. *Signals, Sound, and Sensation*. Modern Acoustics and Signal Processing. American Inst. of Physics, 2004. ISBN 9781563962837. URL `https://books.google.co.uk/books?id=3N72rIoTHiEC`.

Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media, 2001.

Dorien Herremans and Elaine Chew. Tension ribbons: Quantifying and visualising tonal tension. *Second International Conference on Technologies for Music Notation and Representation (TENOR)*, 2:8–18, 05/2016 2016.

Ching hua Chuan and Elaine Chew. A dynamic programming approach to the extraction of phrase boundaries from tempo variations in expressive performances. In *In Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007*, pages 305–308, 2007.

Hui Chi Khoo. *Playing with Dynamics in the music of Chopin, Ph.D. thesis*. Royal Holloway, University of London, 2007.

Rebecca Killick and Idris A. Eckley. changepoint: An R package for changepoint analysis. *Journal of Statistical Software*, 58(3):1–19, 2014. URL `http://www.jstatsoft.org/v58/i03/`.

Rebecca Killick, Idris A Eckley, Kevin Ewans, and Philip Jonathan. Detection of changes in variance of oceanographic time-series using changepoint analysis. *Ocean Engineering*, 37(13):1120–1126, 2010.

Rebecca Killick, Paul Fearnhead, and IA Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012.

Katerina Kosta, Oscar F. Bandtlow, and Elaine Chew. Practical implications of dynamic markings in the score: Is piano always piano? In *Audio Engineering Society (AES) 53rd International Conference on Semantic Audio*, London, UK, 2014.

Katerina Kosta, Rafael Ramrez, Oscar F. Bandtlow, and Elaine Chew. Mapping between dynamic markings and performed loudness: a machine learning approach. *Journal of Mathematics and Music*, 10(2):149–172, 2016. URL `http://dx.doi.org/10.1080/17459737.2016.1193237`.

Jörg Langner and Werner Goebl. Visualizing expressive performance in tempo-loudness space. *Computer Music Journal*, 27(4):69–83, December 2003. ISSN 0148-9267. URL http://dx.doi.org/10.1162/014892603322730514.

Edward W. Large and Caroline Palmer. Perceiving temporal regularity in music. *Cognitive Science*, 26(1):1–37, 2002. ISSN 1551-6709. URL http://dx.doi.org/10.1207/s15516709cog2601_1.

Marc Lavielle. Using penalized contrasts for the change-point problem. *Signal processing*, 85(8):1501–1510, 2005.

Daniel Leech-Wilkinson. Compositions, scores, performances, meanings. *Music Theory Online*, 18(1):1–17, 2012. URL http://mtosmt.org/issues/mto.12.18.1/mto.12.18.1.leech-wilkinson.php.

Daniel Leech-Wilkinson. Cortot's berceuse. *Music analysis*, 34(3):335–363, 2015.

Shengchen Li, Simon Dixon, Dawn Black, and Mark Plumbley. A model selection test on effective factors of the choice of expressive timing clusters for a phrase. *13th Sound and Music Computing Conference (SMC 2016)*, 2016.

Lie Lu, Dan Liu, and Hong-Jiang Zhang. Automatic mood detection and tracking of music audio signals. *IEEE Transactions on audio, speech, and language processing*, 14(1):5–18, 2006.

James Methuen-Campbell. Chopin in performance. In Jim Samson, editor, *The Cambridge Companion to Chopin*. Cambridge University Press, 1994.

Brian C. J. Moore, Brian R. Glasberg, and Thomas Baer. A model for the prediction of thresholds, loudness, and partial loudness. *J. Audio Eng. Soc*, 45(4):224–240, 1997. URL http://www.aes.org/e-lib/browse.cfm?elib=10272.

Brian CJ Moore. Why are commercials so loud? *Noise & Vibration Worldwide*, 36(8):11–15, 2005.

Soren H Nielsen and Esben Skovenborg. Evaluation of different loudness models with music and speech material. In *Audio Engineering Society Convention 117*. Audio Engineering Society, 2004.

Regina Nuzzo. Scientific method: Statistical errors. *Nature*, 506(7487):150–152, 2014. URL http://www.nature.com/news/scientific-method-statistical-errors-1.14700.

Caroline Palmer and Sean Hutchins. What is musical prosody? volume 46 of *Psychology of Learning and Motivation*, pages 245–278. Academic Press, 2006. URL http://www.sciencedirect.com/science/article/pii/S0079742106460072.

E. Pampalk, A. Rauber, and D. Merkl. Content-based Organization and Visualization of Music Archives. In *Proceedings of the ACM Multimedia*, pages 570–579, Juan les Pins, France, December 1-6 2002. ACM.

John Ross Quinlan. Induction of decision trees. *Machine learning*, 1:81–106, 1986.

R.E. Radocy and J.D. Boyle. *Psychological foundations of musical behavior*. Charles C. Thomas, 2003. ISBN 9780398073848. URL https://books.google.co.uk/books?id=tioJAQAAMAAJ.

Bruno H Repp. The dynamics of expressive piano performance: Schumann's träumerei revisited. *The Journal of the Acoustical Society of America*, 100 (1):641–650, 1996.

John Rink. Authentic chopin: history, analysis and intuition in performance. In John Rink and Jim Samson, editors, *Chopin Studies 2*, volume 2. Cambridge University Press, 1994.

John Rink. *Musical Performance: A Guide to Understanding*. Cambridge University Press, 2002. ISBN 9780521788625. URL https://books.google.co.uk/books?id=xaYxRe-5ztIC.

Marko Robnik-Sikonja and Igor Kononenko. An adaptation of relief for attribute estimation in regression. In *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, pages 296–304, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc. ISBN 1-55860-486-3.

María Ros, Miguel Molina-Solana, Miguel Delgado, Waldo Fajardo Contreras, and M. Amparo Vila. Transcribing Debussy's syrinx dynamics through linguistic description: The MUDELD algorithm. *Fuzzy Sets and Systems*, 285: 199–216, 2016. URL http://dx.doi.org/10.1016/j.fss.2015.08.004.

T. Rutherford-Johnson, M. Kennedy, and J. Kennedy. *The Oxford Dictionary of Music*. OUP Oxford, 2012. ISBN 9780199578108. URL https://books.google.co.uk/books?id=ZeOuuAAACAAJ.

J.A. Sadie. *Companion to Baroque Music*. University of California Press, 1998. ISBN 9780520214149. URL https://books.google.co.uk/books?id=Ip6voIceW0AC.

Craig Sapp. *Computational Methods for the Analysis of Musical Structure, Ph.D. thesis*. Stanford University, 2011.

Manfred R Schroeder, Bishnu S Atal, and JL Hall. Optimizing digital speech coders by exploiting masking properties of the human ear. *The Journal of the Acoustical Society of America*, 66(6):1647–1652, 1979.

Floris Schuiling. Composition, improvisation and practical creativity in the performance practice of the instant composers pool. In *Performance Studies Network International Conference, Cambridge*, 2013.

Andrew John Scott and M Knott. A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, pages 507–512, 1974.

L. Henry Shaffer. Performing the F# minor prelude op. 28 no. 8. In John Rink and Jim Samson, editors, *Chopin Studies 2*, volume 2. Cambridge University Press, 1994.

Stanley Smith Stevens. The volume and intensity of tones. *American journal of psychology*, 46:397–408, 1934.

Stanley Smith Stevens and H. S. Terrace. The quantification of tonal volume. *American journal of psychology*, 75:596–604, 1962.

Dan Stowell and Elaine Chew. Maximum a posteriori estimation of piecewise arcs in tempo time-series. In *International Symposium on Computer Music Modeling and Retrieval*, pages 387–399. Springer, 2012.

Adrian Thomas. Beyond the dance. In Jim Samson, editor, *The Cambridge Companion to Chopin*. Cambridge University Press, 1994.

Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

Carlos Vaquero, Ivan Titov, and Henkjan Honing. What score markings can say of the synergy between expressive timing and loudness. In *in proceedings of the European Society for Cognitive Sciences Of Music*, 2017 (to appear).

Anja Volk. Persistence and change: local and global components of metre induction using inner metric analysis. *Journal of Mathematics and Music*, 2 (2):99–115, 2008.

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

Luwei Yang, Elaine Chew, and Khalid Z Rajab. Logistic modeling of note transitions. In *International Conference on Mathematics and Computation in Music*, pages 161–172. Springer, 2015.