Dynamics and Relativity: Practical Implications of Dynamic Markings in the Score

Katerina Kosta[1], Oscar F. Bandtlow[2], Elaine Chew[1]
1. Centre for Digital Music, School of Electronic Engineering and Computer Science, Queen Mary University of London
2. School of Mathematical Sciences, Queen Mary University of London

ABSTRACT

This article 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 (actually 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 twice, there tends to be significant difference in the ways they are interpreted in a recording. Finally we present evidence that pianists with high numbers of outlier recordings, i.e. showing uniqueness in the ways they transition between dynamic markings, tend to co-occur in outlier clusters for a significant number of Mazurkas, meaning they diverge from common practice in similar ways. Applications of these results include expression synthesis and music transcription.

1 Introduction

Musical expression is an important part of music performance analysis. Palmer, in [29], underscores 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 for the purposes of communicating emotion and clarifying structure.

Much research has focused on analysing and modelling the correlation between these acoustic properties, examples include studies on the relationship between tempo and loudness (e.g. [37] and references therein) or pitch and loudness (e.g. [14] and reference therein). However, less attention has been paid to the relation between such properties and their representation, as a means of communication between the composer and the performer. Our study focuses on musical expressivity and aims to understand the connection between loudness levels in recordings and dynamic markings as notated in the 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 [31, p.59], who can choose to render the symbols in unique ways. The audio effect of the refracted work is received by the listener, who in

turn may perceive it through their own individual lenses. This paper focuses on two of these perspectives, that of the composer (as communicated by a specific score edition) and that of the professional performer (as archived in a recording).

Cook in [7, p.14] highlights that in books on classical music there is a distinction between *authors*, i.e. composers and *reproducers*, i.e. performers; The latter ones used to be mentioned when they obscured the original music through "over-interpretation or gratuitous virtuosity." In the case of Chopin, being both composer and performer, his personal performing style was imitated internationally, although it was not his intention to emphasize his originality [10, p. 215]. The above has to be taken into account when we analyse different interpretations of his pieces. As many of the performers might imitate his style, we should acknowledge that in terms of dynamics "all his contemporaries agree in reporting that his dynamics did not exceed the degree of *forte*, without however losing a single bit of shading" [10, p. 215].

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,

$$\boldsymbol{pp} < \boldsymbol{p} < \boldsymbol{mp} < \boldsymbol{mf} < \boldsymbol{f} < \boldsymbol{ff}, \tag{1}$$

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.

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. The goal of our study is to better understand the parameters which define the performers' responses to dynamic markings. Spurred on by our first study in [23], this article presents an expanded and deeper analysis of the nature of the relativity in expression of dynamic levels.

In this study 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 piece, 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 piece, 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

[1]For example, studies have examined the meanings of nuances in utterances of the words "whatever" [4] and "okay" [20].

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 ***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.) For the second type of relativity, for each recording we compare the loudness change between the different consecutive markings in pairs ($\boldsymbol{p}_2$, $\boldsymbol{f}_1$), ($\boldsymbol{f}_1$, ***ff***), (***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 dynamic change vary related to the ordinal distance between dynamic markings?

(iv) How are the same dynamic markings realised as they recur throughout the piece?

Along the direction of analysing general dynamic change behaviour throughout a piece, we also conduct an experiment on analysing specific patterns that emerge on the way different pianists manipulate dynamic changes over time based on the positions of the dynamic markings, with a focus on exceptional interpretations.

A 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 study, we have drawn from the CHARM[2] Chopin Mazurka dataset, 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 [3]. Hence, our choice to focus on Chopin's Mazurkas.

In order to address the four questions raised above, 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 [12, 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.

[2]http://www.charm.kcl.ac.uk, accessed January 2018

This research will enable the building of 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 results can be a useful tool for generating expressive renderings of notated scores. However, one must take into account two main considerations concerning the score markings: firstly, the dynamic markings may simply reinforce the natural predisposition of a music part to be expressed in a particular way based on its structure; secondly, precise annotation by the composer can be present at a non-obvious place [18]. Together with the large amount of information and performance related parameters [5], the challenges above make the modelling of music expression one of today's most important unsolved problems in the description of music audio.

The article is organized as follows: Section 2 describes related work; Section 3 describes the creation of the dataset for the studies of this article; Section 4 presents the results of the studies; Section 5 presents the analysis of clustering results between recordings; followed by conclusions and discussions in Sectiond 6 and 7, respectively.

2 Related research

Extremely 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 Chopin Mazurkas [23].

[15] presents an early study on dynamic contrasts in relation to score dynamic markings. This study analysed data sampled from 60 commercial recordings of choral, orchestral, and piano compositions mostly from the 19th century and from composers including Beethoven, Chopin, and Schumann, amongst others. The dynamic changes within the music samples were recorded using a Bruel and Kjaer Graphic Level Recorder, which enables the continuous measurement of relative intensity changes. Across the music excerpts sampled, the results showed a larger dynamic range from $\boldsymbol{p}$ to $\boldsymbol{f}$, with an average change of 13.42dB, than from $\boldsymbol{f}$ to $\boldsymbol{p}$, with an average change of 11.97dB.

[32] proposed the MUDELD (MUsic Dynamics Extraction through Linguistic Description) algorithm, which uses linguistic description techniques to categorise dynamic labels of separate musical phrases into three levels designated as *piano* ($\boldsymbol{p}$), *mezzo* ($\boldsymbol{m}$), and *forte* ($\boldsymbol{f}$). The basic idea of this study was to extract loudness information from audio input by computing the Root Mean Square of the signal followed by normalisation so that the resulting values lie in $[0, 1]$. The music signal was then manually segmented, each segment was then represented by its average loudness value. Next, "disagreement values" were computed for the distance between the segment's value label and the reference label obtained from the score. As an initial experiment, eight commercial recordings of Debussy's "Syrinx" were retrieved and manually segmented into phrases. The outcome of the experiment indicates that the resulting labels were similar to the score indications.

In [18], a linear basis framework (LBM) is proposed to account for expressive variations as well as the effect of expressive markings in the score on music performances. The approach relies on the creation of a number of hand crafted numerical descriptors, or "basis functions", encoding certain structural aspects of the score. Each of the basis functions is linked to one

score marking and represents the activation of that marking on a scale of 0 (non-active) to 1 (active). The basis functions serve as indicators for note attributes, such as *stacatto* or expressive markings, for gradual (e.g. *crescendo*) or instant (e.g. *piano*) changes. Essentially, LBM provides a way to determine the optimal influence of each of the sets of basis functions with a set of weights that approximate the impact of an expressive parameter.

To test the LBM approach, two experiments were performed; the first examined how accurately expressive dynamics can be represented and predicted using as dataset the Magaloff corpus [13], recordings of Chopin's pieces performed on a Bösendorfer recording piano. The results showed that, for both representation and prediction, the correlation ranges from weak to medium for various combinations of basis functions, as well as for weights computed globally (same in all pieces) or locally (different for each piece). The latter produced better results. Also, prediction variance was substantially lower than the variance observed. The second experiment examined how well the LBM framework reveals differences between performers 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 make direct links between the coefficients and individual performances, independent of the piece. This finding resonates with our results in [24].

The basis framework has been used in the case of recorded orchestral music in [19] where the measure of dynamics is the overall variation of loudness over time. Also, a study of a machine learning approach concerned with the score-based prediction of separate note intensities in performed music was presented in [17].

Our approach differs from prior work in its focus on loudness representation (notated loudness) and the degree to which the representations capture constant and relative dynamic changes as realised in performance.

3 Data Preparation

In this section we present the data that has been created and used for the purpose of this study. More specifically, Section 3.1 describes the dataset, and Section 3.2 describes the loudness parameter that forms the basis of our analysis.

3.1 MazurkaBL Dataset

For this and related studies, we have created the MazurkaBL dataset [21], a collection of score-beat positions and loudness values, with corresponding score dynamic and tempo markings for 2000 recordings of forty-four Chopin Mazurkas. The dataset is publicly available and it can be found in [27]. The audio files used for the creation of this dataset are derived from the Mazurka Dataset[3] which offers a significant number of different interpretations of the same Mazurkas by Chopin as performed by different professional pianists. MazurkaBL focuses on a part of the dataset containing 2000 audio recordings of performances of forty-four pieces. The final number of recordings selected for each Mazurka is shown in Table 1. The audio recordings included in MazurkaBL are the ones that adhere to the repetition instructions in the score; excessively noisy recordings have been excluded.

[3] http://www.mazurka.org.uk, accessed November 2016

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
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
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
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

Table 1: Chopin Mazurkas used in this study, and the number of recordings, and the number of dynamic markings that appear in each Mazurka; the Mazurkas are indexed as "M<*opus*>-<*number*>."

For our analysis, we focus on the following markings:

$$S = \{\boldsymbol{pp},\ \boldsymbol{p},\ \boldsymbol{mf},\ \boldsymbol{f},\ \boldsymbol{ff}\}. \tag{2}$$

pp occurs 63 times, p 234, ***mf*** 21, ***f*** 169, and ***ff*** 43 times in the dataset, giving a total of 530 markings. The number of markings that appear in each piece individually is shown in Table 1.

In the next subsections we explain how we link the beat positions in the score with corresponding ones in each recording, and we describe how we compute the loudness information for each dynamic marking. For simplicity, we focus on the loudness values that correspond to the marking positions, while acknowledging the importance of the dynamic changes in between. The way the data is structured allows us to create a score-based information model.

The recording dates ranging from 1902 to the early 2000's. Furthermore, the score edition used by each performer is not known. As mentioned in [1, 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." Thus, the many pianists could be playing from rather different editions. However, tracing the actual score used in the preparation of each performance is unrealistic. For the purposes of obtaining score-based dynamic markings for this study, we used the edition by Paderewski, Bronarski and Turczynski, as it is one of the most popular and readily available editions. A general overview of differences among score editions on dynamic markings in S reveals the following: a popular case is that a marking might be located in a different position by one or two–beat distance, a non-popular case where a marking is either missed or is additional, a non-popular case where a marking is presented inside a parenthesis while it does not appear in other editions, and a rare case of a marking expressed as a different marking elsewhere.

We created an XML version of the scores and information concerning the exact location of each dynamic marking was extracted automatically using Music21[4] [9].

[4]web.mit.edu/music21, accessed January 2017

3.2 Loudness information for markings

We have already mentioned in the introduction that we focus on audio features that contain information about loudness. More specifically we would like to capture the change in dynamics through a piece of music.

Laboratory attempts to measure loudness in terms of how much louder one sound is than another have used estimations of various loudness distances or ratios. A sone scale of loudness, based on estimates of apparent ratios provides a way to describe apparent loudness relationships [2]. In our research we have used sone measurements to quantify the loudness of audio signals. An advantage of the sone scale is that it is linear: doubling the sone values corresponds to doubling the loudness sensation of a corresponding tone, allowing for more accurate normalisation. Another equally important reason is that we wish to train our models on audio that is pre-processed based on the basic psychoacoustic concept of equal loudness curves, and not to make any other compression or modification. If a *machine* were to *listen* like a human, then it needs an input similar to that of a listener.

For the experiments presented in this article, loudness information is extracted from the audio signal using the `ma_sone` function in Elias Pampalk's Music Analysis toolbox[5]. The specific loudness sensation in sones per critical band is calculated by following the process explained in [30]. Using this procedure, we calculate the power spectrum of the audio signal using a Fast Fourier Transform. We then use a window size of 256 samples, a 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." [30] We calculate the Spectral masking effects using the method described in [33]. Then, we calculate the loudness as dB-SPL units; from these values we calculate the equal loudness levels in Phon, and convert the values to sones following the formula described in [6]:

$$S = \begin{cases} 2^{(L-40)/10}, & P \geq 40 \\ (L/40)^{2.642}, & P < 40, \end{cases} \tag{3}$$

where L is the loudness level in phons, S is the loudness level in sones, P is the absolute dB value of the envelope expressed in phons via stored curves of equal loudness levels. This calculation accounts for the threshold of hearing and the ear's nonlinear and frequency-dependent response to intensity differences. [6]

This pre-processing stage transforms each recording into a dynamic time series (Fig. 1-top, blue curve). The sone values are smoothed by local regression using weighted linear least squares and a 2^{nd} degree polynomial model (the "loess" method of MATLAB's *smooth* function[6]) (Fig. 1—top, brown curve). From the smoothed data, we consider the values of the score-beat positions (yellow x's in Fig. 1—top), which constitutes a sequence $x_n, n \in \mathbb{N}$, where n is the number of score beats in one piece. The sequence is then normalised to $[0, 1]$ by dividing the values by the maximum loudness value of the particular recording. In this way we are able to compare different recording environments which serves our purpose of focusing on relative rather than absolute changes.

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

[6] `uk.mathworks.com/help/curvefit/smooth.html?refresh=true`, accessed 3 January 2018.

For each Mazurka, the first author manually annotated the beats for one recording, and beat positions were transferred automatically to the remaining recordings using a multiple performance alignment heuristic. The alignment uses the algorithm described in [11], which is based on Dynamic Time Warping using chroma features; the heuristic optimises the choice of reference recording. A characteristic of Chopin's music is that it draws inspiration from singing, which translates to a style of piano playing whereby the melody may be displaced from the corresponding beat position in the accompaniment so as to convey fluidity of expressive timing. Related to this, see [16] on the melody lead effect. 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 markings that are analysed are marked in Fig. 1—bottom as black x's. The loudness value corresponding to each marking is the average of that found at the beat of the marking and the two consecutive beats, exceptions being the markings where in the following two beats there was no event; in such cases we considered only the loundess value that correspond to the current beat. More formally, if $\{y_n\} \in \mathbb{R}$ is the sequence of the normalised 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

$$\ell_b = \frac{1}{3} \sum_{i=1}^{3} y_{b+(i-1)}. \tag{4}$$

The reason behind choosing three beats is based on the observation that sometimes the actual change in response to a new dynamic marking does not take place immediately, and can only be observed in the subsequent beat or two. It is clear from the data that loudness varies considerably between one dynamic marking and the next. Thus, we additionally aim to have the smallest window possible to capture dynamic changes in response to a marking. Consequently, we have chosen a three-beat window as the window for study, which for Mazurkas corresponds to a bar of music.

4 Performed Loudness Study

4.1 Ordinal Loudness Sequence Preserved On Average?

In this section we explore the general behaviour of pianists' responses to the dynamic markings in 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 do 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. As some markings appear in the score more than once, we compute the aggregate value for each marking in a recording by taking the average of the loudness values whenever the same marking appears in the score: for a symbol $s \in S$ that appears n

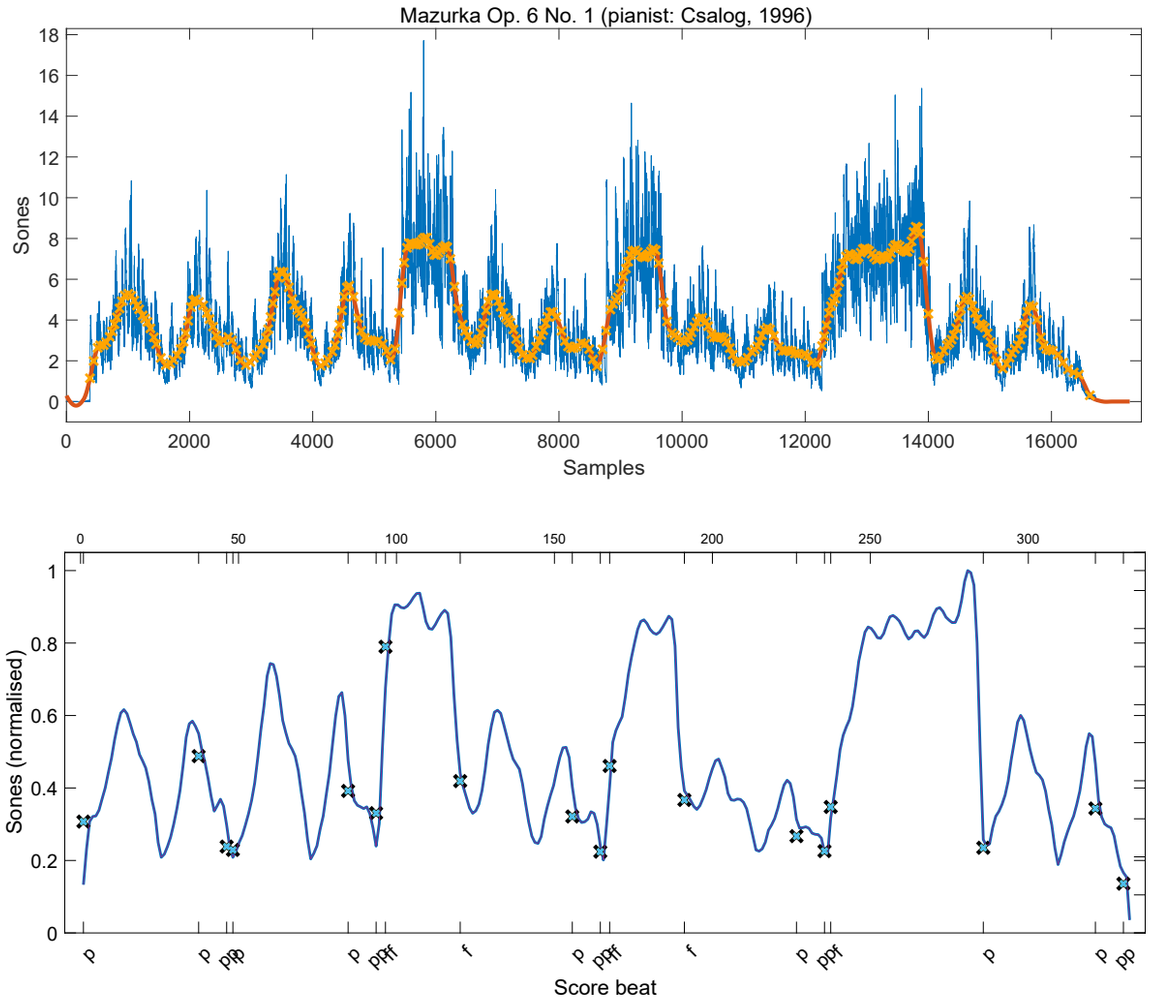

Figure 1: Loudness representation of a single Mazurka Op. 6, no. 1 recording (pianist: Csalog, recording year: 1996). Top: Smoothed curve of raw data of sone values; the score beat positions are highlighted by x's. Bottom: Normalized score beats curve; the averaged final loudness values for each marking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

times in the score at the beat positions $\{s_1, s_2, \ldots, s_n\}$,

$$\mathrm{E}(s) = \frac{1}{n} \sum_{i=1}^{n} \ell_{s_i}. \tag{5}$$

We first consider for each recording the average loudness 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.

For each recording in a particular piece, each unique pair of markings is associated with their respective average performed loudness. Each pair of these values is then compared to the expected ordering of the markings according to the OLS. A Kendall's τ value of 1 indicates perfect rank agreement with the OLS, and a value of -1 indicates the converse. For each piece we compute the mean of τ value, each corresponding to a particular marking pair and derived from a recording of that piece. The results of the mean Kendall's tau rank correlation coefficient test are summarized in Table 2.

The numbers in parentheses in Table 2 are the results of the mean Kendall's tau test following the same process as before. However, now the loudness level per marking is not derived from a three-beat window size, instead it is the average of all beats from the position of the marking until a new marking or indicator for dynamic change appears in the score. The comparison of the values highlights the exceptions to the OLS in Mazurka Op. 33 No. 4 where ***pp*** is softer than ***p*** in most instances when one considers the entire duration of the marking, Mazurka Op. 68 No. 3 where ***ff*** is louder than ***f*** in most cases if we use the bigger window, Mazurkas Op. 50 No. 1 and Op. 56 No. 3 where ***mf*** is louder than ***p***, as well as Mazurka Op. 67 No. 1 where ***pp*** is softer than ***mf***. The case of (posthumous) Mazurka Op. 67 No. 2 is of particular interest as the interpretation of all the markings does not adhere to the OLS in most cases.

The remaining negative mean τ values, highlighted in bold (and red), are further investigated and analysed in the Appendix. In summary, the case studies have shown that 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. As general conclusion, we cannot assign fixed thresholds for the different dynamic markings.

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.

	Pair								
Mazurka	{*pp*,*p*}	{*pp*,*mf*}	{*pp*,*f*}	{*pp*,*ff*}	{*p*,*mf*}	{*p*,*f*}	{*p*,*ff*}	{*mf*,*f*}	{*f*,*ff*}
M06-1	0.94(1)		0.94(1)	1(1)		0.35(0.77)	1(1)		1(0.88)
M06-2						1(1)			
M06-3	0.95(0.95)		1(1)	1(1)		1(1)	1(1)		0.71(0.57)
M07-1	0.71(1)		1(1)	1(1)		1(0.81)	1(1)		1(1)
M07-2							1(1)		
M07-3	1(0.97)		1(1)	1(1)		0.97(1)	0.97(0.97)		0.83(0.69)
M17-1						0.96(1)			
M17-2	0.48(0.68)		0.88(0.72)			0.96(0.64)			
M17-3					0.39(0.94)				
M17-4	0.08(0.67)			1(0.97)			1(0.88)		
M24-1	1(1)	1(1)			0.70(1)				
M24-2	0.96(1)		0.96(1)			0.54(0.95)			
M24-3					0.54(0.85)				
M24-4	0.89(0.96)		1(1)	1(1)		0.93(0.96)	1(1)		1(0.96)
M30-1						1(0.96)			
M30-2						0.92(0.88)			
M30-3	0.41(0.89)		1(1)	1(1)		0.96(0.82)	0.82(0.74)		**-0.33**(**-0.30**)
M30-4	0.64(0.96)		1(1)	1(1)		1(1)	1(1)		0.93(0.20)
M33-1						0.92(0.96)			
M33-2			1(1)	1(1)					0.96(0.80)
M33-3						0.83(1)			
M33-4	**-0.56**(0.97)		0.81(1)			1(1)			
M41-1	0.83(0.94)		1(1)	1(1)		1(1)	1(1)		1(1)
M41-2					0.43(0.81)	**-0.19**(**0.34**)	1(0.62)	**-0.48**(**-0.81**)	1(0.62)
M41-3						0.95(1)	1(0.74)		0.95(0.95)
M41-4	0.52(0.33)	1(0.58)	1(0.97)		0.88(0.27)	1(0.94)		0.94(0.76)	
M50-1					**-0.20**(0.42)	0.91(1)		0.91(0.82)	
M50-2						0.75(0.20)			
M50-3	0.13(0.34)		1(1)	**-0.67**(**-0.91**)		1(0.96)	**-0.94**(**-1**)		**-1**(**-1**)
M56-1					1(0.94)	1(0.94)		0.18(0.29)	
M56-2						0.83(0.96)			
M56-3					**-0.26**(0.10)	1(1)		0.96(0.96)	
M59-1						1(1)			
M59-2	1(1)		1(1)	1(1)		0.96(1)	1(1)		0.89(1)
M59-3						1(1)			
M63-1	0.95(1)		1(1)			0.95(1)			
M63-3						1(0.61)			
M67-1	0.89(0.94)	**-0.43**(0.37)	0.83(0.94)	0.94(1)	**-0.89**(**-0.77**)	0.09(0.49)	0.43(0.66)	0.94(0.77)	0.71(0.49)
M67-2	**-0.48**(**-0.55**)	**-0.29**(**-0.42**)	0.03(**-0.23**)		0.42(**-0.30**)	0.61(**-0.09**)		0.23(**-0.03**)	
M67-3	**-1**(**-0.9**)		0.45(0.65)	0.90(0.80)		1(1)	1(1)		0.90(0.90)
M67-4					0.29(0.81)	0.95(0.90)		0(0)	
M68-1						0.947(1)			
M68-2	**-0.13**(**-0.02**)	0.96(0.96)	1(1)		1(0.91)	1(1)		1(1)	
M68-3						0.95(0.95)	1(1)		**-0.10**(0.86)

Table 2: Mean Kendall's τ values for pairwise comparisons of the average loudness values that correspond to the dynamic markings for each recording and the corresponding values of the markings at the OLS: values of *'3-beat window (marking window)'*. Negative values are highlighted in bold and coloured red.

4.2 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, ($\boldsymbol{p}$, $\boldsymbol{f}$) and ($\boldsymbol{f}$, $\boldsymbol{p}$), 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 1. In this section, we consider only pairs of distinct consecutive markings.

Fig. 2 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 $\boldsymbol{p}$, is expressed by a pianist as ℓ_{k-1}, and the loudness of the k-th dynamic marking, say a $\boldsymbol{f}$, is expressed as ℓ_k, then the log ratio (as summarized in Fig. 2) is $\log(\frac{\ell_k}{\ell_{k-1}})$. The data is found to be more easily compared with the log representation, as the result can be positive or negative depending on moving from a bigger value to a smaller one or the opposite, respectively.

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 3 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 Fig. 2. 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.

Outlier in ascending pair	Mazurka (proportion of outlier pairs)
($\boldsymbol{pp}$,$\boldsymbol{p}$)	M33-4 (0.25) M41-1 (0.25) M67-3 (0.11) M68-2 (0.05)
($\boldsymbol{p}$,$\boldsymbol{mf}$)	M50-1 (0.17)
($\boldsymbol{mf}$,$\boldsymbol{f}$)	M41-2 (0.75)
($\boldsymbol{p}$,$\boldsymbol{f}$)	M41-2 (0.75) M50-1 (0.17) M50-2 (0.14) M67-1 (0.09)
Outlier in descending pair	**Mazurka (proportion of outlier pairs)**
($\boldsymbol{p}$,$\boldsymbol{pp}$)	M17-2 (0.25) M50-3 (0.08) M67-2 (0.14)
($\boldsymbol{f}$,$\boldsymbol{mf}$)	M41-2 (0.75)

Table 3: 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 {$\boldsymbol{p}$, $\boldsymbol{f}$, $\boldsymbol{mf}$, $\boldsymbol{f}$, $\boldsymbol{ff}$}, and in all transitions except

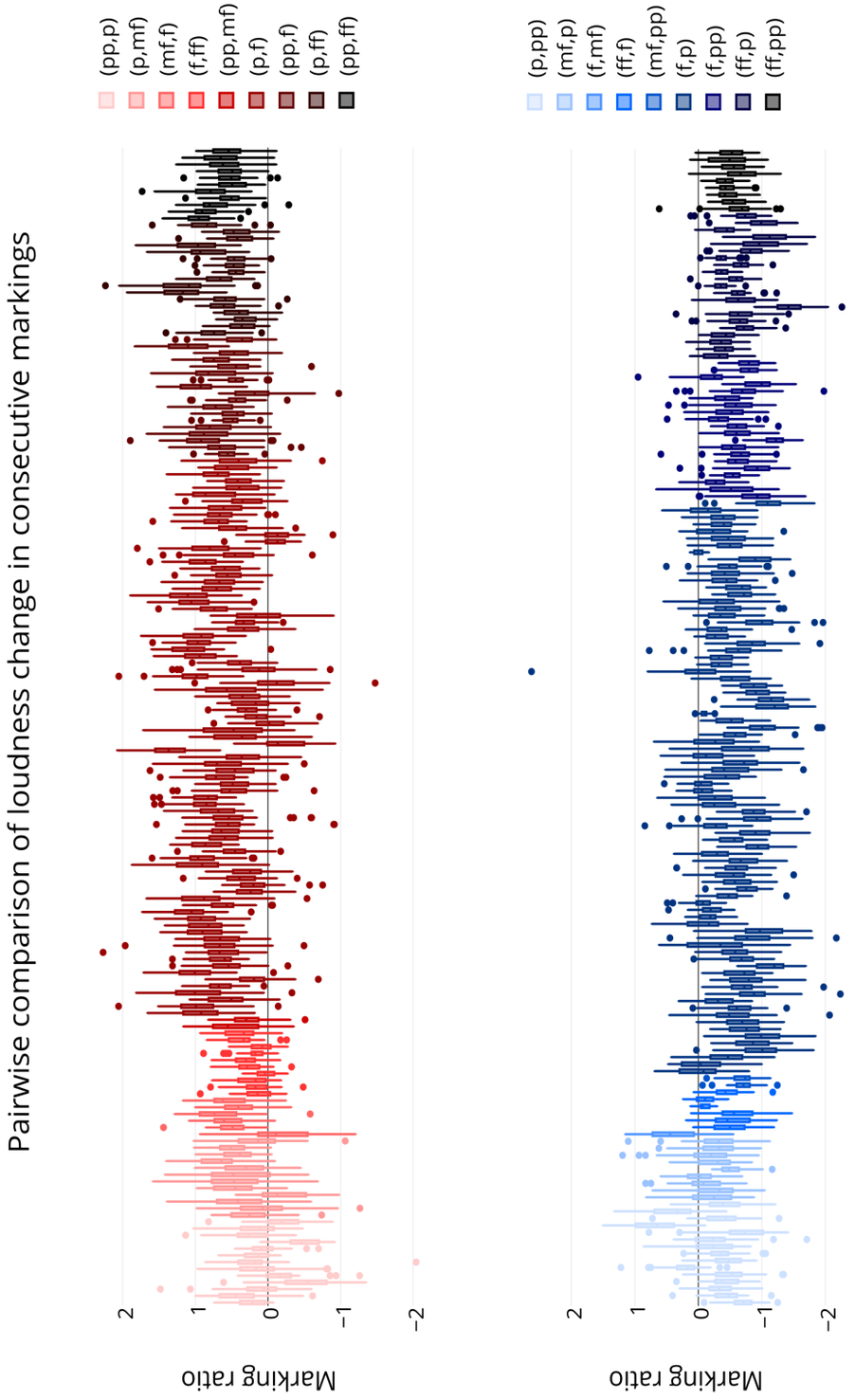

Figure 2: The log ratio of the loudness in the transition from a marking m_{k-1} to a marking m_k in the score sequence (m_{k-1}, m_k) where $\ell_{k-1} < \ell_k$ (top-left) and $\ell_k < \ell_{k-1}$ (bottom-right) following the OLS.

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 Fig. **??** in Section 4.1.

Moving on, Fig. 3 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 standard deviation (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 pianists move from a louder to a softer marking.

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).

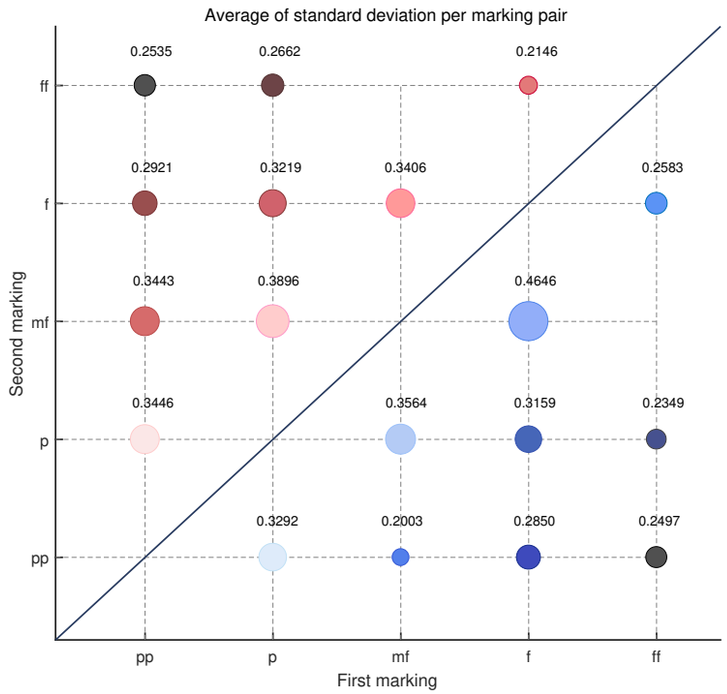

Figure 3: The average standard deviation of the log loudness ratio for each marking—transitions are from the x-axis to the y-axis.

The above observations relate to loudness transitions between consecutive and distinct markings, thus dealing with local behaviour. Having considered the average behaviour 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.

4.3 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 the population means are the same. 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 gauges how well the data supports the null hypothesis. A low P–value suggests that the sample provides enough evidence that the null hypothesis, which is that the population means are the same, can be rejected.

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 [28]. 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 article.

The two sets of ANOVA results are presented in the "P–value 1" and the "P–value 2" columns respectively in Table 4. 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. It is observed that a marking tends to be interpreted in a similar way across a piece having two occurrences of the particular marking, in contrast to markings having more than two occurrences. This phenomenon could be explained more based on features such as the distance between the same markings, the length of the piece, or the number of different markings in the piece.

Two cases exist where the P–value1 is bolded (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.

Fig. 4 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	$\cdot 10^{-10}$	$\cdot 10^{-10}$	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	$\cdot 10^{-10}$	M24-2	5	$8.7497{\cdot}10^{-7}$	$<{\cdot}10^{-10}$
M68-2	5	$\cdot 10^{-10}$	$\cdot 10^{-10}$	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	$3.4907{\cdot}10^{-6}$	$<{\cdot}10^{-10}$	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: 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. The symbol "#" indicates the number of the same markings that appear in the specific Mazurka.

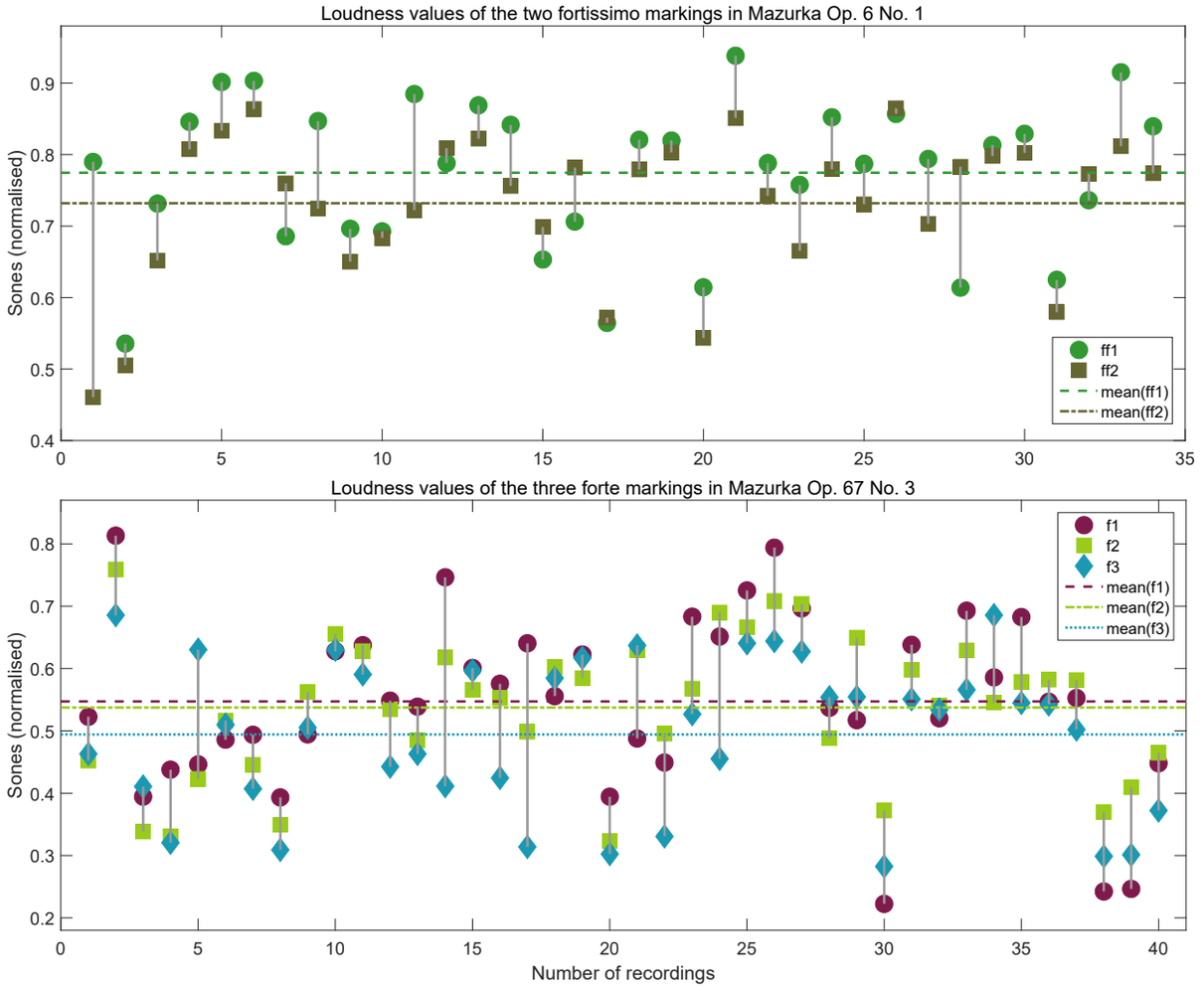

Figure 4: The case of the loudness values of groups of dynamic markings (***ff***'s in Mazurka Op. 6 No. 1—top, and ***f***'s in Mazurka Op. 67 No. 3—bottom) for which the one-way ANOVA shows no significant difference in the marking group means (dotted lines), but the two-way ANOVA, which reflects the 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 Fig. 5. In each plot in Fig. 5, 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 6; 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

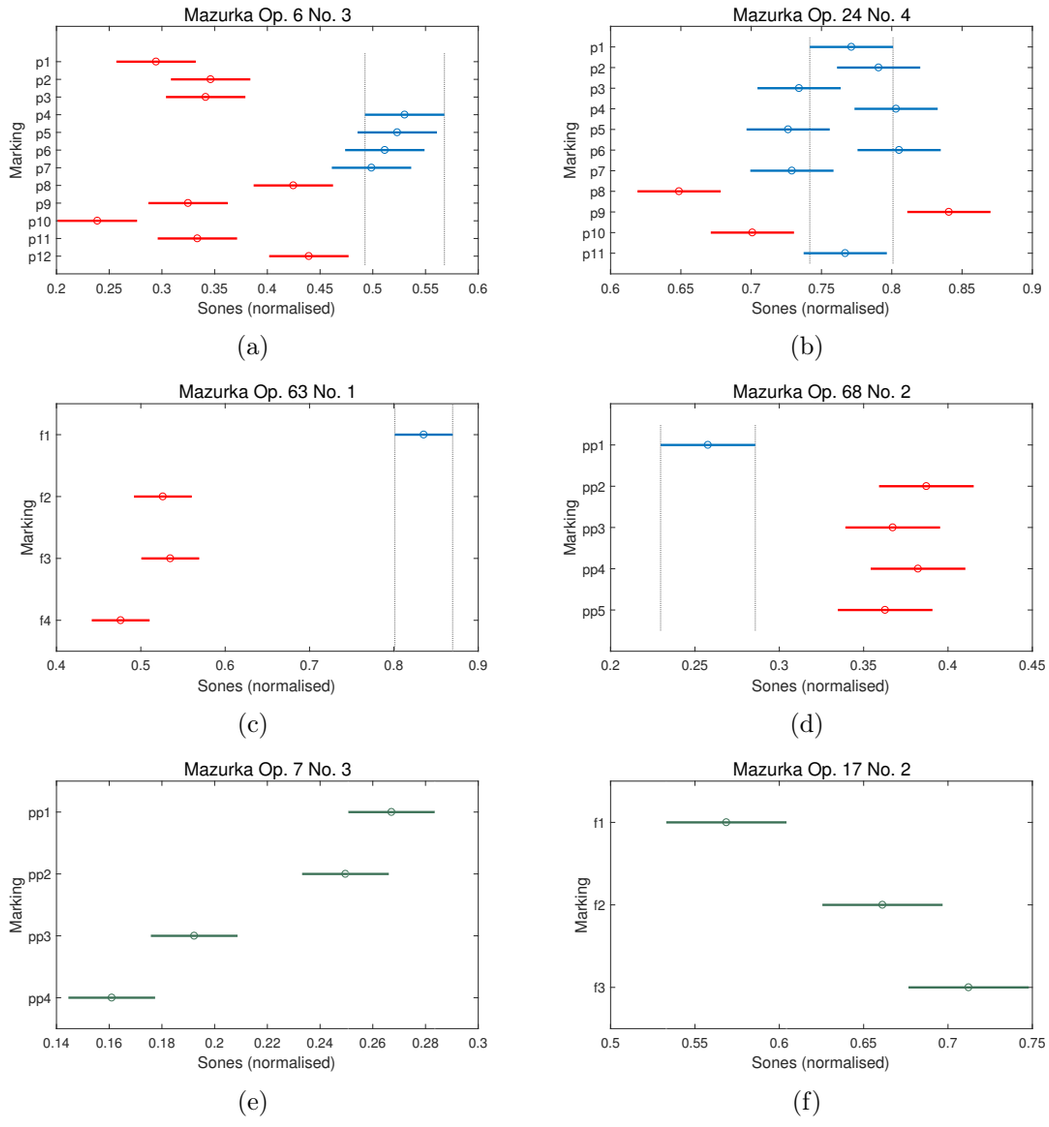

Figure 5: Multiple comparison of the means for the groups of the same markings: (a) twelve ***p***'s in Mazurka Op. 6 No 3, (b) eleven ***p***'s in Mazurka Op. 24 No. 4, (c) four ***f***'s in Mazurka Op. 63 No. 1, (d) five ***pp***'s in Mazurka Op. 68 No. 2, (e) four ***pp***'s in Mazurka Op. 7 No. 3, and (f) three ***f***'s in Mazurka Op. 17 No. 2 .

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 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

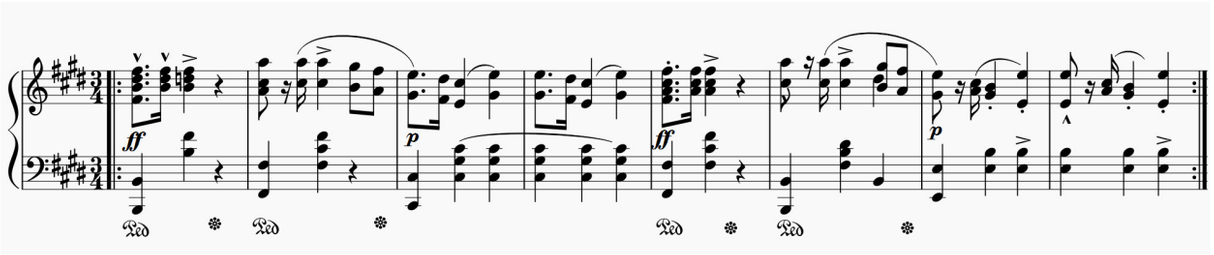

Figure 6: The repeated phrase in Mazurka Op. 6 No. 3 where the markings $\boldsymbol{p}_4$, $\boldsymbol{p}_5$, $\boldsymbol{p}_6$ (repetition of $\boldsymbol{p}_4$), and $\boldsymbol{p}_7$ (repetition of $\boldsymbol{p}_5$) appear.

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.

5 Clustering of dynamic change curves

A question that follows naturally is the degree of variation in dynamic behaviour across a single recording. In this section we present an experiment described in [22] where the MazurkaBL dataset is used. For this study, 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:

$$f(x) = \begin{cases} 2, & x > 0.5 \\ 1, & 0 < x \leq 0.5 \\ 0, & x = 0 \\ -1, & -0.5 \leq x < 0 \\ -2, & x < -0.5 \end{cases}, \tag{6}$$

where the input $x = l_b - l_{(b-1)}$ for every marking position b in each Mazurka. Then we map each recording to a time series of discretised values of the function f. Each recording 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.

We use k-means to cluster the curves obtained. The number of clusters, k, is defined using the "gap statistic" [35]; we further limit k to the range $[3, 8]$ to ensure a reasonable number of recordings appears in each cluster, but also provide the flexibility of detecting clusters that include unusual curves.

The result is a number of loudness behaviour clusters for each Mazurka. Clusters with the least number of recordings are labelled outliers. Then, we create a list of pianists that appear in these outlier clusters. Table 4 presents the top three outlier pianists who have each made recordings of more than thirty 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: Pianists having the highest proportion of recordings in outlier clusters.

Magin is the only pianist whose recordings were the only element in a cluster for Mazurka Op. 7 No. 1 and Mazurka Op. 33 No. 3. Fig. 7 shows the loudness behaviour of Magin's recordings of these two Mazurkas, and how they differ from the other recordings. Fig. 7 shows the centroid for each cluster, thus the discretised dynamic value may not be equal to any one outcome of the function f.

Magin appears to differ from the other recordings in interpreting the last two transitions in Mazurka Op. 7 No. 1, these are from ***f*** to ***pp*** and from ***pp*** to ***f***, over the last three markings out of a total of thirteen markings. This shows that even though the values of Magin's cluster are fairly close to the corresponding centroids of other clusters, the alternative interpretation of the last ***pp*** marking was crucial in separating his recording from others. In Mazurka Op. 33 No. 3 Magin interprets all four markings in a contrarian fashion, thus resulting in the creation of a separate cluster for his unique recording.

Next, we focus on 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 in the same outlier cluster for all Mazurkas. The results are shown in Table 5, 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 6: Pianists whose recordings co-occur in more than twenty outlier clusters.

Table 6 shows higher degrees of similarity in performed loudness between pianist pairs 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. Further work remains to trace their genealogy to determine teacher influences in their playing styles.

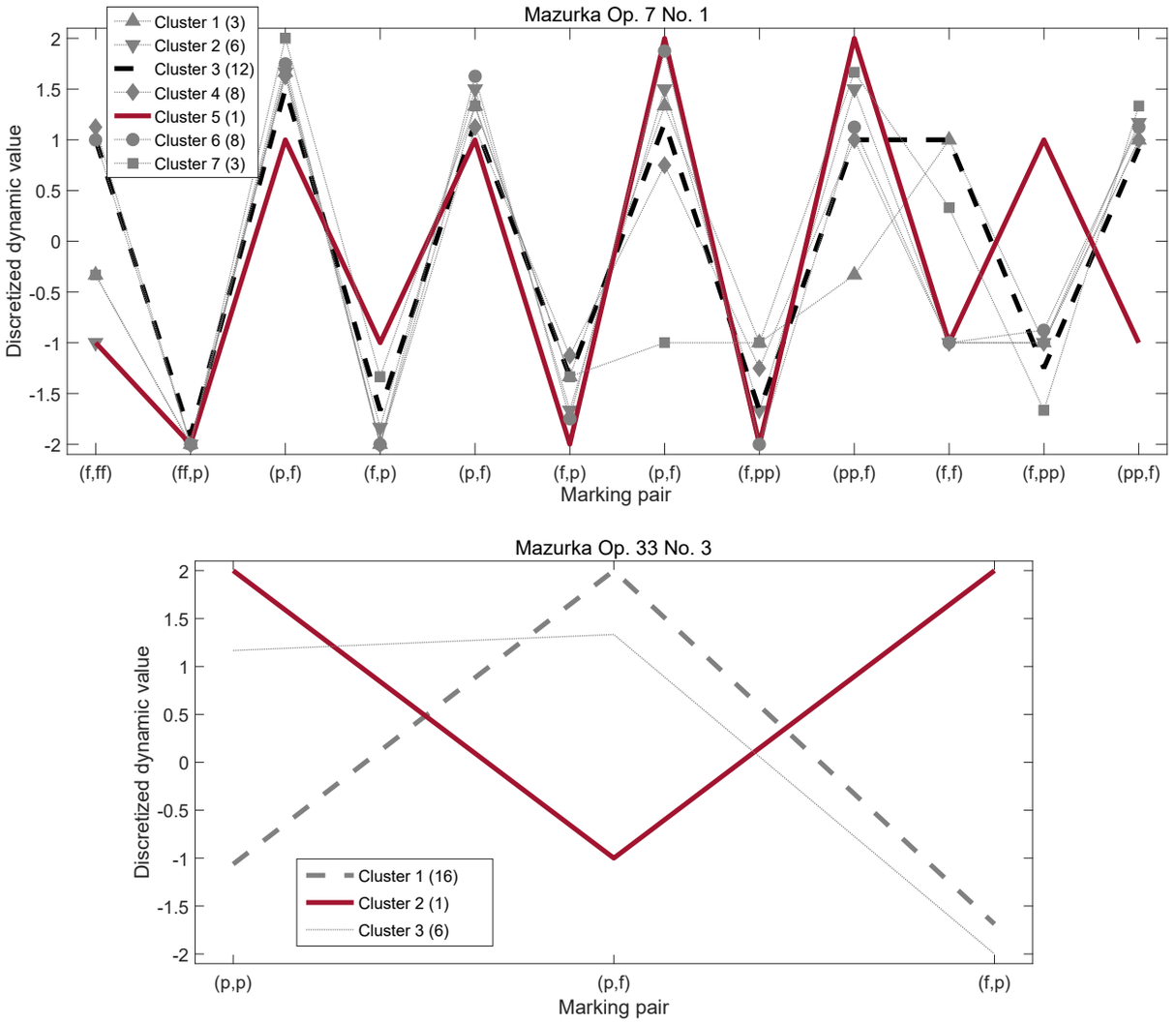

Figure 7: Loudness transition clusters found for Mazurka Op. 7 No. 1 (top) and Op. 33 No. 3 (bottom); in each case, the cluster comprising of only Magin's recording is shown in solid bold lines, the cluster having the highest number of recordings is shown as dotted bold lines.

6 Conclusions

In this article 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 [31, p.4],the manifested sound for the same effect 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 highlighted in this article 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***.

Whole-piece patterns in dynamics are discovered through studying different pianists' interpretation of the markings in the score. We were able to identify outlier loudness interpretation behaviours, and joint behaviours. Although certain pianists exhibit outlier behaviours across a large number of Mazurkas, when these pianists' outlier behaviours co-occur with other pianists, they tend to do so fairly consistently across the Mazurka repertoire.

7 Discussion

The factors that have been created through this study have played an important role in [24]; they constitute the list of features that have been extracted in order to investigate the bi-directional mapping between dynamic markings in the score and performed loudness. The study applied machine-learning techniques 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 on different recordings of the same piece, but fail dismally when trained on the pianist's recordings of other pieces, demonstrating that score features may trump individual style when modelling loudness choices. The evidence suggested that all the features chosen for the prediction and classification tasks—current/previous/next dynamic markings, distance between markings, and proximity of dynamic-related and non-dynamic markings —were relevant. Furthermore, analysis of the results reveal the forms (such as the return of the theme) and structures (such as dynamic marking repetitions) influence the predictability of loudness levels and dynamic markings.

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 resulting loudness. Also the extraction of structural parameters, such as the marking position as it relates to positions within phrase could warrant further investigation; there are many ways to interpret a music piece utilizing both compositional and expressive parameters, with respect to structural aspects such as phrasing [34] and pitch [14].

With regard to tempo changes; [36] examined tempo-loudness interactions at specific score markings over a subset of the MazurkaBL dataset, and investigated how including information on one parameter impacted prediction of the other. The authors considered score markings indicating loudness or tempo change, and the model included score, tempo, and loudness-related features. When considering recordings of the same Mazurka, experiments showed that considering loudness-related features did not improve prediction of tempo change. However, adding tempo-related features did result in marginal improvement in predicting loudness change.

Other future directions include the exploration of new ways to capture the loudness level corresponding to a marking using different loudness detection techniques.

It is important to remain conscious of the limitations of score–based analyses: deviations from the score can occur in performance, such as playing notes in an octave different from that written or applying rubato where not indicated. More general, working from score-based analysis to recording may be considered as declaring "off limits all those aspects of performance that cannot be directly related to notational categories", eliminating the unnotatable aspects of rhetoric, persuasion, and expressive effects having little or nothing to do with theoretical structure [8, p. 233].

Other deviations from the score occur when performers prefer an aberrant interpretation style which deviates from the score. From the standpoint of the researcher, it can be unclear whether "the expressive deviations measured are due to deliberate expressive strategies, music structure, motor noise, imprecision of the performer, or even measurement errors" [25], for "even well-known scores that appear to have widely-recognised meanings are changing all the time, not simply as general performance style changes but in their characterisation, leading to a change in their perceived nature." [26]

It is not clear what an analysis of the dynamics of a performance represents when based on an audio recording. Do the dynamics result from 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 hear, which could be different from what was actually played. The type of instrument and the room acoustics may also affect the end result. 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 comments 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.

Acknowledgements

This research has been funded in part by a Queen Mary University of London Principal's studentship. The authors would like to thank Jordan Smith, Ph.D., for advice on statistical tools used for this research.

References

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

[2] Jacob Beck and William A. Shaw. "Ratio-estimations of loudness intervals". In: *Journal of the acoustical society of America* 80 (1967), pp. 59–65.

[3] 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`.

[4] 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.

[5] Erica Bisesi and Richard Parncutt. "Second Vienna Talk". In: *Proceedings of the Second Vienna Talk, University of Performing Arts Vienna*. 2010, pp. 26–30.

[6] R. A. W. Bladon and Björn Lindblom. "Modeling the judgment of vowel quality differences". In: *The Journal of the Acoustical Society of America* 69.5 (1981), pp. 1414–1422.

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

[8] Nicholas Cook et al. *The Cambridge companion to recorded music*. Cambridge University Press, 2009.

[9] 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*. 2010, pp. 637–642.

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

[11] 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)*. Taipei, Taiwan, 2009, pp. 1869–1872.

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

[13] Sebastian Flossmann et al. "The Magaloff project: An interim report". In: *Journal of New Music Research* 39.4 (2010), pp. 363–377.

[14] Anders Friberg, Roberto Bresin, and Johan Sundberg. "Overview of the KTH rule system for musical performance". In: *Advances in Cognitive Psychology* 2.2-3 (2006), pp. 145–161.

[15] John M Geringer. "Loudness estimations of noise, synthesizer, and music excerpts by musicians and nonmusicians." In: *Psychomusicology: A Journal of Research in Music Cognition* 12.1 (1993), p. 22.

[16] Werner Goebl. "Melody lead in piano performance: Expressive device or artifact?" In: *The Journal of the Acoustical Society of America* 110.1 (2001), pp. 563–572.

[17] Maarten Grachten and Florian Krebs. "An Assessment of Learned Score Features for Modeling Expressive Dynamics in Music". In: *IEEE Transactions on Multimedia* 16.5 (Aug. 2014), pp. 1211–1218. ISSN: 1520-9210.

[18] Maarten Grachten and Gerhard Widmer. "Linear Basis Models for Prediction and Analysis of Musical Expression". In: *Journal of New Music Research* 41.4 (2012), pp. 311–322. eprint: http://dx.doi.org/10.1080/09298215.2012.731071. URL: http://dx.doi.org/10.1080/09298215.2012.731071.

[19] Maarten Grachten et al. "Towards computer-assisted understanding of dynamics in symphonic music". In: *IEEE Multimedia: Special Issue on Multimedia Technologies for Enriched Music Performance, Production and Consumption* 24.1 (2017), pp. 36–46.

[20] Agustín Gravano et al. "On the role of context and prosody in the interpretation of 'okay'". In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistic (ACL)*. 2007, pp. 800–807.

[21] Katerina Kosta, Oscar F. Bandtlow, and Elaine Chew. "MazurkaBL: Score-aligned loudness, beat, expressive markings data for 2000 Chopin Mazurka recordings". In: *(to appear)*. 2018.

[22] Katerina Kosta, Oscar F. Bandtlow, and Elaine Chew. "Outliers in performed loudness transitions: an analysis of Chopin Mazurka recordings". In: *Proceedings of the 14th International Conference for Music Perception and Cognition (ICMPC)*. San Fransisco, California, USA, 2016, pp. 601–604.

[23] 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.

[24] Katerina Kosta et al. "Mapping between dynamic markings and performed loudness: A machine learning approach". In: *Journal of Mathematics and Music* 10.2 (2016), pp. 149–172. eprint: http://dx.doi.org/10.1080/17459737.2016.1193237. URL: http://dx.doi.org/10.1080/17459737.2016.1193237.

[25] Jörg Langner and Werner Goebl. "Visualizing Expressive Performance in Tempo-Loudness Space". In: *Computer Music Journal* 27.4 (Dec. 2003), pp. 69–83. ISSN: 0148-9267. URL: http://dx.doi.org/10.1162/014892603322730514.

[26] Daniel Leech-Wilkinson. "Compositions, scores, performances, meanings". In: *Music Theory Online* 18.1 (2012), pp. 1–17. URL: http://mtosmt.org/issues/mto.12.18.1/mto.12.18.1.leech-wilkinson.php.

[27] *MazurkaBL dataset*. https://github.com/katkost/MazurkaBL. Accessed: 2018-01-03.

[28] Regina Nuzzo. "Scientific method: Statistical errors". In: *Nature* 506.7487 (2014), pp. 150–152. URL: http://www.nature.com/news/scientific-method-statistical-errors-1.14700.

[29] Caroline Palmer and Sean Hutchins. "What Is Musical Prosody?" In: vol. 46. Psychology of Learning and Motivation. Academic Press, 2006, pp. 245–278. URL: http://www.sciencedirect.com/science/article/pii/S0079742106460072.

[30] E. Pampalk, A. Rauber, and D. Merkl. "Content-based Organization and Visualization of Music Archives". In: *Proceedings of the ACM Multimedia*. Juan les Pins, France: ACM, Dec. 2002, pp. 570–579.

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

[32] María Ros et al. "Transcribing Debussy's Syrinx dynamics through Linguistic Description: The MUDELD algorithm". In: *Fuzzy Sets and Systems* 285 (2016), pp. 199–216. URL: http://dx.doi.org/10.1016/j.fss.2015.08.004.

[33] Manfred R Schroeder, Bishnu S Atal, and JL Hall. "Optimizing digital speech coders by exploiting masking properties of the human ear". In: *The Journal of the Acoustical Society of America* 66.6 (1979), pp. 1647–1652.

[34] J. Sundberg. *Gluing tones: grouping in music composition, performance and listening*. Kungl. Musikaliska akademiens skriftserie. Royal Swedish Academy, 1992. ISBN: 9789185428731. URL: https://books.google.co.uk/books?id=eM45AQAAIAAJ.

[35] Robert Tibshirani, Guenther Walther, and Trevor Hastie. "Estimating the number of clusters in a data set via the gap statistic". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.2 (2001), pp. 411–423.

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

[37] Gerhard Widmer and Werner Goebl. "Computational Models of Expressive Music Performance: The State of the Art". In: *Journal of New Music Research* 33.3 (2004), pp. 203–216.

8 Appendix

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. 8 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. Also, by observing the changes in tempo, we notice that the performers tend to play slower at the locations of the ***ff*** and the second ***f***.

In Fig. 9, the first two bars show the score area where the 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 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 Fig. 10 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 ***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 Fig. 10. 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.

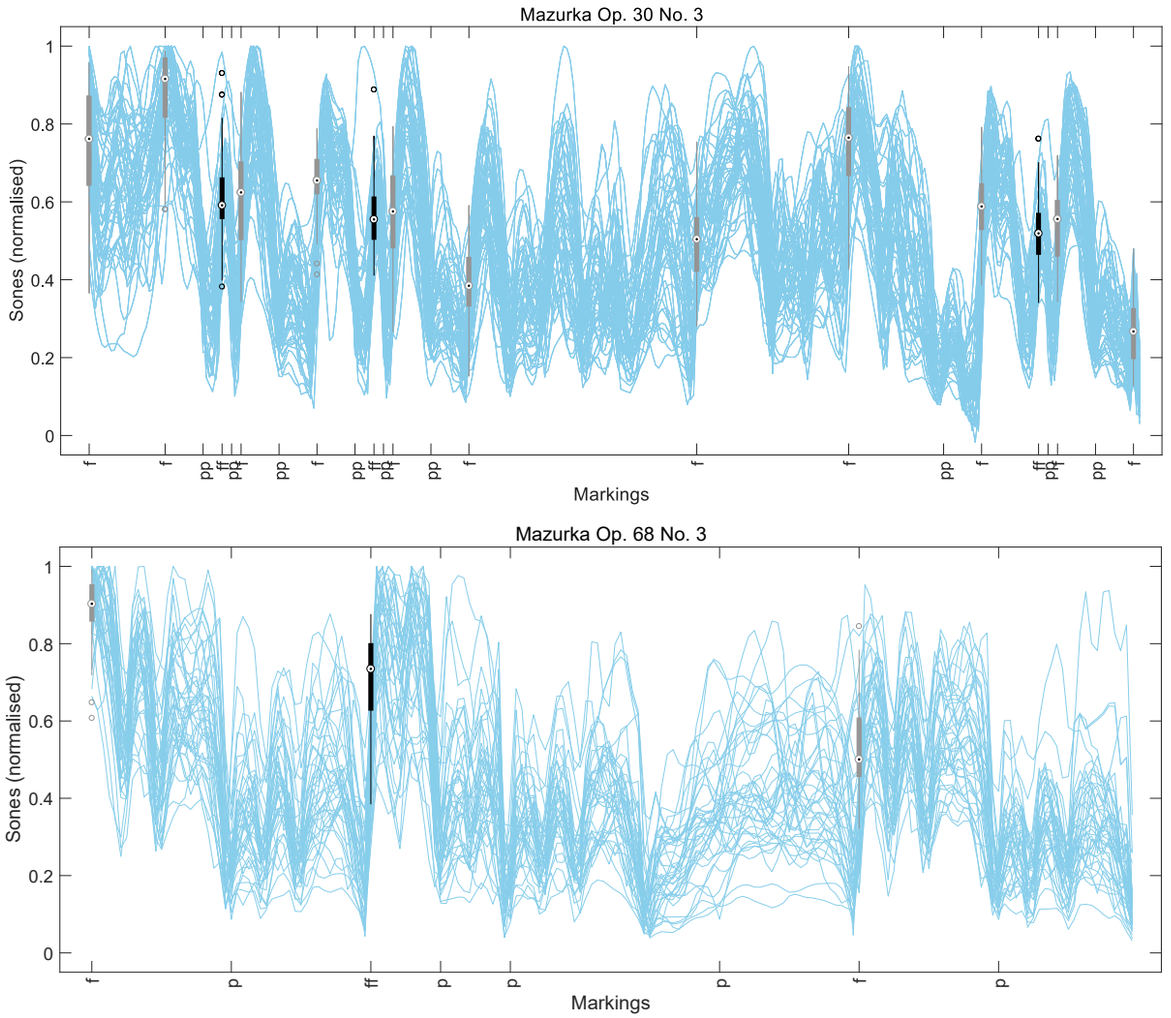

Figure 8: The Mazurka cases where only the ***f*-*ff*** pair has negative τ value. The dynamic values of the markings belonging to the pair in Mazurka Op. 30 No. 3 (top), and Mazurka Op. 68 No. 3 (bottom) are presented as box plots. Positions of other markings $s \in \mathrm{S}$ appear in x-ticks as in score sequence. At the background the curves represent the score—beat dynamic values per recording.

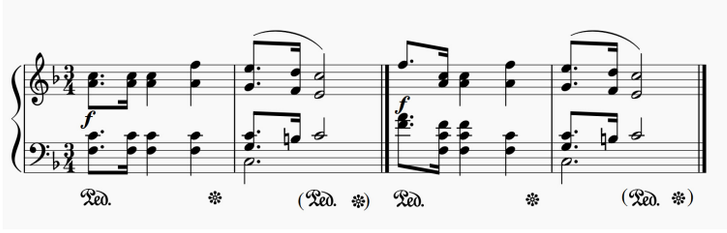

Figure 9: The repeated phrase where the ***f*** appears in Mazurka Op. 68 No. 3 (left part: score beats 1–2, right part: score beats 133–134.)

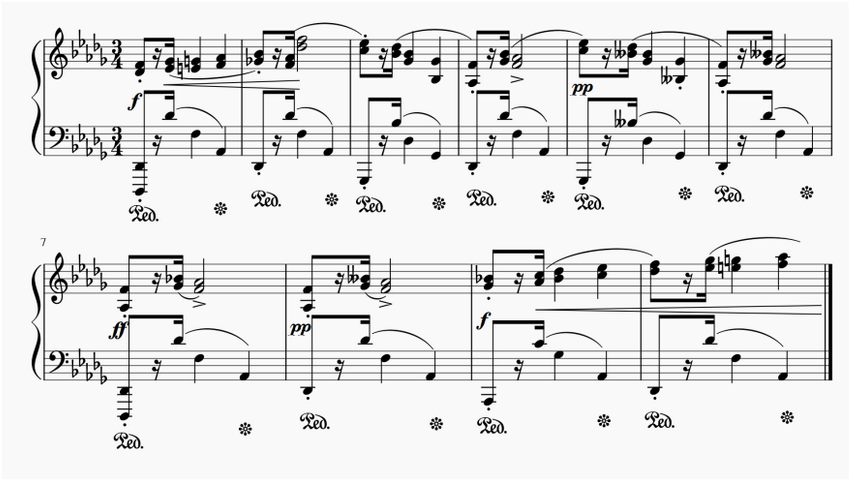

Figure 10: The repeated phrase where the ***ff*** appears in Mazurka Op. 30 No. 3 and its relation to the neighbouring markings ***pp***, and ***f*** (score beats 25–54, 73–102, and 283–312.)

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). Fig. 11 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 Fig. 11. Many bars separate the ***pp*** and the ensuing ***p***, as shown in Fig. 12, 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 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 Fig. 13 shows, every ***pp*** is preceded by a sustained note in ***sf***, which might be the reason for the higher loudness level.

In Mazurka Op. 68 No. 2, there is a middle part, where there are four closely clustered humps, as they can be observed from the background curves in Fig. 11, which correspond 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.

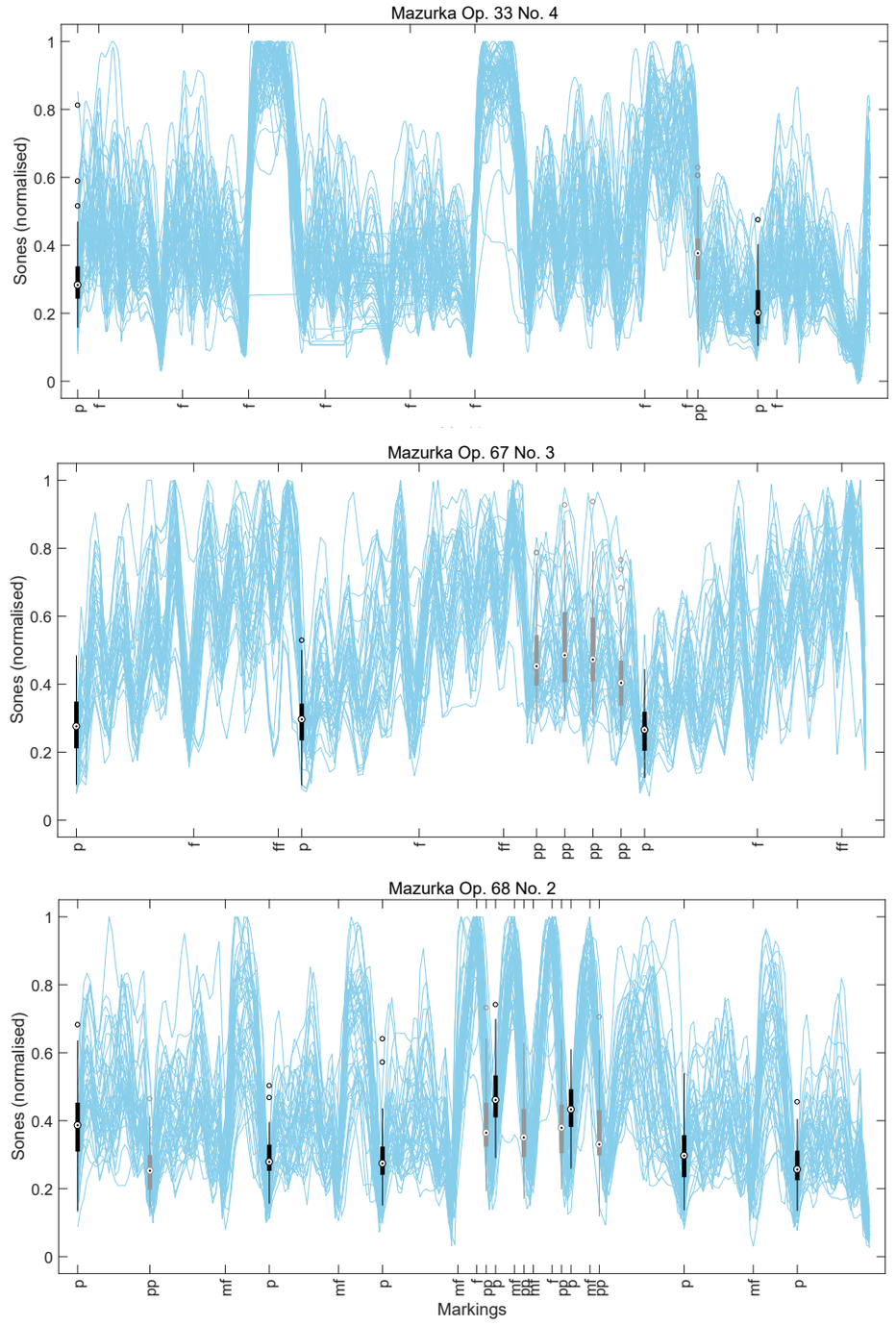

Figure 11: The Mazurka cases where only the ***pp*-*p*** pair has negative τ value. Box plots of the dynamic values of the markings belonging to the pair in Mazurkas Op. 33 No. 4, Op. 67 No. 3, and Op. 68 No. 2 are presented. Positions of other markings $s \in$ S appear in x-ticks as in score sequence. At the background the curves represent the score–beat dynamic values per recording.

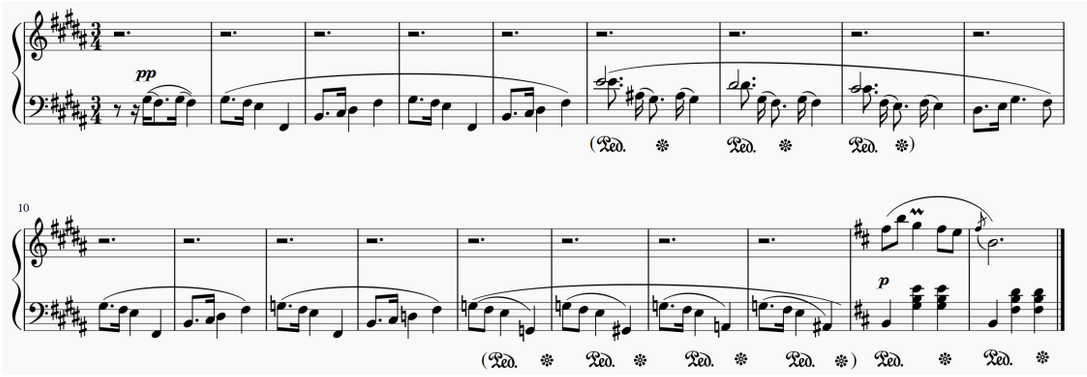

Figure 12: The position of the ***pp*** marking in Mazurka Op. 33 No. 4. The response of the recordings at this position is higher than the response at the position of the ***p*** marking on average.

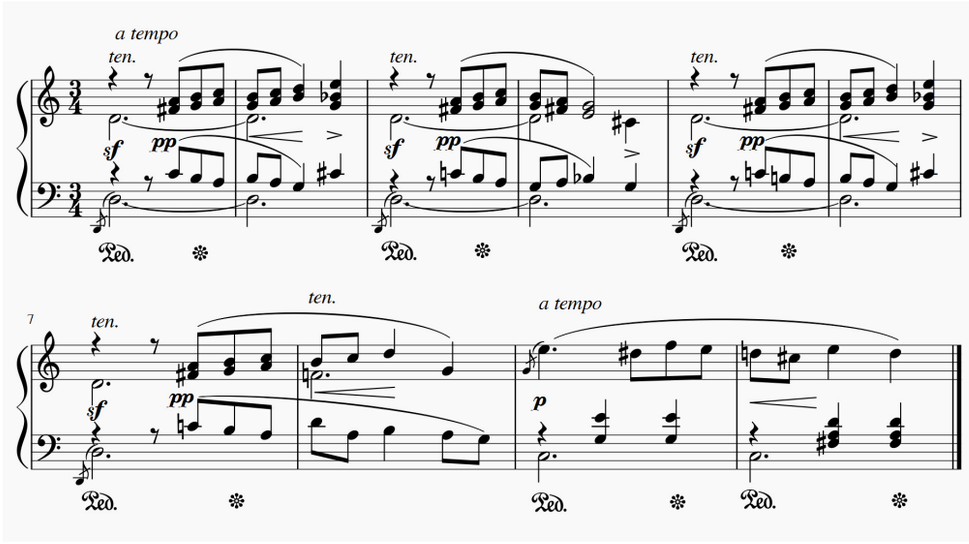

Figure 13: The ***pp*** markings in Mazurka Op. 67 No. 3. The response of the recordings at these positions is higher than the response at the positions of the ***p*** markings on average.

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). Fig. 14 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.

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 in Fig. 14 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. Fig. 15 shows the score

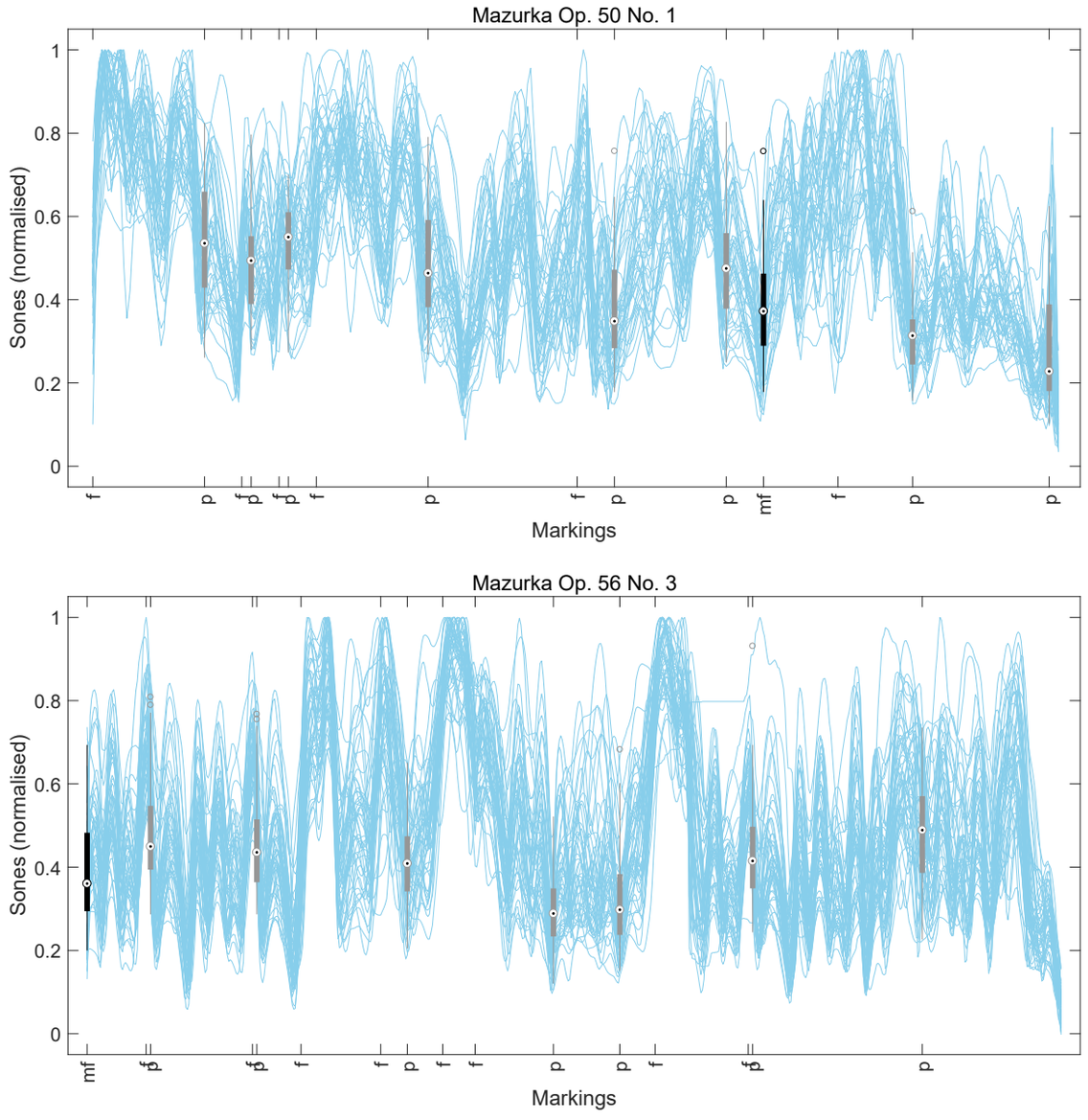

Figure 14: The Mazurka cases where only the ***p*-*mf*** pair has negative τ value. The dynamic values of the markings belonging to the pair in Mazurka Op. 50 No. 1 (top), and Mazurka Op. 56 No. 3 (bottom) are presented as box plots. Positions of other markings $s \in \mathrm{S}$ appear in x-ticks as in score sequence. At the background the curves represent the score–beat dynamic values per recording.

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

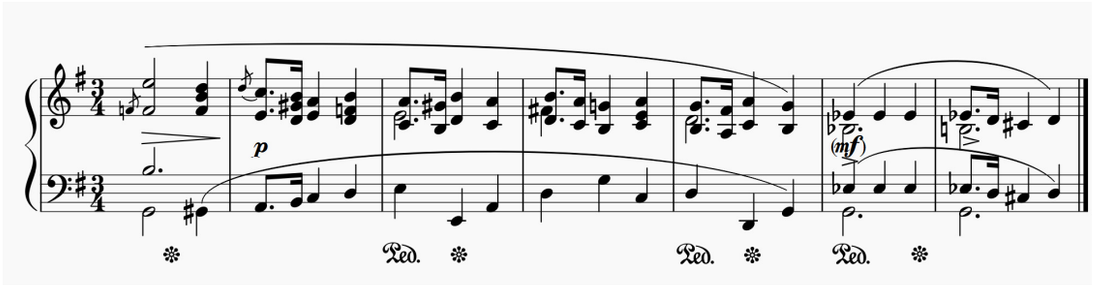

Figure 15: The ***mf*** marking in Mazurka Op. 50 No. 1 and its relation with the preceded ***p*** marking.

bar immediately before the ***p***. When the 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). Fig. 16 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.

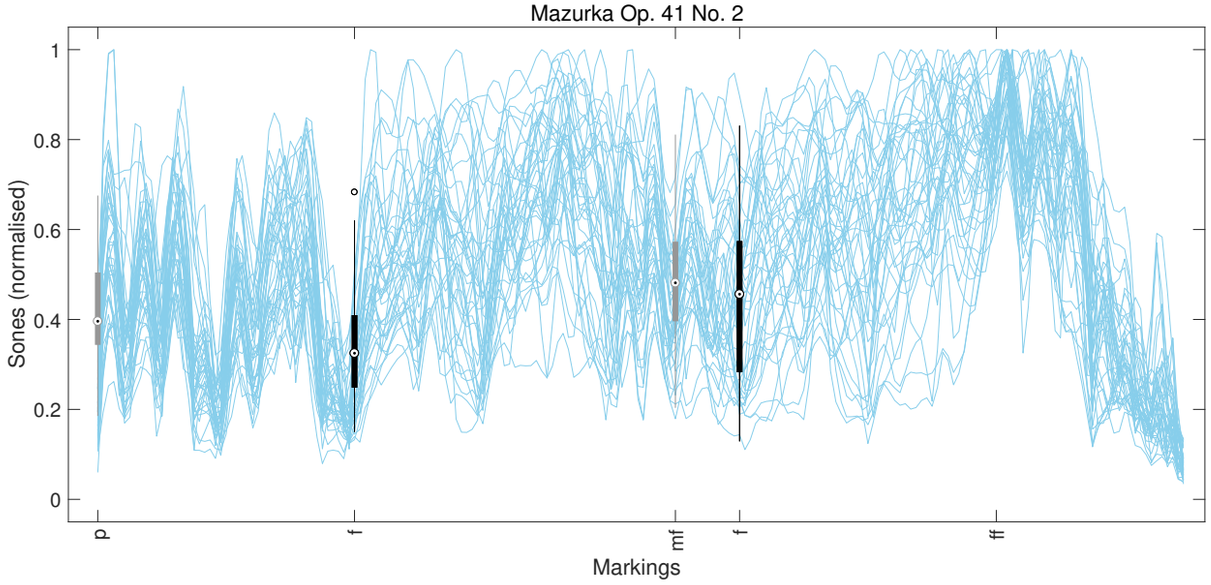

Figure 16: Box plots of the dynamic values of the markings belonging to the pairs ***p*-*f***, and ***mf*-*f*** in Mazurka Op. 41 No. 2. Positions of other markings $s \in S$ appear in x-ticks as in score sequence. At the background the curves represent the score–beat dynamic values per recording.

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 Fig. 17 the location of the two $\boldsymbol{f}$'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 $\boldsymbol{f}$. This is followed by a short ***Crescendo***, which allows the $\boldsymbol{f}$ 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 $\boldsymbol{f}$. The $\boldsymbol{p}$ 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 $\boldsymbol{mf}$ marking, which can be seen at the bottom part of Fig. 17, is inside parentheses, which suggests freedom on how it could be interpreted.

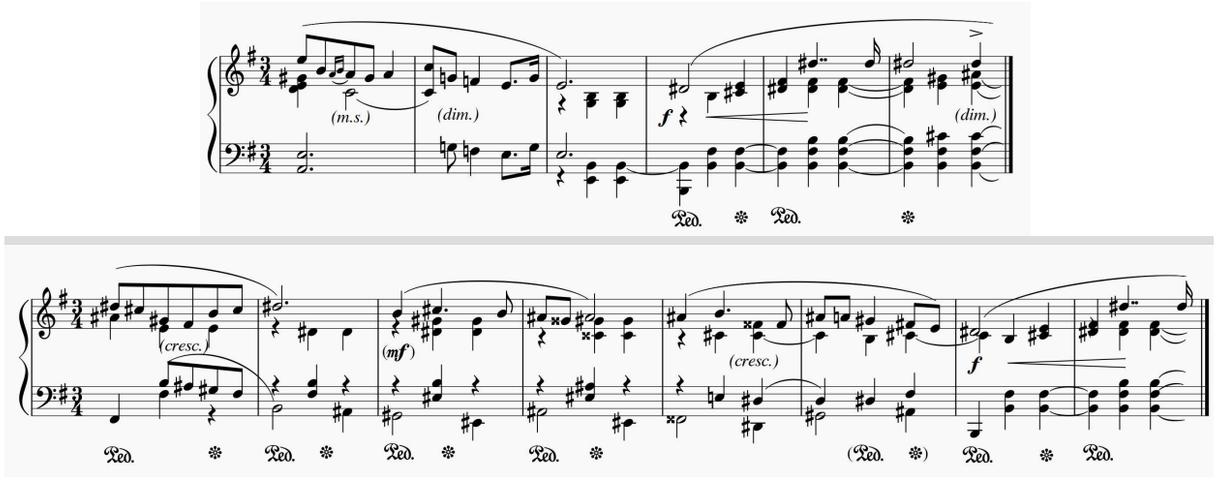

Figure 17: The location of the first $\boldsymbol{f}$ marking (top), and the location of the second $\boldsymbol{f}$ marking in Mazurka Op. 41 No. 2 (bottom), score-beats 40–57, and 103–126, respectively.

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. Fig. 18 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***.

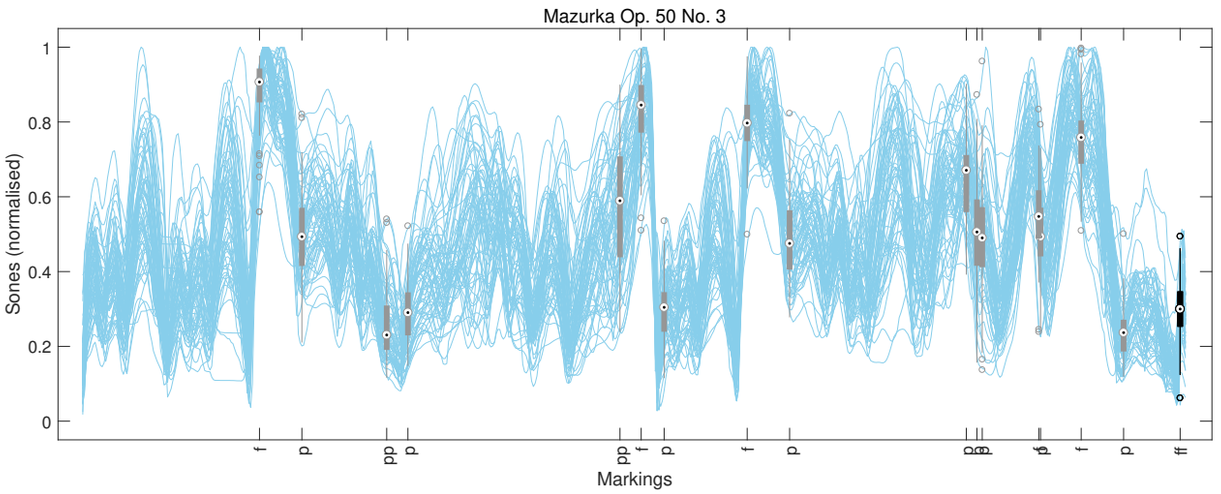

Figure 18: Box plots of the dynamic values of the markings belonging to the pairs (***pp***, ***ff***), (***p***, ***ff***), and (***f***, ***ff***) in Mazurka Op. 50 No. 3. Positions of other markings $s \in$ S appear in x-ticks as in score sequence. At the background the curves represent the score–beat 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). Fig. 19 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.

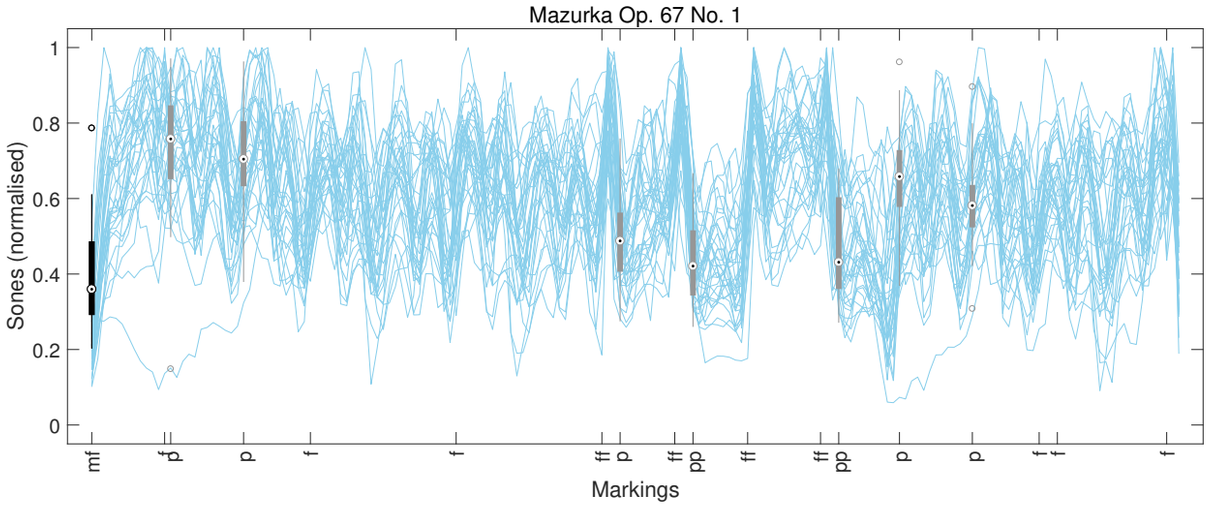

Figure 19: Box plots of the dynamic values of the markings belonging to the pairs (***pp***, ***mf***) and (***p***, ***mf***) in Mazurka Op. 67 No. 1. Positions of other markings $s \in$ S appear in x-ticks as in score sequence. At the background the curves represent the score–beat dynamic values per 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). Fig. 20 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 Fig. 20. 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.

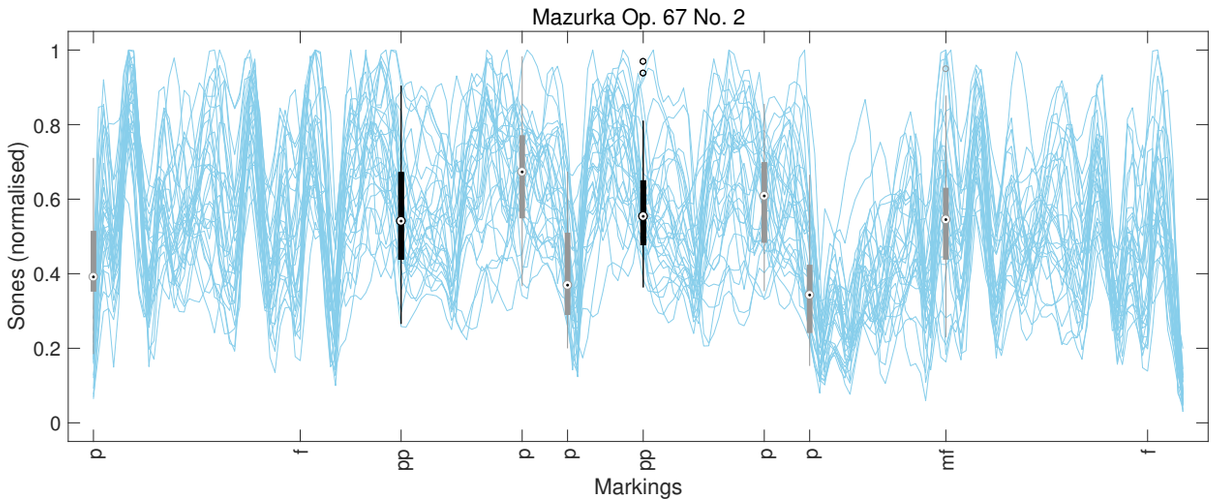

Figure 20: Box plots of the dynamic values of the markings belonging to the pair ***pp***–***p*** and the pair ***pp***–***mf*** in Mazurka Op. 67 No. 2. Positions of other markings $s \in \mathrm{S}$ appear in x-ticks as in score sequence. At the background the curves represent the score–beat dynamic values per recording.

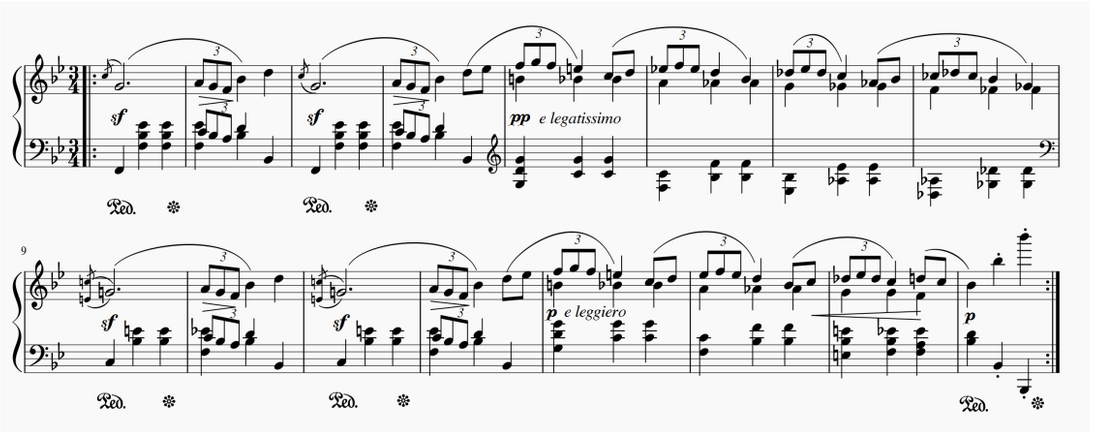

Figure 21: The repeated phrase in Mazurka Op. 67 No. 2 which includes the ***pp*** marking, and the two ***p*** markings that follow up.