MazurkaBL: SCORE-ALIGNED LOUDNESS, BEAT, AND EXPRESSIVE MARKINGS DATA FOR 2000 CHOPIN MAZURKA RECORDINGS

Katerina Kosta
Centre for Digital Music,
Queen Mary University of London,
London, UK
katkost@gmail.com

Oscar F. Bandtlow
School of Mathematical Sciences,
Queen Mary University of London,
London, UK
o.bandtlow@qmul.ac.uk

Elaine Chew
Centre for Digital Music,
Queen Mary University of London,
London, UK
elaine.chew@qmul.ac.uk

ABSTRACT

Large-scale analysis of expressive performance—with focus on how a performer responds to score markings—has been limited by a lack of big datasets of recordings with accurate beat and loudness information with score markings. To bridge this gap, we created the MazurkaBL dataset, a collection of score-beat positions and loudness values, with corresponding score dynamic and tempo markings for 2000 recordings of forty-four Chopin Mazurkas. MazurkaBL forms the largest annotated expressive performance dataset to date. This paper describes how the dataset was created, and variations found in the dataset. For each Mazurka, the recordings were first aligned to the score and one to another to facilitate the transfer of meticulously created manual beat annotations from one reference to all other recordings. We propose a multi-recording alignment heuristic that optimises the reference audio choice for best average alignment results. Loudness values in sones are extracted and analysed; we also provide the score position of dynamic and tempo markings. The result is a rich repository of score-aligned loudness, beat, and expressive marking data for studying expressive variations. We further discuss recent and future applications of MazurkaBL and future directions for database development.

1. INTRODUCTION

The musical score provides an incomplete representation of a composer's intended expressions for the rendering of a piece. How a performer responds to these instructions can vary widely, and has increasingly become an important area of study in recent years. However, systematic analyses of score-informed performance data has been beset by a lack of large datasets with appropriate information, such as synchronisation between performance and score information, and between performances, essential for comparing audio features and prosodic decisions along with score representation. Synchronisation is often done through beat alignment. This is particularly problematic for music with large tempo and timing deviations as current automatic beat tracking methods perform poorly for such music. Alignment between highly expressive music audio and symbolic score information is also fraught with error, requiring manual intervention. The problems are typically circumvented by manual annotation, which does not scale well to large datasets.

As a result, only a limited number of datasets exist for highly expressive music that is score-aligned and synchronised with expressive features; of these, few have large numbers of recordings of the same pieces or do so with only a handful of pieces. Table 1 shows a representative sample of such datasets, together with the expressive information layers they provide. As can be seen, there is a lack of a systematic collection of annotations for a large number of recordings that represented a range of interpretations of the same music pieces.

To bridge this gap, we created the MazurkaBL dataset, which augments 2000 recordings from the CHARM [1] Chopin Mazurka Project database with expressive information layers containing score-beat positions, loudness values, and locations and labels of score-based dynamic and tempo markings. The Mazurka Project database has been the subject and object of a few previous studies. For example, Sapp [6] created hierarchical scape plots for visual analysis of tempo and loudness similarity at multiple timescales. The dataset also provided material for testing beat tracking algorithms (eg. [7] and [8]) and for creating robust tempo-based novelty detection functions by harnessing simultaneous analyses of multiple recordings of the same piece [9].

The rationale for focusing on Chopin's Mazurkas is not only because the Mazurka dataset exists. For the majority of pianists, and indeed other instrumentalists as well, the Romantic repertoire presents a wealth of expressive possibilities [10]. The reason for indexing the recordings by score beat information and expressive markings is because the score encapsulates the composer's intentions while the recording reflects the performer's interpretation of the notated score. Each symbol—be it a note, dynamic marking, indication of articulation, or phrase grouping—can have a variety of possible interpretations. In performance studies, the original score is considered to be refracted through the performer [11, p.59], who can choose to render the symbols in unique ways. In order to understand expressivity, it is important to be able to have recordings, and hence au-

[1] http://www.charm.rhul.ac.uk/index.html

Dataset	*Description*	*No.*	*Expression annotations*
MazurkaBL	Forty-four Chopin Mazurkas (audio)	2000	beat, loudness, expr marks
Mazurka-CS [1]	Five Chopin Mazurkas (audio)	239	beat, loudness, phrase, expr marks
Magaloff Project [2]	Complete works of Chopin (Bösendorfer)	155	Midi-score alignment
Saarland Music Dataset [3]	Selected piano pieces (Disklavier, audio)	50	Midi-audio alignment, pedal
EEP Dataset [4]	String quartet movements w mocap	23	bowings
QUARTET Dataset [5]	Intonation, dynamics, phrasing exercises (audio, mocap, video)	95	bowings

Table 1: Datasets annotated with expression markings and parameters.

dio features, aligned to the score so that comparisons can be drawn between the performer's choices and the composer's notations.

The choice of annotating beats, from which one can infer tempo and timing information, and providing loudness values synchronised with notated expressive markings follows that of [1].[2] In order to obtain reliable measurements and scalable analyses, we rely on computational audio analysis tools, which despite their imperfections are becoming standard tools for empirical musicologists [12, p. 225–233]. The large scale in which we were able to deploy the beat annotation and calculation of tempo and loudness values was made possible by a state-of-the-art audio-to-audio alignment technique [13]. Only one recording was painstakingly annotated with beat information, and that annotation transferred to all other recordings. A multi-alignment heuristic, described later in this paper, optimised the choice of reference recording for the alignment procedure.

A large dataset facilitates systematic and empirical studies aimed at understanding the range of expressive possibilities proffered by a score. It also enables the design of robust statistical models that can capture the range of possible expressive variations. A big dataset will allow scholars to discern what constitutes a typical style for the performance of a piece. Knowledge of this performance style can, in turn, constrain parameters in models of expressive performance. It also allows researchers to identify what constitutes an outlier in performance style.

The paper is organised as follows: Section 2 gives an introduction to the development of expressive notation and its use in Chopin's works, Section 3 describes the MazurkaBL dataset, Section 4 presents the method created for obtaining score beat positions from audio recordings, Section 5 presents the method used to extract loudness information from the audio, Section 6 describes studies that have used the MazurkaBL dataset, and Section 7 offers some future directions.

2. DEVELOPMENT OF EXPRESSIVE NOTATION AND CHOPIN'S WORKS

This section gives a brief introduction to the concept of music notation, the symbolic representation of music in written form, so that it can be reproduced, as it developed in European classical music, and the development of expressive notation in Chopin's works.

The neumatic notation—from the Greek word "neuma" meaning "gesture" or "sigh"—was the system of musical notation used from the 7th to 14th century. It evolved from grave and accute accents to a system of precise indications of pitch for singing [14]. Referring to a study by Sam Barrett [15], which posits that neumatic notation is more than a memory aid, being a "reflexive tool for disciplined knowing", Cook [16, pg.11] concludes that music is "conceived platonically, as an abstract and enduring entity that is reflected in notation".

Developments to music notation as we know it today mainly involved changes on the representation of the duration and pitch of the notes that are sounded. Innovations included the development of notational symbols for different playing techniques and performance actions. Giovanni Gabrieli (1554-1612) was the first composer to specify dynamics in a score, in the *Sonata pian e forte* from the Sacrae symphoniae (1597) [17, pg. 28–29]. Annotation of dynamics, such as ***p*** for *piano*, "has remained relatively constant, although contemporary composers have explored its extremes." [14]

Next we consider the use of expressive notation in Chopin's works. Chopin's compositions can be best understood through 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 [18, p.150]. [19] suggests that Chopin was influenced by late baroque and pre-classical composers; however, J. S. Bach's imprint can be found in his later works.

Searching for the characteristics that make a performance 'musical', Shaffer in [20] analyses recordings of Chopin's Prelude Op. 28 No. 8 in F# minor, examining the structural tension and the variations in tempo and dynamics to decide whether a performance "conveys an insight into the musical meaning" [20, p.184]. The combination of melodic, harmonic and rhythmic processes identify structure, while operating on different levels, interacting within and perhaps across the levels. The results of the study show the use of a phrasing gesture where there is an acceleration and increase in dynamics into a musical unit (such as a phrase) and the respective 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 harmonisation in positions where ***ff*** and ***p*** markings appear, which helps emphasise the dynamic contrast.

[2] See also `http://mazurka.org.uk/info/excel/beat/` and `http://mazurka.org.uk/info/excel/dyn/gbdyn/` for beat and dynamic information on the Mazurka project.

In terms of dynamics, Thomas [18] 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 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.

With regard to Chopin's own performed dynamics, Chopin himself preferred pianos capable for depicting refined nuances rather than ones constructed based on providing acoustic sharpness and high intensity sounds [21]. Although markings such as ***ff*** and ***fff*** appear in his works, "all his contemporaries agree in reporting that his dynamics did not exceed the degree of *forte*, without however losing a single bit of shading" [22, 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 a "dominant fanfare" [18]. In the case of features related to timing, a characteristic of Chopin's music is that it draws inspiration from singing, which translates to a bel canto style of piano playing [23, p.216]. This style offers a strong sense of rubato by keeping a more steady rhythm with the left hand while freeing the other to push forward or hold back. Carl Mikuli, one of his pupils, "complimented Chopin's rubato for its naturalness and its 'unshakeable emotional logic'" ([24, p.91]).

3. SYNOPSIS OF THE DATASET

The MazurkaBL dataset [3] was created from 2000 selected recordings from the CHARM Mazurka dataset. The audio recordings cover a total of forty-four different Chopin Mazurkas. Table 3 shows the Chopin Mazurkas and the number of recordings of each Mazurka included in the dataset. MazurkaBL contains a table for each Mazurka in `.csv` (comma separated value) text format that includes the score beat positions (details in Section 4) in seconds per recording. Also, it contains a separate table for each Mazurka that includes the loudness information (details in Section 5) per score beat per recording. In both table formats, the rows represent the number of score beats and the columns represent the index of the recordings of the particular Mazurka. The recordings have been labeled using the same pianist-ID as in the Mazurka dataset. For each Mazurka another table has been created that includes the name of an expressive marking annotation found in the score and the number of score beat position where it is located. The score markings extracted are listed in Table 2.

We have included recordings in which the performer followed the repetitions designated in the score, and excluded ones that do not. We also excluded noisy recordings. By noisy recordings, we mean recordings with distortion artifacts (some old recordings) or live recordings with audience sounds that could not be removed. Following this cleanup process, the remaining Mazurkas and recordings were not included if the total number of recordings did not exceed twenty.

Dynamics
Markings: 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 2: Score markings having to do with dynamics and tempo or timing.

The recordings date from 1902 to the early 2000s. There is no information available on the score edition used by each performer. Tracing the actual score used in the preparation of each performance is an impossible task. Multiple editions of Chopin's Mazurkas exist; as noted in [25, p.56], "since most of [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." Even the (arguably) most widely used editions of Peters, Schirmer, and Augener bear the marks of later edits.

For the purposes of obtaining score-based tempo and dynamic markings, we used the Paderewski, Bronarski and Turczynski edition as it is one of the most popular and readily available editions. A comparison of dynamic markings across different score editions reveals a few differences. The most common reason for a difference between editions arises from a slight displacement in marking position of usually only one or two beats. Less commonly, if a location typically does not have any dynamic marking, an outlying edition may have one there, presented directly or inside parentheses. On a rare occasion, a marking that appears in most editions may be replaced by a completely different one in a maverick edition.

We encode each Chopin Mazurka score in XML format using Musescore [4] and we extract the location of each tempo and dynamic marking using the Music21 software package [26], the result of which was verified manually. 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" [25, p.53]. However this marking could not be included in our XML edition as it is not supported by the Music21 software.

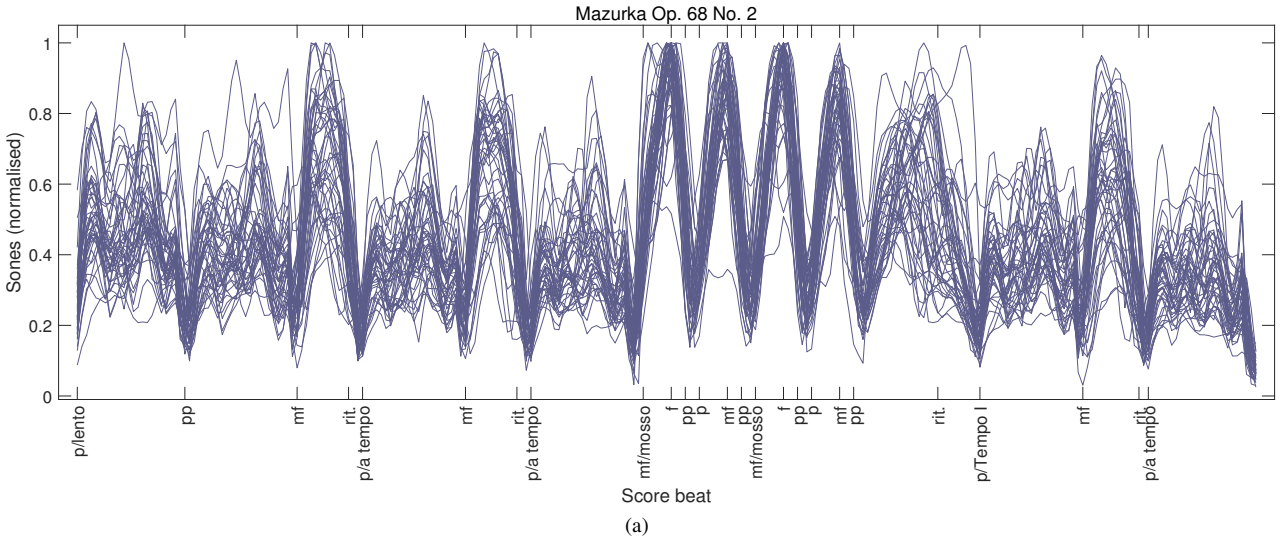
Figure 1 graphs the score-aligned loudness and inter-beat-interval (IBI) values for all 48 recordings of Mazurka Op. 68

[3] The dataset is publicly available and it can be found at: `https://github.com/katkost/MazurkaBL`. For copyright reasons, it does not include the audio files.

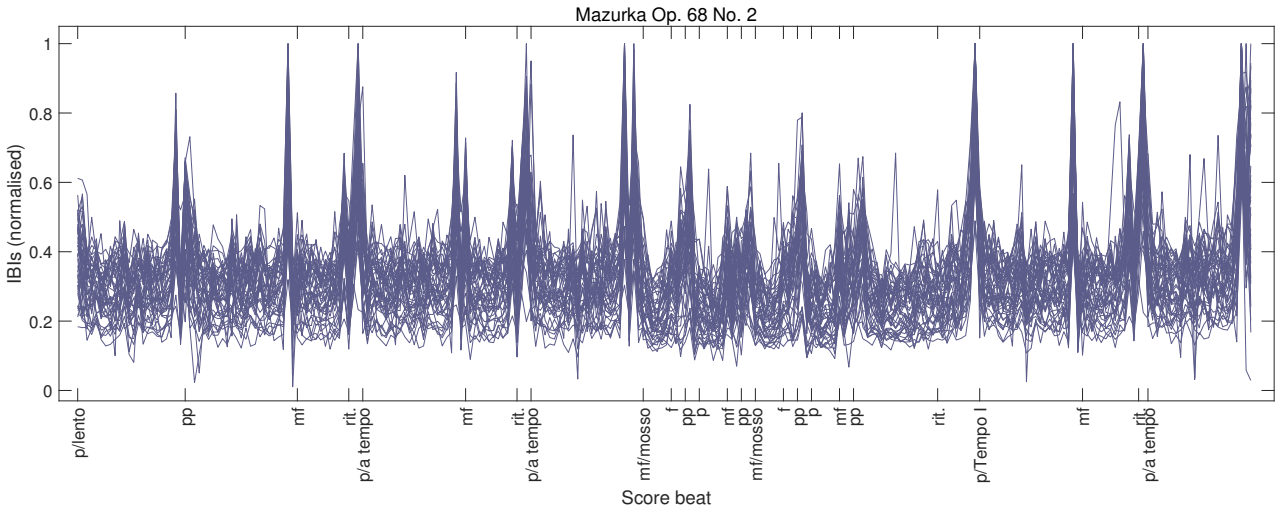
[4] `http://www.musescore.org`

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 3: Chopin Mazurkas used in this study and the number of recordings for each one. Mazurkas are indexed as "M<*opus*>-<*number*>."



(a)



(b)

Figure 1: Raw time-series representation of the MazurkaBL dataset for Mazurka Op. 68 No. 2. (a) shows a plot of the dynamic values in sones and (b) the Inter-Beat-Interval (IBI) per score beat for all 48 recordings, each presented as a separate curve. Expressive markings show on the x-axis at their corresponding locations in the score.

No. 2 from the MazurkaBL dataset. Each recording's loudness and IBI values were re-scaled to the range $[0, 1]$. Each recording is represented as an individual time-series curve of either the sone values for dynamics (a) or the IBI values for timing (b). By inspection, regions of agreement and parts where greater variation occurs are immediately apparent, as are the regions where certain outliers can be found. Similar interactive plots for all Mazurkas are avail-

able online [5] where it is possible to include or exclude particular curves separately, provide details of the exact values as well as the name of the pianist per curve, and zoom in to regions of interest.

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

4. SCORE BEAT INFORMATION

The position of score markings can be specified using the musical time axis of beats and measures. To study how a specific pianist realises a given marking in a performance, we need to locate its corresponding position in the recording in seconds. A common way to do this is to manually annotate the position of each musical beat in each available recording by tapping while listening to the music [1] and using specialised tools such as Sonic Visualiser [6] to check and correct the results. While manual annotations are typically quite reliable and accurate, creating them is highly time consuming and labour intensive. For example, for this research, the manual annotation and correction by inspecting the spectrogram of a single recording of Mazurka Op. 6 No. 2, which is approximately three minutes long, 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 rendition of a music piece, such synchronisation 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 annotated recording the *reference audio*. Its beat positions are transferred automatically to all the remaining recordings using a multiple recording alignment heuristic described in the next sections.

The approach to use a *reference* recording in an alignment procedure is not new—see, for example, [7] and [8]—and it has been shown to provide a significant stabilising effect on alignment accuracy. In this study, the multiple alignment heuristic calls the pairwise alignment algorithm by Ewert et al. [13], which applies Dynamic Time Warping (DTW) 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 an average onset error of 44 ms for piano recordings.

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. The choice of reference audio directly impacts the accuracy of the alignment. Intuitively, if an audio is an outlier, highly different from all the others in the set, it is a poor choice as a *reference* audio for accurate alignment to all other recordings. In order to determine the best choice of a *reference audio*, we created a ground truth dataset, which consisted of all forty-two recordings of Mazurka Op. 6 No. 2, each manually annotated with score beat positions. We computed the optimal *reference audio*, then determined its properties and designed a heuristic to automatically select this *reference audio* for other Mazurkas.

The goal of the multiple recording alignment heuristic is to optimise 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 4.1 we present an analysis of the *reference audio* properties, and in Section 4.2 we present a heuristic to detect the optimal *reference audio*.

4.1 Optimal *reference audio* choice

For this section, we use as ground truth our manual annotations of score-beat positions in all forty-two recordings of Mazurka Op. 6 No. 2. As a rule, in our manual annotations, we have chosen to follow the melody line so as to capture the lyricism of the rubato in the piano playing. Here, our goal is to determine the audio file (*reference audio*) that, when aligned and its beat annotations transferred to other audio recordings in the set, predicts most accurately the score-beat positions of the other recordings.

For this experiment, we removed silences in the beginnings and ends of all recordings by discarding any audio at the beginning and end in which the loudness value was < 0.002 sones (more information about the extraction of the sone values is given in Section 5). There are a total of 288 beats; no notes were struck on 11 of these beats. The alignment procedure calls the algorithm described in [13] for audio-to-audio alignment and the annotations (beat positions) from each candidate reference audio recording were transferred to all other recordings in a pairwise fashion.

Let n be the number of recordings. We thus obtain a total of $n \times (n-1)$ new sets of annotations generated from all the candidate reference audio files. To determine the audio that performed best in providing the alignments with the lowest beat prediction error, we compared the predicted beat positions to the annotated beat positions, the ground truth. The Jarque-Bera test showed that not all sets of prediction errors 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. For the sets of median values, we implemented the non-parametric Friedman test, where the small p-value ($p = 3.1546 \times 10^{-31}$) indicates that at least one column's sample median is significantly different from the others. The multiple comparison test shows the audio with the lowest median error value, which we interpret to be the best *reference audio*, to be Sztompka (1959), highlighted in bold in Figure 2, followed closely by the median error value of Kiepura's (1999) recording. Note that the y-axis is oriented so that the lowest values are at the top.

[5] https://goo.gl/xC5LcY

[6] http://sonicvisualiser.org/

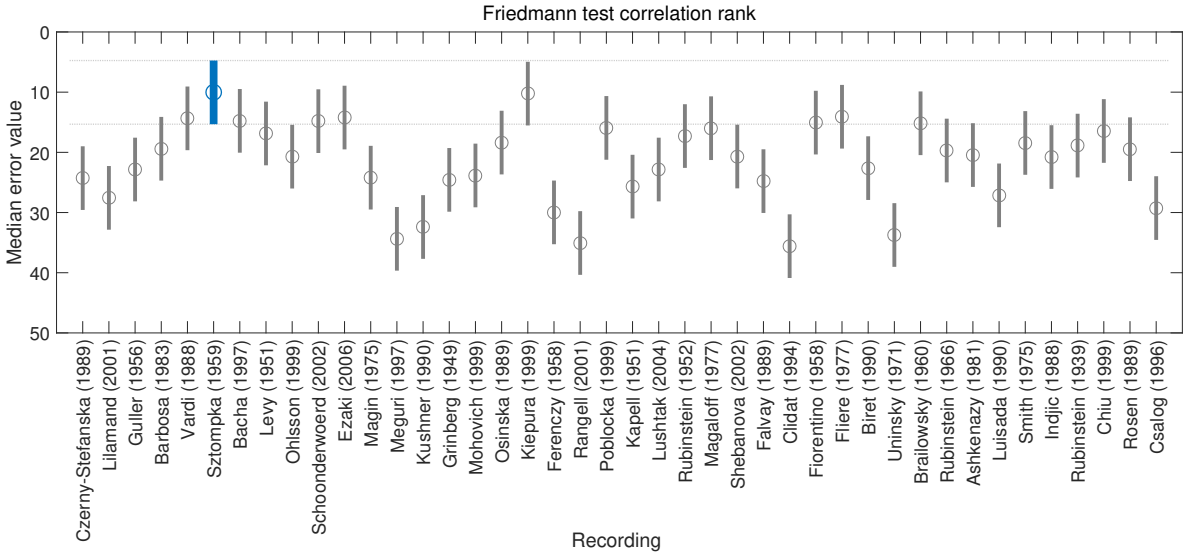


Figure 2: Error bars for the median beat prediction error when each of the 42 recordings of Mazurka Op. 6 No. 2 served, in turn, as the *reference audio*. x-axis shows pianist and recording year. The Friedmann correlation rank test showed Sztompka (1959) to be the recording with the lowest correlation rank and it is identified as the optimal *reference audio,* followed closely by Klepura (1999). Dotted horizontal lines mark the error bar limits of Sztompka (1959).

4.2 *Reference audio* selection heuristic

We first set up a fitness measure for a *reference audio* choice. The pairwise alignment algorithm [13] produces a match between two audio files, say i and j, using dynamic time warping. The alignment 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 from the two audio files. We compute the Euclidean distance between each pair of the dynamic time warped audio files as follows:

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

where $m \in N$ is the size of the vectors. In this way, each audio has a profile corresponding to its alignment to all other audio recordings, $\mathbf{d}_i = [d_{i,j}]$. 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 with the minimum average distance to other audio files, which, at the same time, does not exhibit extreme differences to more than two other audio recordings as measured by the norm distance. In this way, after exploring alternative values of outliers, a test on Mazurka Op. 6 No. 2 identified the same *reference audio* as that found using the exact method of Section 4.1. Mathematically, the problem of finding the *reference audio* can be expressed as one of solving the following problem:

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

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 evaluate the method using the ground truth created using Mazurka Op. 6 No. 2. For each candidate *reference audio*, we compared the *reference audio*-derived beat positions with the manually annotated beat positions for the remaining forty-one recordings of the Mazurka. The average error was found to be 30.7 ms.

4.3 Evaluation of score beat positions

Several approaches for evaluating alignment procedures exist—see, for example, [27] and references therein. For alignment procedures that do not follow a *reference* recording, such as in [28], the number of beats that are created may not be the same as the number of beats in the ground truth; thus, evaluation metrics different from that in this study may be employed.

For this study, in order to evaluate the beat positions of the MazurkaBL dataset, we compare them with the manual annotations provided by the Mazurka project. The Mazurka project provides publicly available manual annotations for 63 recordings of Mazurka Op. 17 No. 4, 64 recordings of Mazurka Op. 24 No. 2, 34 recordings of Mazurka Op. 30 No. 2, 95 recordings of Mazurka Op. 63 No. 3, and 50 recordings of Mazurka Op. 68 No. 3. The intersection of these with the recordings in MazurkaBL provides pairs of aligned positions for 48, 54, 30, 62, and 42 recordings of the respective Mazurkas mentioned for comparison. The results of the comparison in terms of mean and standard deviation of the beat difference (in milliseconds) are presented in Table 4.

The average beat difference between our manual beat annotations in the *reference audio* and the manual beat annotations of the corresponding recording from the Mazurka

Piece (# beats)	Diff mean (ms)	Diff std (ms)
M17-4 (395)	85	150
M24-2 (360)	69	119
M30-2 (193)	66	41
M63-3 (229)	71	61
M68-3 (180)	80	69

Table 4: Summary statistics for the difference between MazurkaBL (alignment-based beat transfer from manually-annotated reference audio) and Mazurka project (all manual) beat annotations.

Piece (# beats)	Diff mean (ms)	Diff std (ms)
M17-4 (395)	65.7	86.6
M24-2 (360)	64.2	21.9
M63-3 (229)	63.8	21.7
M68-3 (180)	57.5	33.7

Table 5: Summary statistics for the difference between the manual beat annotations of the MazurkaBL *reference audio* and the manual annotations of the corresponding recording from the Mazurka project.

Project is given in Table 5. Beat annotations of the Mazurka recording of Op. 30 No. 2 corresponding to the *reference audio* for that Mazurka in MazurkaBL was not available.

Table 4 shows that the beat annotations of the *reference audio* and of the annotations transferred from the *reference audio* for Mazurka Op. 17 No. 4 differ most from the corresponding manual annotations of the Mazurka project. The information provided in Table 5 shows how much manual annotations may differ from one annotator to the next; this may reflect a difference in the chosen criteria for marking beats.

5. LOUDNESS INFORMATION

In the MazurkaBL dataset, the loudness time series is extracted from each recording using the *ma_sone* function in Pampalk's Music Analysis toolbox [7]. The loudness time series is expressed in sones. There are two reasons we choose the sone values as a measure of dynamics. The sone scale is psycho-acoustically linear, so we can more readily and accurately normalise the values across different recorded environments. Furthermore, without having to apply any audio compression or modification, the sone calculations automatically pre-processes the audio intensity values based on the psychoacoustic concept of equal loudness curves.

The specific loudness sensation in sones per critical band is calculated by following the process explained in [29]. 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." [29] We also calculate the spectral masking effects, based on the research presented in [30]. Then we calculate the loudness in dB-SPL units, and from these values we calculate the equal loudness levels in phons via stored curves of equal loudness level. Next, from the phon values, we detect the values in sones, following the calculation described in [31], according to which the loudness level S in sones can be calculated from the loudness levels L in phons using the formula:

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

the rationale being that "in this way the threshold of hearing and the nonlinear and frequency-dependent response of the ear to intensity differences are taken into account." [31]

The sone values are smoothed by local regression using a weighted linear least squares and a 2nd degree polynomial model (the "loess" method of MATLAB's *smooth* function [8]). The loudness time series for each recording is normalised to $[0, 1]$ by dividing the values of a recording by the maximum loudness value of that particular recording.

6. RECENT APPLICATIONS OF MazurkaBL

This section presents some studies that have used the MazurkaBL dataset and briefly describes their findings.

The set of markings {***pp***, ***p***, ***mf***, ***f***, ***ff***} were studied in [32], which explored the absolute meanings of the dynamic markings change as a function of the intended (score defined) and projected (recorded) dynamic levels, and that of the surrounding musical context. The analysis revealed a (sometimes) wide range of realisations of the same dynamic markings throughout a recording of a piece. Reasons for this counter-intuitive 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. The analysis showed that, transitions from a louder to a softer marking, between markings of high intensity, and between markings of high contrast, tend to be more consistent. For markings that appear in the score more than once, most often than not, there was significant variation in the ways the markings were interpreted.

Offering a different perspective, [33] addressed the question of whether changes in dynamics, as automatically identified by statistical change-point algorithms, corresponded to dynamic markings. The assumption was that a dynamic marking indicated a point of change, and thus served as ground truth on which to evaluate the change-point algorithms. The results show that significant dynamic score markings do indeed correspond to change points, and evidence suggests that change points in score positions without dynamic markings serve to bring prominence to structurally salient events or to events the introduce a change in tempo.

A subset of the MazurkaBL dataset was used in [34] to investigate the bi-directional mapping between dynamic

[7] www.pampalk.at/ma/documentation.html

[8] http://uk.mathworks.com/help/curvefit/smooth.html?refresh=true

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 modeling 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.

Finally, [35] describes another study that applied machine learning techniques to a subset of the MazurkaBL dataset. The goal of this study was to examine tempo-loudness interactions at specific score markings over a set of recordings, and to investigate how including information about 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. As before, the predictions failed when the model was trained on loudness or tempo change information from recordings of multiple Mazurkas performed by the same pianist.

7. FUTURE DIRECTIONS

We have presented MazurkaBL, a new dataset for expressive music performance studies, comprising of 2000 beat-aligned recordings of forty-four Chopin Mazurkas overlaid with loudness information and score markings pertaining to tempo/timing and dynamics.

We provide material to quantitatively investigate what the score notation represents from a performer's perspective. Future tools providing different ways of visualising the dataset can bring insights that lead to a new notation system that represents changes in expression. Once important changes have been identified, symbols can be chosen to signify these changes and the representation can serve as a tool for comparing and analysing performances.

Much research has focused on proposing and establishing the relationship between dynamics and timing variations (see, for example [36] and references therein.) These studies range from establishing simple rules such as louder passages tend to be faster [37] to audio-synchronised animations of expressive parameters in tempo-loudness space [38]. Musical timing and amplitude has also been linked to subjective ratings of emotionality, for example in [39]. Timing and loudness variations in a music performance form critical cues for the identification of core music features such as phrase boundaries—see, for example, [40], [41], and [42]). The MazurkaBL dataset opens up many more avenues for explorations of this kind, and on a much larger scale.

Some future directions include expanding the list of score markings such as pedaling, and including audio features such as timbre. Further analytical studies could investigate gradual changes such as the analysis of *crescendo* or *diminuendo*. Also the same approach of large-scale annotation of score-beat information can be applied to other audio recordings of music by other composers, for other instruments, and of other genres.

Acknowledgments

This research was funded in part by a Queen Mary University of London Principal's PhD studentship. The authors are also grateful to Sebastian Ewert for providing the code for the audio-to-audio alignment algorithm.

8. REFERENCES

[1] C. Sapp, "Comparative analysis of multiple musical performances," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Vienna, Austria, 2007, pp. 497–500.

[2] S. Flossmann, W. Goebl, M. Grachten, B. Niedermayer, and G. Widmer, "The Magaloff project: An interim report," *Journal of New Music Research*, vol. 39, no. 4, pp. 363–377, 2010.

[3] M. Müller, V. Konz, W. Bogler, and V. Arifi-Müller, "Saarland Music Data (SMD)," in *Late-Breaking and Demo Session of the International Society for Music Information Retrieval Conference (ISMIR)*, Miami, USA, 2011.

[4] M. Marchini, R. Ramirez, P. Papiotis, and E. Maestre, "The Sense of Ensemble: A Machine Learning Approach to Expressive Performance Modelling in String Quartets," *Journal of New Music Research*, vol. 43, no. 3, pp. 303–317, 2014.

[5] P. Papiotis, "A computational approach to studying interdependence in string quartet performance," Ph.D. thesis, Univesitat Pompeu Fabra, 2016.

[6] C. Sapp, "Hybrid numeric/rank similarity metrics for musical performance analysis." in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Philadelphia, USA, 2008, pp. 501–506.

[7] A. Arzt and G. Widmer, "Real-time music tracking using multiple performances as a reference," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Malaga, Spain, 2015, pp. 357–363.

[8] S. Wang, S. Ewert, and S. Dixon, "Robust and efficient joint alignment of multiple musical performances," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2132–2145, 2016.

[9] M. Müller, T. Prätzlich, and J. Driedger, "A cross-version approach for stabilizing tempo-based novelty detection." in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Porto, Portugal, 2012, pp. 427–432.

[10] A. Benetti Jr., "Expressivity and musical performance: Practice strategies for pianists," in *Performance Studies Network International Conference*, Cambridge, UK, 2013. [Online]. Available: http://www.cmpcp.ac.uk/wp-content/uploads/2015/11/PSN2013_Benetti.pdf

[11] J. Rink, *Musical Performance: A Guide to Understanding*. Cambridge University Press, 2002.

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

[13] S. Ewert, M. Müller, and P. Grosche, "High resolution audio synchronization using chroma onset features," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei, Taiwan, 2009, pp. 1869–1872.

[14] T. Rutherford-Johnson, M. Kennedy, and J. Kennedy, *The Oxford Dictionary of Music*. OUP Oxford, 2012.

[15] S. Barrett, "Reflections on music writing: Coming to terms with gain and loss in early medieval latin song," in *Vom Preis des Fortschritts: Gewinn Und Verlust in der Musikgeschichte*, A. Haug and A. Dorschel, Eds. Universal Edition, 2008.

[16] N. Cook, E. Clarke, L.-W. Daniel, and J. Rink, *The Cambridge Companion to Recorded Music*. Cambridge University Press, 2009.

[17] J. Sadie, *Companion to Baroque Music*. University of California Press, 1998.

[18] A. Thomas, "Beyond the dance," in *The Cambridge Companion to Chopin*, J. Samson, Ed. Cambridge University Press, 1994.

[19] J.-J. Eigeldinger, "Placing chopin: Reflections on a compositional aesthetic," in *Chopin Studies 2*, J. Rink and J. Samson, Eds. Cambridge University Press, 1994, vol. 2.

[20] L. H. Shaffer, "Performing the F# minor prelude op. 28 no. 8," in *Chopin Studies 2*, J. Rink and J. Samson, Eds. Cambridge University Press, 1994, vol. 2.

[21] J. Methuen-Campbell, "Chopin in performance," in *The Cambridge Companion to Chopin*, J. Samson, Ed. Cambridge University Press, 1994.

[22] A. Einstein, *Music in the Romantic Era: A History of Musical Thought in the 19^{th} Century*. New York : W.W. Norton, 1975.

[23] J. Rink, "Authentic chopin: History, analysis and intuition in performance," in *Chopin Studies 2*, J. Rink and J. Samson, Eds. Cambridge University Press, 1994, vol. 2.

[24] H. C. Khoo, "Playing with dynamics in the music of chopin," Ph.D. dissertation, Royal Holloway, University of London, 2007.

[25] E. Bailie, *Chopin: A Graded Practical Guide*, ser. Pianist's Repertoire. Kahn & Averill, 1998.

[26] M. S. Cuthbert and C. Ariza, "music21: A Toolkit for Computer-Aided Musicology and Symbolic Music Data," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Utrecht, Netherlands, 2010, pp. 637–642.

[27] T. Prätzlich and M. Müller, "Triple-based analysis of music alignments without the need of ground-truth annotations," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 2016, pp. 266–270.

[28] P. Grosche, M. Müller, and C. S. Sapp, "What makes beat tracking difficult? a case study on chopin mazurkas." in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Utrecht, Netherlands, 2010, pp. 649–654.

[29] 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, 2002, pp. 570–579.

[30] M. R. Schroeder, B. S. Atal, and J. Hall, "Optimizing digital speech coders by exploiting masking properties of the human ear," *The Journal of the Acoustical Society of America*, vol. 66, no. 6, pp. 1647–1652, 1979.

[31] R. A. W. Bladon and B. Lindblom, "Modeling the judgment of vowel quality differences," *The Journal of the Acoustical Society of America*, vol. 69, no. 5, pp. 1414–1422, 1981.

[32] K. Kosta, O. F. Bandtlow, and E. Chew, "Dynamics and relativity: Practical implications of dynamic markings in the score," *Journal of New Music Research*, 2018 (to appear).

[33] ——, "A change-point approach towards representing musical dynamics," in *Mathematics and Computation in Music*. London, UK: Springer Lecture Notes in Artificial Intelligence 9110, 2015, pp. 179–184.

[34] K. Kosta, R. Ramirez, O. F. Bandtlow, and E. Chew, "Mapping between dynamic markings and performed loudness: A machine learning approach," *Journal of Mathematics and Music*, vol. 10, no. 2, pp. 149–172, 2016.

[35] C. Vaquero, I. Titov, and H. Honing, "What score markings can say of the synergy between expressive timing and loudness," in *Abstracts from European Society for Cognitive Sciences Of Music Conference*, Ghent, Belgium, 2017.

[36] G. Widmer and W. Goebl, "Computational Models of Expressive Music Performance: The State of the Art," *Journal of New Music Research*, vol. 33, no. 3, pp. 203–216, 2004.

[37] B. H. Repp, "The dynamics of expressive piano performance: Schumann's Träumerei revisited," *The Journal of the Acoustical Society of America*, vol. 100, no. 1, pp. 641–650, 1996.

[38] S. Dixon, W. Goebl, and G. Widmer, "The Performance Worm: Real Time Visualisation of Expression based on Langner's Tempo-Loudness Animation," in *Proceedings of the International computer music conference (ICMC)*, Gothenburg, Sweden, 2002.

[39] A. Bhatara, A. K. Tirovolas, L. M. Duan, B. Levy, and D. J. Levitin, "Perception of emotional expression in musical performance," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 37, no. 3, 2011.

[40] C.-H. Chuan and E. Chew, "A Dynamic Programming Approach to the Extraction of Phrase Boundaries from Tempo Variations in Expressive Performances," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Vienna, Austria, 2007, pp. 305–308.

[41] E. Cheng and E. 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*, vol. 37, no. 4, pp. 325–338, 2008.

[42] D. Stowell and E. Chew, "Maximum a Posteriori Estimation of Piecewise Arcs in Tempo Time-Series," in *From Sounds to Music and Emotions – International Symposium on Computer Music Modeling and Retrieval (CMMR)*. London, UK: Springer LNCS 7900, 2012, pp. 387–399.