TISMIR

Revisiting Expressive Timing in Piano Performance at Scale: A Distant Listening, Multi-Corpus Study

RESEARCH ARTICLE

]u[ubiquity press

PATRICIA HU
HUAN ZHANG
SILVAN DAVID PETER
CARLOS EDUARDO CANCINO-CHACÓN
SIMON DIXON
GERHARD WIDMER

*Author affiliations can be found in the back matter of this article

ABSTRACT

Empirical expressive performance studies are typically performed on small, curated datasets. In this study, we present a large-scale analysis of expressive timing in classical piano performance using a 'distant listening' approach across six MIDI datasets totaling over 600 hours of music. We revisit three core aspects of expressive timing drawn from empirical performance research: (1) the relational invariance of timing patterns with respect to tempo; (2) the influence of score markings, register, and interval consonance on articulation; and (3) the role of asynchrony as a means of expressive accentuation. Our findings indicate that timing patterns do not scale proportionally with tempo; however, weak interactions between tempo and metrical position suggest that both factors jointly contribute to shaping timing patterns. Articulation analysis confirms associations with score markings and register, although the relationship with interval consonance is less consistent. Melody lead tendencies are consistently observed across datasets, though our findings indicate that those occurring on the beat level may primarily reflect dynamic accentuation. Overall, our results provide empirical validation for certain established theories of expressive performance while challenging others, illustrating how large-scale computational analysis can complement traditional controlled studies by offering broader insights into performance timing.

CORRESPONDING AUTHOR:

Patricia Hu
Institute of Computational Perception, Johannes Kepler University, Linz, Austria
patricia.hu@jku.at

KEYWORDS:

distant listening, expressive timing, piano performance

TO CITE THIS ARTICLE:

Hu, P., Zhang, H., Peter, S.D., Cancino-Chacón, C.E., Dixon, S., & Widmer, G. (2026). Revisiting Expressive Timing in Piano Performance at Scale: A Distant Listening, Multi-Corpus Study. *Transactions of the International Society for Music Information Retrieval*, 9(1), 329–346. DOI: https://doi.org/10.5334/tismir.317

1 INTRODUCTION

Music performance analysis involves understanding and interpreting the nuances of expressive musical renditions. This analytical process allows researchers, musicians, and music enthusiasts to study the details of a performance, uncovering stylistic choices and interpretations that contribute to the richness of musical expression. In the past, most work in expressive performance analysis using computational methods has been exploratory in nature (Lerch et al., 2021), conducted on relatively small-scale performance data measured on special MIDI-controllable instruments (Desain and Honing, 1994; Goebl et al., 2010; Palmer, 1989; Repp, 1990), or (semi-)automatically extracted features from audio recordings (Kosta et al., 2015; Sapp, 2008).

Recent releases of high-quality performance datasets (Hu and Widmer, 2023; Peter et al., 2023) and advances in transcription algorithms and the introduction of large-scale transcribed performance data (Kong et al., 2021; Zhang et al., 2022) have enabled new possibilities for studying expression in music performances. In this article, we propose a 'distant listening' approach to expressive music performance analysis using both transcribed audio data and recorded MIDI data. In particular, we will revisit a set of hypotheses related to expressive timing that have been proposed in the literature and reassess their validity at scale using a 'distant listening' approach and data from (predominantly) real-world (as opposed to laboratory) performances. To apply such an approach to the domain of expressive music performance research, we look at six MIDI piano performance datasets note-aligned to their MusicXML scores, covering a wide range of piano solo work from the Common Practice Period and totaling over 600 hours of music and 14M performed notes. In this way, we wish to provide stronger empirical support, on 'ecologically valid' data, for some selected hypotheses from performance research—or identify possible problems with these hypotheses in the context of 'real-world' music performance data.

The motivation for this article is twofold: (1) to revisit previously observed hypotheses on expressive timing at scale, with the aim of drawing more general conclusions about the performance patterns of expressive timing and its various components, and (2) to demonstrate the potential of large-scale distant listening analysis as a complement to more detailed and controlled musicological investigations. We provide all analysis scripts at: https://github.com/huispaty/expressive_timing.[1]

Building on preliminary observations from empirical research on expressive performance, particularly in relation to expressive timing, we aim to explore three research questions previously raised in seminal quantitative performance studies (Desain and Honing, 1994; Palmer, 1989, 1996b; Repp, 1990, 1994). Leveraging the size of analyzed datasets, we seek to examine the following three questions: (1) *Are timing patterns relationally invariant to tempo?*; (2) *How do score markings, interval consonances, and octave register influence articulation?*; and (3) *Do note asynchrony values reflect melody lead and melody lag patterns?* We chose these three questions to systematically investigate how tempo, articulation, and asynchrony each contribute to expressive timing in classical piano performance, allowing us to better understand the distinct and combined roles these parameters play in shaping musical expression.

2 RELATED WORK

We structure this related work section around our three research questions, dedicating one subsection to each. Each subsection covers the literature relevant to its corresponding experiment.

2.1 RELATIONAL INVARIANCE

In the first experiment, we ask the question of whether note inter-onset intervals (IOIs) (termed 'timing patterns' or 'timing profiles' by Repp (1990, 1992a, 1994) and Desain and Honing (1994)) scale proportionally with global tempo or, alternatively, if a change in global tempo necessitates a change in timing patterns. This question can be formalized as a hypothesis test. The *null hypothesis (H0)* corresponds to the *relational invariance hypothesis*, which assumes that variations in timing patterns remain relationally invariant under moderate changes in global tempo (Repp, 1994). In contrast, the *alternative hypothesis* (HA) is the *tempo-specific timing hypothesis*, which states that timing patterns are intrinsically linked to tempo such that different (phrase-)structural levels become salient at different global tempi (Honing, 2007).

Evidence has been found both in favor of the relational invariance hypothesis (H0, in the following) and its alternative tempo-specific timing hypothesis (HA). Repp (1994) compared recordings by two pianists playing Schumann's 'Träumerei' three times at three different tempi on a computer-controlled piano and found IOIs to scale roughly in proportion with global tempo, concluding relational invariance to hold up. Following the same procedure, Desain and Honing (1994) studied recordings of the theme and first variation of the Beethoven variations WoO 70, played by two pianists again at three different tempi, and found significant interactions between global tempo and IOIs, concluding that timing patterns do not scale proportionally with global tempo. Contrasting their results with the then-recent results by Repp (1994), the underlying musical structure, the range of tempi considered, and finally methodological issues related to IOI measurement were listed as potential factors having influenced the analysis.

2.2 ARTICULATION: SCORE MARKINGS, INTERVAL CONSONANCE, AND REGISTER

Prior empirical performance studies have proposed several musical factors that systematically influence performed articulation. Bresin and Battel (2000) investigated articulation strategies in the second movement of Mozart's Piano Sonata KV.545, comparing passages marked *legato* and *staccato*. They showed that articulation markings are reflected in measurable differences in the ratio between performed note duration and IOI, providing evidence for a *score-marking hypothesis*.

Register has also been proposed as an influencing factor. Due to the acoustic properties of the piano, higher-pitched tones decay faster than lower ones (Benade, 1977). Based on this, Repp (1995) observed that pianists tend to adapt their articulation across registers to achieve perceptually consistent continuity, motivating a *register hypothesis*.

A further proposal concerns interval consonance. Repp (1996b) suggested that dissonant intervals may be articulated with less overlap than consonant ones, forming an *interval consonance hypothesis*. However, this hypothesis has so far only been supported by controlled listening experiments and not yet confirmed on performance data.

2.3 CHORD ASYNCHRONY

In expressive piano performance, temporal asynchronies between voices, i.e., where notes notated at the same score onset are played non-simultaneously, either within one or across both hands, are a well-documented phenomenon. These asynchronies can occur in various forms, most prominently as *melody lead* (also known as *bass lag*), where the melody voice precedes accompanying voices by 20–30 ms, or *bass anticipation* (also known as *bass lead*), where the bass note occurs approximately 50 ms earlier than other chord notes (Goebl, 2001; Goebl et al., 2010; Palmer, 1996b).

A central question concerns the underlying mechanism of these asynchronies: do they arise from intentional temporal accentuation or from dynamic emphasis? This alternative *velocity hypothesis* proposes that observed onset differences may be mechanical consequences of playing melody notes louder, rather than deliberate timing choices. While early studies found weak or inconsistent correlations between MIDI velocity and onset timing (Palmer, 1989), later work demonstrated a strong relationship between voice-leading asynchronies, and dynamic differentiation (Repp, 1996a). Goebl (2001) further showed that melody lead disappears almost entirely when measuring finger-key contact times, implying that dynamic accentuation, rather than timing intention, may drive observed onset differences.

Beyond melody lead, research has documented systematic variations in asynchrony patterns according to musical structure. Between-hand asynchronies vary with texture, tempo, and metrical position: Performances of pieces of predominantly chordal textures show less asynchrony than those with sustained melodic lines, and faster tempi tend to reduce both the magnitude and variability of asynchrony (Goebl et al., 2010). Likewise, both melody lead and bass anticipation have been observed to be particularly pronounced on metrically strong positions (Palmer, 1996b), followed by other on-beat and least frequently off-beat metrical positions (Goebl et al., 2010).

3 DISTANT LISTENING

Using the term *distant listening*, we connect to the concept of distant reading introduced by Moretti in the field of literary analysis (Moretti, 2013). Distant reading refers to a hermeneutic technique of analyzing global patterns and trends in large amounts of textual data (as opposed to close interpretation of individual instances—so-called close reading). While close reading entails the thorough observation of central themes and their evolution by closely reading and following the structure of the source text, distant reading abstracts itself from the actual textual content of a given work and instead aims to find global patterns and features from single or multiple text(s) (Jänicke et al., 2015).

Distant reading is used for various purposes within the digital humanities, e.g., to identify patterns, trends, and recurring features at a scale beyond the reception capacity of an individual; to map large-scale literary history; to investigate the non-canonical, under- or unread works; to generate hypotheses for close reading or, vice versa, to validate and generalize hypotheses from close reading; or to provide an abstract, macroscopic view of a given corpus.

For our work, we operationalize distant listening as a method that inherits two of these purposes: the analysis of musical performances at scale and the validation of established close reading hypotheses. On the other hand, we do work with canonical data—at least in terms of repertoire—and omit exploratory, pattern finding, longitudinal, or macroscopic approaches.

In both digital literature studies and digital musicology, the underlying source data for distant analysis is of textual or notated form—that is, *encoded*—which lends itself to analysis through a wealth of standardized methods. Digital research of such data can often lead to practical tools such as visualizations or interfaces that provide an abstract view of a given corpus by illustrating the hierarchies or relationships between textual elements (Piez, 2005). Such tools may allow for filtering, zooming, grouping, and direct information retrieval and, as such, support close reading in flexible ways (Cayless, 2005). Large-scale analytically annotated corpora of musical scores (Hentschel et al., 2025; Pugin, 2015) likewise can enable similar exploratory and educational tools (Gotham et al., 2023).

All of the above presupposes that the phenomenon under study exists in an encoded, stable form. Musical performance, on the other hand, is largely ephemeral in its nature. Though we have the means to capture and reproduce (mostly: playback) a certain performance, there is no agreed-upon encoding of performances, and neither established taxonomies nor means to apply standardized methods for analysis exist. In this work, we do not seek to address this gap. Instead, we aim at a scaling of quantitative music performance research using bespoke processing methods to address existing hypotheses. By examining performance patterns at this scale, we hope to lend empirical support to established claims about broader principles of musical expression.

The difference in scale necessitates a crucial methodological shift. While most prior research examined a certain performance phenomenon in a (mostly) within-performer, within-piece design, we follow a between-performer, multi-corpus approach, where performances stem from different individuals and performance contexts. Although this introduces greater variability than controlled laboratory studies, it better captures the diversity of real-world musical practice and enables the study of performance phenomena in ecologically valid, real-world performance contexts.

This shift away from laboratory settings toward large-scale, real-world data raises legitimate methodological concerns regarding reliability, control, and sampling—concerns similar to those debated in music-perception research on web-based experiments (Honing and Ladinig, 2008; Honing and Reips, 2008; Kendall, 2008). In laboratory studies, researchers carefully select and recruit participants; in distant listening, this controlled sampling is replaced by the availability and characteristics of existing data. Likewise, the controlled conditions of laboratory experiments give way to the inherent variability of real-world recordings, which differ in recording quality and performance context. However, as argued in the web-based experiments debate, we believe the potential of distant listening for empirical performance studies—with regards to scale, versatility, and ecological validity—outweighs these concerns, provided they are appropriately addressed through proper experimental design and critical interpretation of the results.

Proper experimental design and sampling strategies can be ensured by carefully aligning methodological choices with the specific research questions at hand (Honing and Reips, 2008). The more substantial challenge lies in data reliability. Large-scale approaches using real-world data necessarily introduce greater variance across multiple dimensions: recordings vary in acoustic quality; performances differ in their interpretive goals and contexts; and, most importantly, large-scale approaches rely on large-scale data predominantly generated through automatic annotation. Rather than viewing the increased variance introduced by real-world data as noise to be eliminated, we argue it should be explicitly modeled and accounted for in the analysis.

Given these trade-offs between control and ecological validity, we do not propose distant listening as a replacement for laboratory studies but instead as a complementary approach. Just as distant reading in literary studies has revealed large-scale patterns in textual corpora that close reading alone could not detect, we propose distant listening to find more global, abstract patterns in musical expression that may remain unapproachable in small-scale investigations. As a first step in this direction, we revisit established hypotheses from the empirical performance research literature concerning expressive timing and reassess their validity across multiple performed corpora.

4 ANALYZED CORPORA

4.1 DATA FORMAT

To facilitate the analysis of multiple datasets, we used a standardized pipeline based on scores in MusicXML format and piano performances as MIDI files. Note-level alignments between score and performance stem either from manual curation or from state-of-the-art symbolic alignment algorithms (Peter, 2023; Peter et al., 2023) and are encoded in the match file format (Foscarin et al., 2022).

4.2 CORPORA

The corpora vary in size, performance skill level, repertoire, and measurement precision. Three stem from recordings by professional pianists, each focusing on the oeuvre of a specific composer. The remaining three datasets encompass a broader range of Western Common Practice solo piano music, including compositions from the Baroque to the late Romantic era. See Table 1 for an overview.

The *Magaloff/Chopin* corpus (Flossmann et al., 2010) is a collection of on-stage recordings of the complete works of F. Chopin for solo piano by world-class pianist Nikita Magaloff on a Bösendorfer SE computer-monitored grand piano. The recordings were converted from proprietary Bösendorfer SE to standard MIDI format and subsequently note-level–aligned with their corresponding scores. The complete corpus consists of 155 performances and 336,581 played notes and represents approximately 10 hours of music.

The *Zeilinger/Beethoven* corpus (Cancino-Chacón et al., 2017) consists of performances of 31 movements from nine Piano Sonatas by L. v. Beethoven performed by the Austrian concert pianist Clemens Zeilinger and subsequently aligned using the same curation protocol as *Magaloff/Chopin*. This dataset comprises over 70,000 performed notes, amounting to just over 3 hours of music.

The *Batik-plays-Mozart* corpus (Hu and Widmer, 2023) contains 36 movements from 12 Piano Sonatas by W. A. Mozart performed by Viennese concert pianist Roland Batik and subsequently note-level–aligned to a standard

Dataset	Pieces	Performances	Duration	MIDI	Repertoire
Magaloff/Chopin (Flossmann et al., 2010)	155	155	8 h 17 m	recorded	Essentially all of Chopin's solo piano work
Zeilinger/Beethoven (Cancino-Chacón et al., 2017)	30	30	2 h 40 m	recorded	9 piano sonatas: Ops. 2, 13, 14, 26, and 27 (early); Ops. 31 and 53 (middle); and Ops. 109 and 110 (late)
Batik-plays-Mozart (Hu and Widmer, 2023)	36	36	3 h 46 m	recorded	12 piano sonatas: KV279–284, KV330–333, KV457, and KV533
Vienna4×22 (Goebl, 1999)	4	88	2 h 18 m	recorded	Excerpts from 4 pieces by Chopin, Mozart, and Schubert
(n)ASAP (Peter et al., 2023)	235	1067	94 h 30 m	recorded	Common Practice Period piano solo work by 15 composers
ATEPP (aligned subset) (Zhang et al., 2022)	320	5094	524 h 48 m	transcribed	Solo piano work by 25 composers, ranging from Baroque to the Modern era
Total	780	6330	636 h 10 m		

Table 1 Overview of datasets used for our distant listening approach, totaling over 600 hours of music.

edition (the New Mozart Edition) of the score, which is further linked to musicological annotations (Hentschel et al., 2021). The corpus contains approximately 102,400 performed notes, totaling nearly 4 hours of music.

The *Vienna 4×22* dataset (Goebl, 1999) contains 22 different performances by piano students and professors from the University of Music and Performing Arts Vienna, each of excerpts of four different pieces from the Classical–Romantic repertoire: the Etude Op. 10, No. 3 and the Ballade Op. 38 by F. Chopin, the first movement of the Sonata KV331 by W. A. Mozart, and the Deutscher Tanz D. 783 No. 15 by F. Schubert. The dataset contains close to 44,000 performed notes.

The *(n)ASAP* dataset (Peter et al., 2023) is the largest fully note-aligned collection of solo piano performances, originating from the Yamaha Piano-e-competition and comprising over 7M annotated notes and nearly 100 hours of music by 15 composers from the Common Practice Period.

ATEPP (Zhang et al., 2022) is a large-scale dataset of transcribed expressive piano performances by virtuoso pianists. The dataset contains 5094 aligned performances (approx. 525 hours) by 49 pianists and 25 composers ranging from the Baroque to the Modern period. The alignments were obtained using the current state-of-the-art offline symbolic alignment method (99% average test F-score) from Peter (2023).

5 EXPERIMENT 1: ARE TIMING PATTERNS RELATIONALLY INVARIANT TO GLOBAL TEMPO?

5.1 HYPOTHESIS

We tested **H0 (relational invariance)** against **HA (tempo-specific timing)** by examining whether timing patterns scale proportionally with tempo changes (supporting H0) or instead change systematically in a non-proportional manner (supporting HA). For a detailed discussion on the related debate, see Section 2.1.

5.2 METHOD

As pointed out by Desain and Honing (1994), the measurement and calculation of note IOIs is not trivial, as parallel (joint note onset and chord) and ornamental (grace note) structures in particular can substantially bias the results toward or against the relational invariance hypothesis.

We combined the measurement procedures proposed by Repp (1994) and Desain and Honing (1994) and computed IOIs as follows: we excluded grace notes and selected, for parallel structures, either the onset of the melody note where voicing information is available or the earliest onset where it is not. We then defined the IOI as the physical time difference between two successive onsets, which we normalized by the nominal score position distance between the corresponding note pair (e.g., a quarter note if the onsets are a quarter note value–equivalent apart, an eighth note if they are an eighth note apart, etc.). To standardize proportional tempo changes across varying average tempi and to better reflect perceptual sensitivity to tempo and timing variations, we lastly applied logarithmic transformation. We chose this hybrid approach because the two earlier studies reached conflicting conclusions, and we argue that the combination of score normalization and logarithmic scaling, without excluding notes immediately surrounding grace notes, allows us to analyze the timing–tempo relationship without bias toward invariance.

To answer the question of whether timing patterns scale proportionally with tempo, we performed two analyses. First, we compare within-tempo group correlations to slow–fast between-group correlations. Following

Desain and Honing (1994), we avoided averaging IOIs across performances before correlation computation. Instead, for each piece, we computed all pairwise correlations between individual performances, separately for within-tempo group pairs (e.g., slow vs. slow) and between-tempo group pairs (e.g., slow vs. fast). We then averaged these pairwise correlations within each comparison type and report the mean and standard deviation across pieces. If relational invariance (H0) holds, between-group correlation should be at similar levels as within-group correlation. In line with the reference study, we tested for significance using the Fisher Z-transform with $\alpha = 0.05$.

Second, to test whether timing scales multiplicatively with tempo (H0) or exhibits tempo-specific patterns (HA), we analyzed log-transformed IOI sequences using separate analyses of variance (ANOVA) per piece, with tempo group (slow, medium, fast) and metrical strength (offbeat, beat, downbeat) as fixed factors. The log-transformation converts multiplicative relationships to additive ones, allowing us to distinguish the hypotheses: H0 would predict only main effects (parallel patterns across tempos), while HA would predict a significant tempo × metrical strength interaction (non-parallel patterns).[2]

Note that our analysis compares timing pattern variations across performances by different performers, unlike the within-performer timing variation approaches of Desain and Honing (1994) and Repp (1990). We chose to apply the same method nonetheless, as empirical evidence suggests that systematic effects of musical structure (e.g., phrase boundaries, harmonic tension) tend to emerge consistently across performers—that is, there seems to be a shared cognitive strategy for timing variations related to the underlying musical structure (Desain and Honing, 1989; Langner and Goebl, 2003; Repp, 1992b).

5.3 RESULTS

For our analysis, we considered all pieces for which there are at least three performances at three distinct estimated global tempi. We used a crude estimate of the global tempo of a performance in beats per minute (BPM) following Desain and Honing (1994), using the duration of a piece divided by the number of beats it contains. We then sorted the performances and split them into three groups (slow/medium/fast) based on their tertiles.[3] Figure 1 illustrates the distribution of tempi across 1326 performances of 80 Chopin pieces in the ATEPP dataset (where each performance's tempo is normalized by the median tempo of all performances of the same piece), along with the resulting tertile-based grouping.

5.3.1 Correlational evidence

Table 2 presents an excerpt of the correlational evidence—that is, the comparison of within-tempo group to between-group correlations for performances of compositions by Chopin and Beethoven in the three datasets studied.

We chose Chopin and Beethoven for their contrasting approaches to rhythm and tempo: in performance, Beethoven's music, while rhythmically inventive and dramatic, generally calls for a more disciplined and structured approach to tempo, with limited use of rubato.

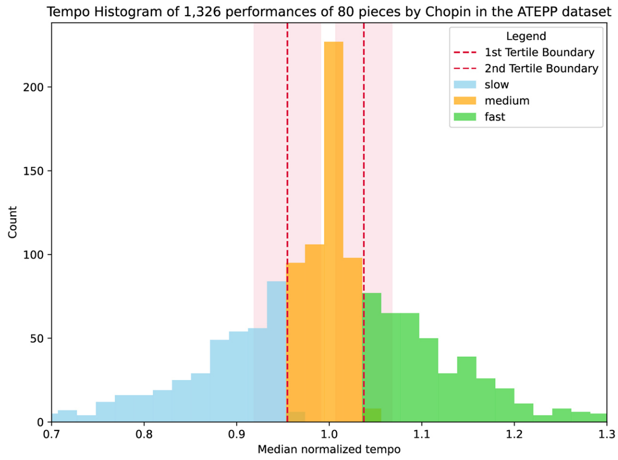

Figure 1 Tempo histogram over median normalized tempi across all (1326) performances of (80) pieces by Chopin in the ATEPP dataset, with the resulting tertile-based tempo grouping. Red dashed lines indicate the mean positions of the first and second tertile boundaries, while the shaded regions represent the corresponding standard deviations.

Composer	Dataset	Tempo	Fast	Medium	Slow
Beethoven	ATEPP	fast	0.22	–	–
		medium	0.26	0.30	–
		slow	0.26	0.33	0.39
	(n)ASAP	fast	0.45	–	–
		medium	0.49	0.51	–
		slow	0.49	0.54	0.60
Chopin	ATEPP	fast	0.35	–	–
		medium	0.33	0.35	–
		slow	0.32	0.33	0.43
	(n)ASAP	fast	0.29	–	–
		medium	0.32	0.35	–
		slow	0.33	0.35	0.38
	Vienna4×22	fast	0.80	–	–
		medium	0.72	0.81	–
		slow	0.80	0.80	0.81

Table 2 Average correlations between performances of pieces by Beethoven and Chopin per dataset. For these two composers, mean within-group correlations (main diagonal) were slightly higher than the mean between-group correlations (lower triangle).

Performers are expected to maintain the integrity of Beethoven's rhythmic patterns. In contrast, Chopin's works invite a freer and more pronounced use of rubato, with performers often stretching and compressing the tempo to heighten emotional expression. This contrast provides an interesting perspective on how different composers and musical eras influence tempo and expressive timing patterns. Together, Beethoven and Chopin also account for a substantial portion of our analyzed corpora, composing over one-third of all pieces and more than half of performances.

It can be seen that both the within- and between-group correlations are considerably higher in the Vienna4×22 dataset than in both (n)ASAP and ATEPP. The lower correlation values observed in (n)ASAP and ATEPP may partly reflect greater expressive variability across performers, suggesting that timing patterns are not uniformly preserved across tempo conditions in more diverse repertoire. This could be further compounded by the fact that pieces in (n)ASAP and ATEPP are generally longer than the short excerpts in Vienna4×22, potentially leaving more room for performers to diverge in their local timing decisions over the course of a piece. Another noteworthy observation is that all correlations of performances estimated as slow (both within and between group) are higher than the mean within medium or within fast tempo group correlations, indicating that timing patterns become more stable and reproducible at slower tempos, where the relative impact of measurements and motor timing noise are reduced.

For Beethoven's pieces in the ATEPP dataset, the average within-group correlation ($\rho_{\mu_w} = 0.30$, $\rho_{\sigma_w} = 0.07$; lower triangle mean in the correlation matrix) is slightly higher than the mean between-group correlation ($\rho_{\mu_b} = 0.28$, $\rho_{\sigma_b} = 0.03$; main diagonal mean). Likewise, performances by Beethoven in the (n)ASAP dataset show a slightly higher within-group correlation ($\rho_{\mu_w} = 0.52$, $\rho_{\sigma_w} = 0.06$) compared to the between-group correlation ($\rho_{\mu_b} = 0.51$, $\rho_{\sigma_b} = 0.02$).

For the Chopin pieces, the performances in Vienna 4×22 show higher mean correlations both within and between groups ($\rho_{\mu_w} = 0.81$, $\rho_{\sigma_w} = 0.005$, $\rho_{\mu_b} = 0.81$, $\rho_{\sigma_b} = 0.038$), which likely reflects both the shorter length of the performances (piece excerpts) and the more laboratory-like recording conditions. Performances in the ATEPP and (n)ASAP datasets have comparable levels of within- and between-group correlations, with the mean between-tempo group correlation being slightly lower.

We also compared individual pairwise within-group correlations to between-group correlations between the slow and fast tempo groups for statistically significant differences. Specifically, we tested whether the between-group correlation is significantly lower than the within-group correlation, using first the slow tempo group as the within-group 'benchmark,' then the fast tempo group.

Table 3 reports the results aggregated by composer and dataset. For each piece, we computed the proportion of pairwise comparisons where the between-group correlation is significantly lower than the within-group correlation and report the mean and standard deviation of these proportions. As can be seen, the choice of reference group substantially affects the results. When using the slow tempo group as the within-group reference, a much larger proportion of comparisons show lower between-group correlations, providing stronger evidence against the relational invariance hypothesis (H0). As discussed above, however, the generally higher level of within-group correlations observed in the slow tempo group could reflect greater measurement robustness at slower tempi. The relatively low proportion of comparisons where between-group correlations are significantly lower than within-group correlations for the fast tempo group suggests that relational invariance (H0) may hold.

Overall, the results of the correlational comparison provide only weak and inconclusive evidence against relational invariance. As shown in Table 3, the proportion of pieces showing significantly lower between-group correlations remains modest and is sensitive to the choice of within-group reference. Given that slow-tempo within-group correlations are likely inflated by greater measurement robustness rather than genuine timing stability, we consider the fast-tempo reference results to be the more conservative and reliable benchmark.

As briefly outlined above, in both Repp (1994) and Desain and Honing (1994), the measurements at the three tempi originated from repeated performances by the same person, respectively. This is not the case for Vienna4×22 and (n)ASAP, where each performance stems from a different individual. In the 197 pieces (6112 performances) analyzed in the ATEPP dataset, 156 pieces (79.19%) include multiple performances by the same performer within the same tempo group, while

				Within = Slow		Within = Fast	
Composer	**Dataset**	**pc**	**pfc**	**$\%_\mu$**	**$\%_\sigma$**	**$\%_\mu$**	**$\%_\sigma$**
Beethoven	ATEPP	25	770	53.95	14.20	26.55	10.40
	(n)ASAP	28	212	58.79	25.39	25.30	29.20
Chopin	ATEPP	80	1326	45.46	27.71	36.80	27.59
	(n)ASAP	26	271	39.42	27.18	15.50	11.09
	Vienna4×22	2	44	21.79	8.82	11.90	11.17

Table 3 Mean and standard deviation values of the proportion of pairwise correlation comparisons per piece where the between-group correlation (slow vs. fast tempo) is significantly lower than the within-group correlation ($\alpha = 0.05$), reported separately for the slow and fast groups, each serving in turn as the within-group benchmark. Columns pc and pfc indicate piece and performance count by that composer in the respective dataset.

126 pieces (63.96%) include performances by the same performer across different tempo groups. The frequent presence of the same performer within a single group may help to explain the higher within-group correlations observed in this dataset. It is expected that performances by the same individual, whether within or across tempo groups, yield higher correlations than those between different performers. While the aforementioned studies focus on how a single performer maintains timing patterns across tempi, we studied the more generic performance space of human interpretation through performances from real-world musical contexts. This broader perspective allows us to examine how timing patterns generally change in relation to global tempo across a diverse set of interpretations.

5.3.2 ANOVA

Table 4 summarizes the piecewise ANOVA results for metrical strength, tempo level, and their interaction across performances of pieces by Beethoven and Chopin from different datasets. The table reports the absolute and relative proportions of significant effects for each factor, as well as for their interaction. Note that the critical test of H0 against HA lies in the interaction term between metrical strength and tempo: if timing patterns scale proportionally with tempo (H0), the way metrical strength shapes timing should remain stable across tempo groups, and the interaction should be nonsignificant. Conversely, if timing patterns change qualitatively with tempo (HA), a significant interaction would be expected.

At the level of individual factors, it can be seen that both metrical strength and tempo significantly influence the timing profile across a large proportion of pieces. This underscores the importance of both meter and tempo-related cues in shaping timing patterns. However, the interaction between metrical strength and tempo reaches significance only in a minority of the pieces examined.

		Metrical Strength		Tempo Group		Interaction	
Composer	**Dataset**	n	%	n	%	n	%
Beethoven	ATEPP	23	92.00	25	100.00	2	8.00
	(n)ASAP	19	65.52	21	72.41	5	17.24
Chopin	ATEPP	54	67.50	70	87.50	14	17.50
	(n)ASAP	24	92.31	23	88.46	2	7.69
	Vienna4×22	2	100.00	2	100.00	0	0.00

Table 4 Excerpt of the analysis of variance results, in terms of absolute and relative amounts of performances of works by Beethoven and Chopin for which a significant effect ($P < 0.05$) for Metre/Tempo was found, shown separately by dataset. Both metrical strength and tempo group significantly affect timing patterns, but, in most pieces, their influence is largely additive rather than interactive.

In light of the relational invariance hypothesis (H0), the results of the correlational comparison and the ANOVA together suggest a rather nuanced picture. The correlational evidence is weak and inconclusive: while between-group correlations tend to be lower than within-group correlations, this difference is modest and partly attributable to greater measurement robustness at slower tempi rather than genuine timing stability. The ANOVA results are more informative in this regard—both metrical strength and tempo systematically influence timing patterns, consistent with relational invariance (H0). Nonetheless, the significant interaction between metrical strength and tempo observed in a low number of performances suggests that, in a few cases, the effect of tempo on timing patterns is not uniform across metrical position, which we interpret as weak evidence in favor of tempo-dependent scaling (HA).

6 EXPERIMENT 2: HOW DO SCORE MARKINGS, INTERVAL CONSONANCES, AND OCTAVE REGISTER INFLUENCE ARTICULATION?

6.1 HYPOTHESIS

Based on prior literature, we tested the following hypotheses in Experiment 2: **H1 (Score-marking hypothesis):** Notes marked as *legato* exhibit greater overlap (higher key overlap ratio [KOR]) than those marked as *staccato*. **H2 (Register hypothesis):** Articulation varies systematically with pitch register, with higher registers encouraging increased overlap. **H3 (Interval consonance hypothesis):** Consonant intervals exhibit greater overlap than dissonant intervals. The second and third hypotheses remain empirically underexplored in real-world performance data and are evaluated at scale in this section.

6.2 METHOD

The three factors investigated in this experiment—score markings, octave register, and interval consonance—were selected on the basis of established empirical findings and theoretical proposals in the performance literature. Score-articulation markings (*legato*/*staccato*) are the most direct link between compositional intent and performed overlap (Bresin and Battel, 2000). Octave register is included because piano tone decay accelerates in higher registers, and prior work shows that pianists compensate by increasing key overlap to maintain perceptual continuity (Repp, 1995). Interval consonance is included following Repp (1996b), who proposed that consonant intervals may be held with greater overlap than dissonant ones; this hypothesis has not previously been tested at scale in real-world performance data, making it a natural candidate for a large-scale replication.

Key overlap ratio (KOR): To study articulation, we investigated the key overlap ratio originally proposed by Repp (1995) and also used by Bresin and Battel (2000). First, we measured the key overlap time (KOT), which is the amount of overlap time between successive piano tones that belong to the same perceptual stream, e.g., the melody line. Since post-key press sound decays over time, brief key overlap is typically perceived not as simultaneity but as increased continuity. The KOR measures the amount of KOT in terms of the IOI between the same respective adjacent notes, and therefore presents a unit-less, numerical measure of their relationship. As KOT is dependent on the interval size between the notes and varies with IOI (Repp, 1995), it is more natural to account for the ratio and investigate KOR in the setting of diverse datasets (Bresin and Battel, 2000). In short, the greater the KOR, the more *legato*-like two adjacent notes will be perceived. *Staccato* notes, on the other hand, have a negative KOR. Figure 2 illustrates both measures: the top figure shows two *legato*-marked notes, which have a KOT of 0.43 ms, an IOI of 0.57 ms, and correspondingly a KOR of $\frac{KOT}{IOI} = \frac{0.43}{0.57} = 0.75$. For the bottom figure, a negative KOT for a *staccato*-marked note results in a KOR of $\frac{-0.27}{0.57} = -0.47$.

Since our analysis relies on score-voicing information, we chose to include the Vienna4×22, Batik, Magaloff, and Zeilinger datasets, where the source score editions are either explicitly known or have undergone careful editorial scrutiny. In contrast with Bresin and Battel (2000), who examined the full right hand, and Repp (1995), who selected individual voices, we restricted our analysis to note transitions in the melody voice. This choice is motivated by the fact that the melody voice typically forms a continuous sequence of successive notes, making note-to-note overlap measures well-defined and interpretable.

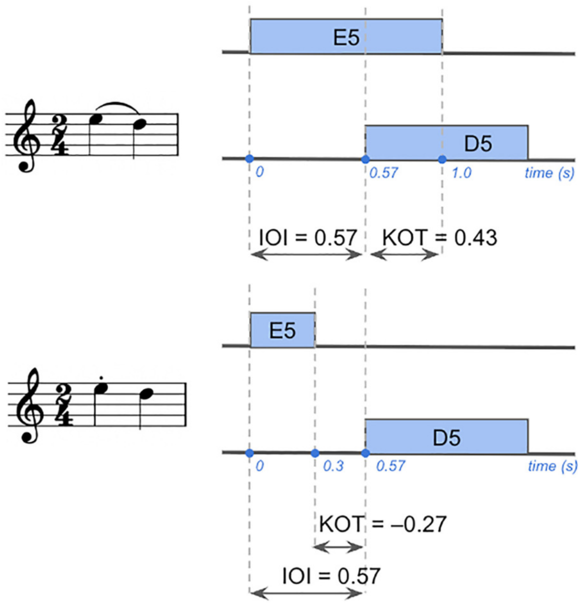

Figure 2 Illustration of key overlap time and inter-onset interval for *legato* (top) and *staccato* (bottom) scenarios. Absolute timestamps of the score events are shown in blue.

Other voices, such as bass or inner voices, often contain sustained tones or discontinuous entries, where key overlap is less interpretable.

While the original KOR analysis by Bresin and Battel (2000) focused on the finger and key actions in their articulation, we augmented our analysis with *sound-off* articulation, which computes the prolonged note offset resulting from sustained pedal usage. For datasets where MIDI sustain control (CC64) is available and reliable (Vienna4×22, Batik, and Magaloff)[4], we determined a dataset-specific *sound-off* threshold—the CC 64 value at which the dampers make contact with the strings.

The sound-off threshold corresponds to a physically fixed position for a given piano: the point at which dampers touch the strings when the pedal is partially depressed. To estimate this position, we aggregated all CC 64 values recorded across all pieces (and pianists) in each dataset and inspected the resulting histogram (Figure 3). During legato re-pedaling, the performer briefly holds the pedal at the damping position before re-pressing; this creates a characteristic accumulation—a local mode—in the intermediate range of the CC 64 distribution, between the dominant peak at the resting position (CC ≈ 0 for Batik and Magaloff; CC ≈ 15 for Vienna4×22, whose Bösendorfer CEUS sensor records an analog signal rather than a binary on/off) and the fully-pressed peak at CC 127. We identify the threshold as the modal value of the lightly smoothed histogram of control values within the intermediate range 50–100, yielding dataset-specific thresholds of 76 (Vienna4×22), 75 (Batik), and 60 (Magaloff). Notes whose key offsets occur while the pedal control remains above this threshold are treated as acoustically sustained, and their effective note offsets are extended until the subsequent pedal-release timestamp, identified as a time point where the CC 64 signal exhibits a large decrease (i.e., decreasing pedal values). KOR is then computed as described above, with statistics reported separately for note transitions occurring under pedal-off and pedal-on conditions.

6.3 RESULTS

We begin by analyzing the KOR of notes based on their articulation score markings (*legato* and *staccato*), following Bresin and Battel (2000). For the Batik and Magaloff datasets, we lacked information about the specific score editions used by the pianists and whether they differ from our reference scores (the Henle Urtext Edition for Magaloff and the New Mozart Edition for Batik). However, it is reasonable to assume that the majority of articulation markings align across carefully edited scores used by professional pianists; hence, with few exceptions, the annotations in the performers' scores likely match ours.

Articulation distinctions such as *legato* and *staccato* are, in theory, conveyed through variations in the amount of note overlap between adjacent notes. Figure 4 illustrates the distribution of note-level KORs using violin

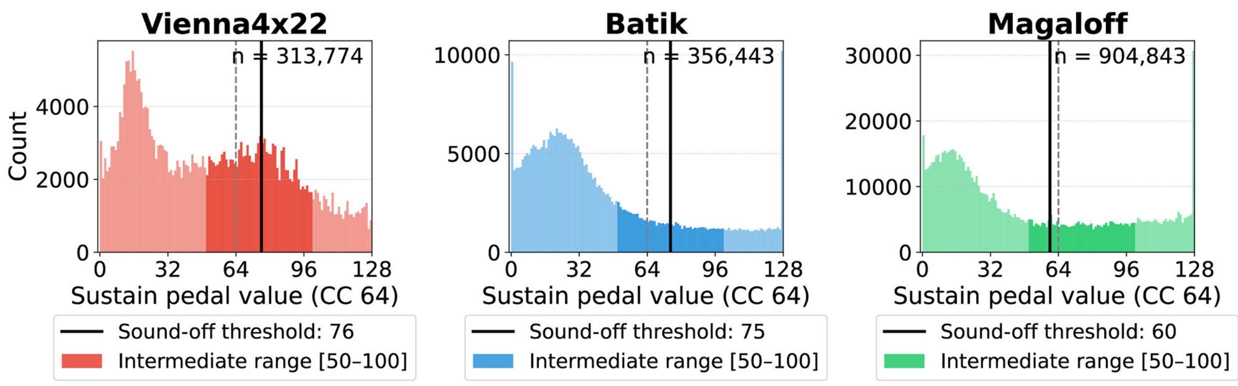

Figure 3 Distribution of sustain pedal values (CC 64) aggregated across all pieces (and pianists) for each dataset. The shaded region marks the intermediate range (control values 50–100) used to locate the modal accumulation corresponding to the physical damping point. The solid vertical line shows the estimated sound-off threshold (Vienna4×22: 76, Batik: 75, Magaloff: 60); the dashed line marks the conventional default of 64.

plots, comparing *legato* and *staccato* markings. For the *staccato* group, we include notes marked with *staccato* and *staccatissimo* dots, while the *legato* group includes notes covered by slurs (excluding the last note of a slur group).

Overall, the differences between the *legato* and *staccato* groups are significant for all datasets: Vienna4×22 ($F(1, 5492) = 142.0, P < 0.05$), Batik ($F(1, 11957) = 537.3$, $P < 0.05$), and Magaloff ($F(1, 55651) = 152.1$, $P < 0.05$). These results indicate a consistent articulation-dependent effect on KOR under the sound-off (pedal-released) condition. The Batik dataset shows the strongest distinction, suggesting that the performer interprets Mozart's articulation markings with particularly clear differentiation between *legato* and *staccato* passages, while Magaloff shows the least-pronounced differences between articulation groups.

In addition to score markings, consonance is another factor influencing the overlap between neighboring notes. Figure 5 shows the distribution of KORs for neighboring note pairs, grouped by interval. We computed intervals in semitone steps modulo 12 and analyzed them separately by direction (upward or downward). The interval groups, ranging from unison to tritone, are

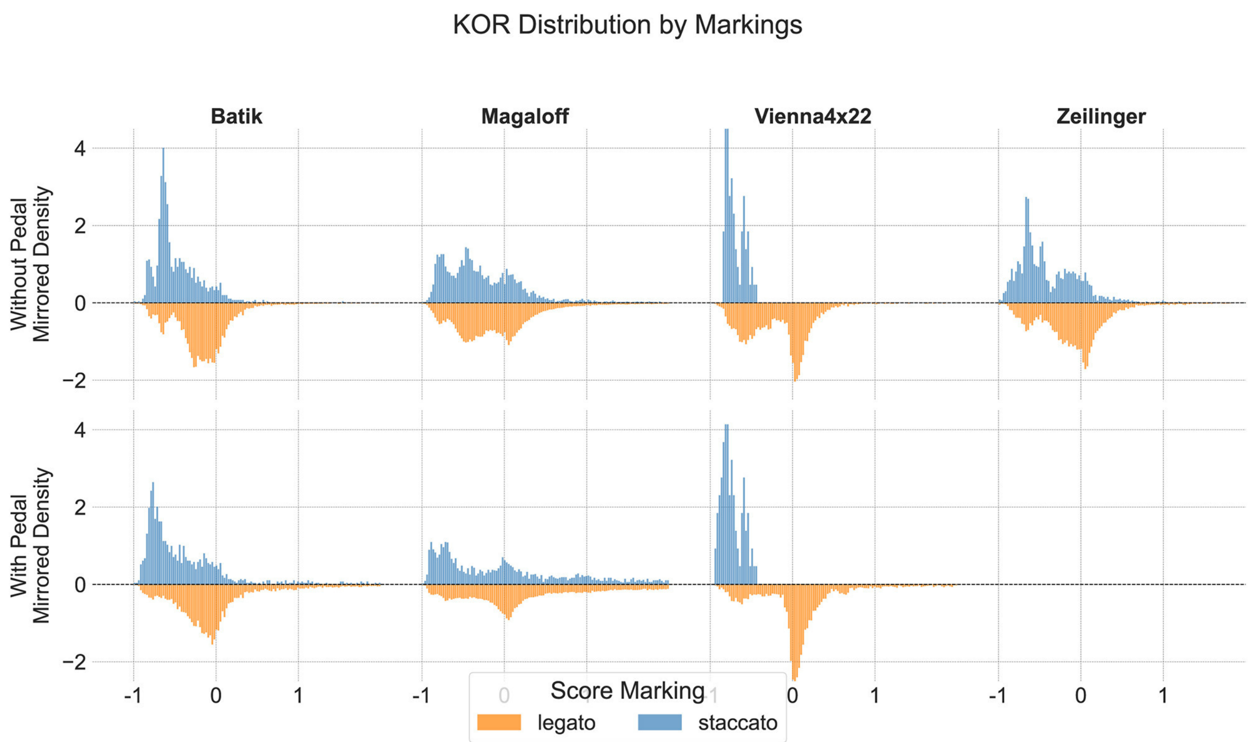

Figure 4 Key overlap ratio distribution of *legato*- and *staccato*-marked notes in different datasets. The vertical axis shows a mirrored kernel density estimate histogram for visual comparison between articulation groups. *Legato* values indicate mirrored density and do not represent negative probabilities.

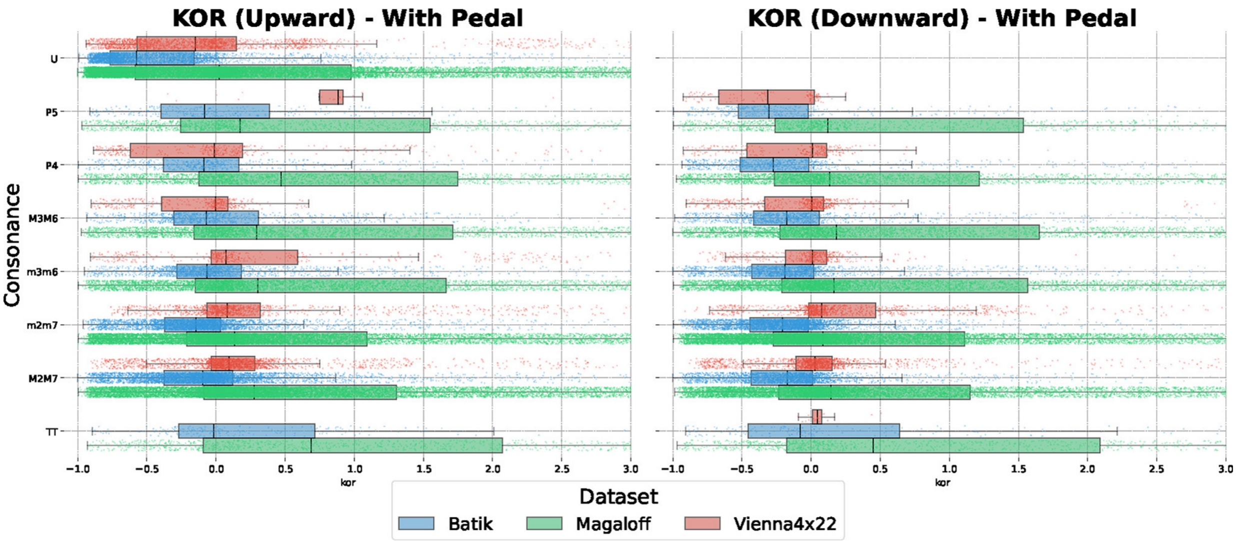

Figure 5 Boxplots of the key overlap ratio for all note transitions, grouped by interval consonance level across datasets. The unison (same pitch) transitions are grouped in the left (upward) side.

ordered by increasing levels of dissonance according to Western common-practice harmony.

None of the performances in the datasets analyzed support the premise that consonant note pairs overlap more than dissonant ones. While ANOVA tests indicate significant differences in mean KOR across interval groups, Pearson's correlation coefficient between mean KOR and consonance level (with unison = 8 and tritone = 0) reveals an insignificant negative trend: $r(6) = -0.50, P = 0.21$ for upward intervals and $r(5) = -0.54, P = 0.21$ for downward intervals. However, it is important to note that these interval categories are highly unbalanced in the performed repertoire.

In terms of the impact of direction, we also found that upward transitions generally contain more overlaps than downward transitions among all datasets, with the difference being significant for all intervals. The interval with the largest difference is the perfect fourth, a historically controversial interval with regard to consonance, where upward leaps have an average KOR (with pedal sustain) of 0.72 ± 1.35 compared to 0.46 ± 1.26 for downward leaps, resulting in a 56% difference. This observation may imply that the upward intervals sound more stable than the downward intervals, leading to the articulation choice of a longer overlap.

Building on the findings of Repp (1995), we investigated how key overlap varies with register in performed notes. His scale-listening experiment demonstrated that increased overlap in the upper register was associated with improved legato perception, while the lower register tolerated less overlap, presumably due to the instrument's inherent acoustic decay. However, in Repp's pianist performance experiment, no significant effects of register were found. We investigated this relationship in a real-world performance context by categorizing melody note KORs based on the octave register of both notes (with A0 marking the pitch class A in register 0).

As shown in Figure 6, our analysis reveals an overall positive relationship between register and overlap articulation. In the Magaloff dataset, the Pearson correlation between register and the mean KOR computed for each register is $r(5) = 0.96, P < 0.05$. The Zeilinger ($r(3) = 0.83, P = 0.08$) and Vienna4×22 ($r(1) = 0.26, P = 0.83$) datasets, which contain compositions within a narrower register range, also show a weak, nonsignificant positive trend. In contrast, the Batik dataset defies this pattern, showing a significant negative correlation ($r(4) = -0.90, P = 0.02$) between register and overlap time.

Treating octave register as a categorical factor, a one-way ANOVA reveals significant differences in mean

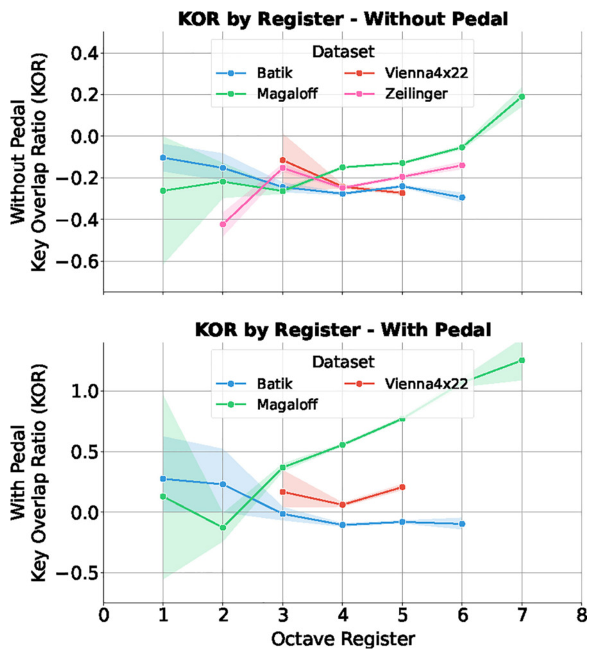

Figure 6 Relationship plot between the octave register and key overlap ratio across datasets. The shaded area represents the 95% confidence interval, interpolated between discrete points for each octave.

KOR across registers ($F = 173.4$, $P < 10^{-200}$). Post-hoc Holm-corrected pairwise comparisons indicate that these differences are systematic across register pairs, despite substantial imbalance in register sample sizes.

Overall, these findings suggest that, in real-world piano performances, higher registers tend to encourage more overlapping articulation, potentially compensating for the faster decay of high-pitched notes. However, this positive relationship is only statistically significant for the Magaloff dataset ($r(5) = .96, p < .05$); the Batik dataset shows a significant *negative* register–KOR relationship ($r(4) = -0.90, P = 0.02$), reflecting a repertoire-specific effect, while the Zeilinger and Vienna4×22 datasets show nonsignificant correlations, indicating that this relationship is likely influenced by repertoire characteristics and individual performer interpretation, rather than being a universal performance rule.

7 EXPERIMENT 3: CHORD ASYNCHRONY—MELODY LEAD OR MELODY LAG?

7.1 HYPOTHESIS

Building on prior work (see Section 2.3), we investigate the following hypotheses on asynchrony: **H1 (Metrical hierarchy)**: Asynchrony magnitude increases on metrically stronger positions, with downbeats showing the largest root mean squaree (RMS) values, followed by beats, and off-beats. **H2 (Tempo hypothesis)**: RMS asynchrony decreases with faster tempi. **H3 (Texture hypothesis)**: RMS asynchrony decreases in pieces with higher chordal density. **H4 (Voice-specific asynchrony hypothesis)**: Melody notes systematically precede other voices (melody lead), while bass patterns vary by metrical position, with bass anticipation most pronounced on downbeats. **H5 (Velocity hypothesis)**: Observed asynchronies arise from dynamic accentuation rather than intentional timing, such that melody asynchrony correlates positively with velocity differences between melody and accompaniment voices.

7.2 METHOD

We restricted our analysis to polyphonic score events where at least three non-grace or non-ornamental notes are performed simultaneously. For each such event, we calculated the pairwise onset time differences between all notes and computed the RMS of these differences as a measure of overall asynchrony (Palmer, 1989). Specifically, for a chord with n notes, we computed all $\binom{n}{2}$ pairwise differences, squared them, took their mean, and extracted the square root. This RMS measure, expressed in milliseconds, captures the overall temporal spread of notes within a chord, regardless of voice assignment.

To examine melody (bass) asynchronies specifically, we computed the onset timing difference between the melody (bass) voice and the mean onset time of all remaining notes in the chord. Where voice annotations are unavailable, we followed the common heuristic of assigning the highest and lowest pitch notes to the melody and bass voices, respectively (Palmer, 1989).

To distinguish temporal accentuation from the velocity hypothesis, which suggests that asynchronies arise mechanically from dynamic differences rather than intentional timing displacement, we computed the velocity difference between the melody note and the mean velocity of the remaining notes at each onset and measured the correlation between these timing and velocity differences.

7.3 RESULTS

With respect to H1 (metrical hierarchy), as a general measure of temporal variability across staves, Figure 7 shows mean asynchrony measured in milliseconds (lower x-axis) for selected composers, grouped by metrical position. The upper x-axis shows polyphonic density—that is, the proportion of events at each metrical position that contain three or more simultaneous notes, providing textural context for interpreting asynchrony patterns. To address potential transcription artifacts while preserving performance variability, we excluded polyphonic events with RMS asynchrony values exceeding 200 ms.

RMS asynchrony is comparable on downbeat and beat positions but less pronounced on off-beat positions. Similarly, chordal density is comparable on downbeat and beat positions, particularly for Romantic compositions, and lower on off-beat positions across all compositional periods. Across all metrical positions, performances of works by Chopin and Schubert exhibit the greatest degree of asynchrony.

To address H2 and H3, we performed two separate linear regression analyses. For H2 (tempo effects), we use estimated tempo (in BPM) as the independent variable and asynchrony as the dependent variable. For H3 (texture effects), we used texture density (proportion of joint note onsets with three or more non-grace notes, over all onsets in a given piece) as the independent variable and asynchrony as the dependent variable. We conducted these analyses at the performer level for performer-specific datasets and at the piece level for multi-performer datasets, and we assessed statistical significance at $P < 0.05$.

For tempo effects on overall asynchrony, we found 76 (out of 248) significant tempo–asynchrony relationships. Of these, 43 were negative (mean slope = –0.0007) and 33 were positive (mean slope = 0.0019). These slopes are very small, indicating that the overall influence of performed tempo on asynchrony is negligible, even when significant.

Looking at the effect of musical texture, we found only 11 (out of 248, or 4.4%) regressions to be statistically significant, which is close to the rate of Type I errors. Moreover, the large magnitude of regression coefficients

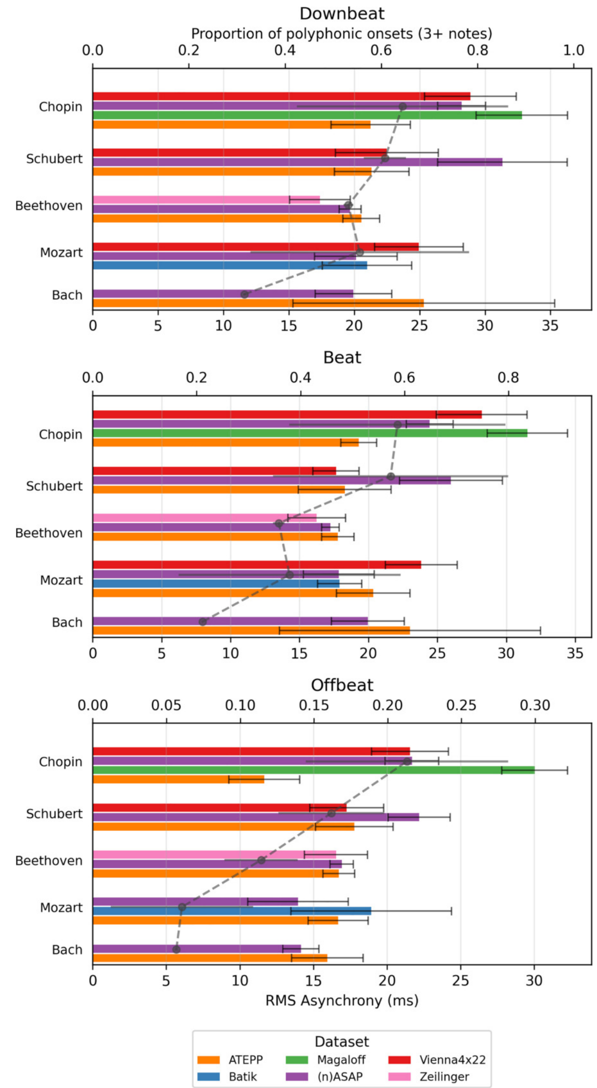

Figure 7 Mean root mean square (RMS) asynchrony (in ms; lower axis and colored bars) and polyphonic density (expressed as a proportion; upper axis and black dashed line) by composer and metrical position: downbeat (top), beat (middle), and offbeat (bottom). Polyphonic density indicates the proportion of events at each metrical position containing 3+ simultaneous notes. RMS error bars show 95% confidence intervals across performers for mixed-performer datasets or alternatively across pieces for single-performer datasets. Polyphonic density error bars show standard deviation across pieces.

suggests numerical instability in the model. When analyzing texture effects for each hand separately, we found 67 significant relationships for the right hand (staff 1: 21 positive with mean slope = 0.0035; 46 negative with mean slope = –0.0007) and 64 for the left hand (staff 2: 29 positive with mean slope = 0.0007; 35 negative with mean slope = –0.0007). Overall, the results indicate that the influence of texture on asynchrony is slightly negative and generally negligible.

With respect to H4 (voice-specific asynchrony hypothesis), we investigated melody and bass note timing with respect to their tendency to lead or lag at different metrical positions. Bass notes tend to lag (59.81%) more often than they lead (39.29%). Across all datasets and metrical positions, bass lags have a mean magnitude of approximately 14.9 ms (SD = 18.1 ms), with comparable values observed at downbeat, beat, and offbeat positions. Bass leads are slightly larger in magnitude, averaging 21.8 ms (SD = 35.3 ms). Lead time magnitudes are greatest at downbeat positions (mean = 28.8 ms) and decrease at weaker metrical positions (beats and offbeats).

Melody notes tend to lead more often than they lag, with an approximate 70:30 lead-to-lag ratio when measured relative to all (both hand) accompaniment notes and roughly 80:20 relative to right-hand accompaniment notes only. Lead-time magnitudes are similar in magnitude across all metrical positions, and, on average, amount to 17.39 ms (SD = 18.13 ms) across both hands and 16 ms (SD = 17.99 ms) within the right hand. Lag times are greater at downbeat positions when measured across both hands, and less so when measured within the right hand only.

Addressing H5 (velocity hypothesis), Figure 8 illustrates the relationship between melody lead (as the timing difference between the melody voice and the mean onset time of other notes at the same onset, measured in milliseconds) and MIDI velocity differences (binned into four-unit bins), within the right hand (top) and across both hands (bottom), grouped by metrical position (columns), for a subset of our analyzed corpora (ATEPP, (n)ASAP, and Vienna4×22). Note that this figure only includes polyphonic events with positive velocity differences. Across all datasets, cases where the melody note is played softer than the accompanying notes is relatively rare, accounting for between 2% and 16% of polyphonic events across hands and between 2% and 13% within the right hand. The lowest rates are observed in Vienna4×22 and Batik, while the higher proportions occur in (n)ASAP and ATEPP. As can be seen in Figure 8, both within-hand and cross-hand melody leads show a positive correlation with velocity differences (as computed on raw, unbinned velocity difference values), consistent with the velocity hypothesis that timing deviations between melodic and accompanying notes at shared score onsets may reflect dynamic accentuation strategies.

8 DISCUSSION: DISTRIBUTIONAL DIFFERENCES BETWEEN TRANSCRIBED AND RECORDED MIDI DATA

Recent advances in automatic piano transcription have made it possible to curate large-scale transcribed MIDI datasets, such as ATEPP (Zhang et al., 2022), which we included in our analysis. While state-of-the-art transcription models (Kong et al., 2021) achieve impressive accuracy, we acknowledge that transcribed data may exhibit distributional differences compared to precisely captured performance measurements.

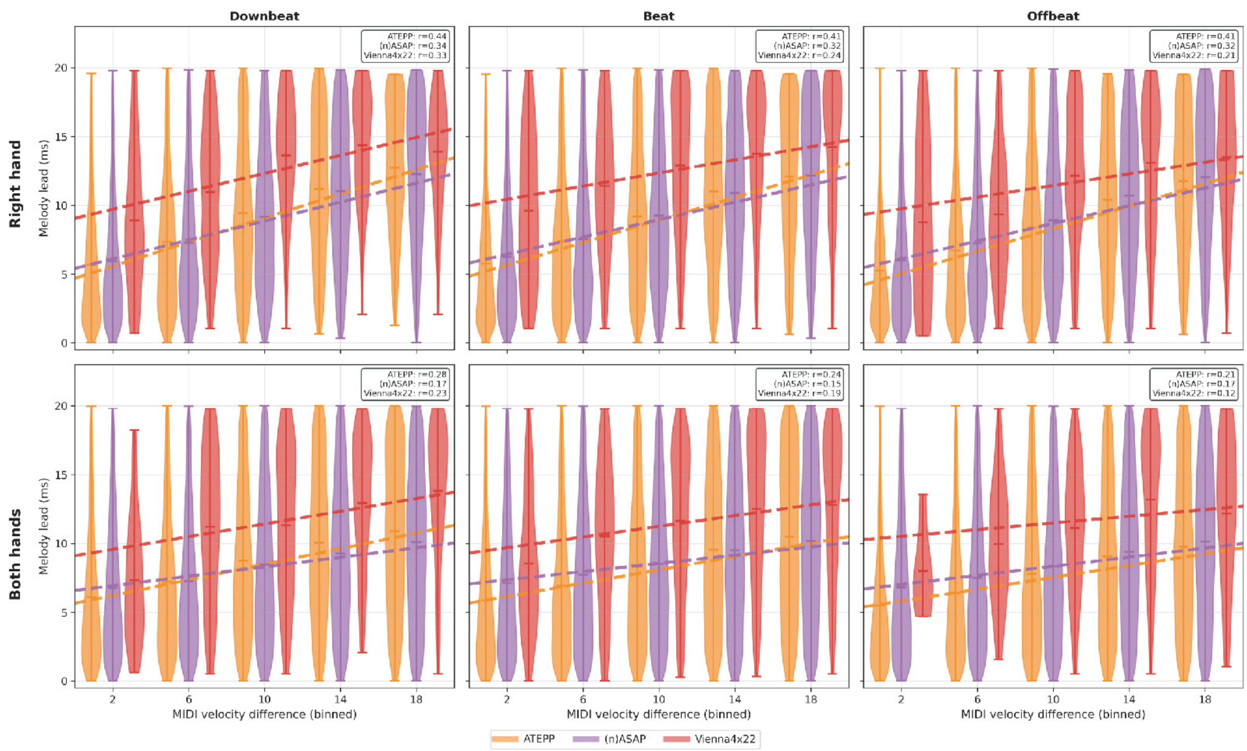

Figure 8 Melody lead time (ms) as a function of MIDI velocity difference between melody and accompaniment (binned at four-unit intervals), shown for right hand only (top) and both hands (bottom), at downbeat, beat, and offbeat positions. Violin plots show the distribution per velocity bin across three corpora (ATEPP, (n)ASAP, and Vienna4×22); dashed lines are dataset-specific regression fits on unbinned data. Melody lead increases with velocity difference, most strongly for the right hand at downbeat positions.

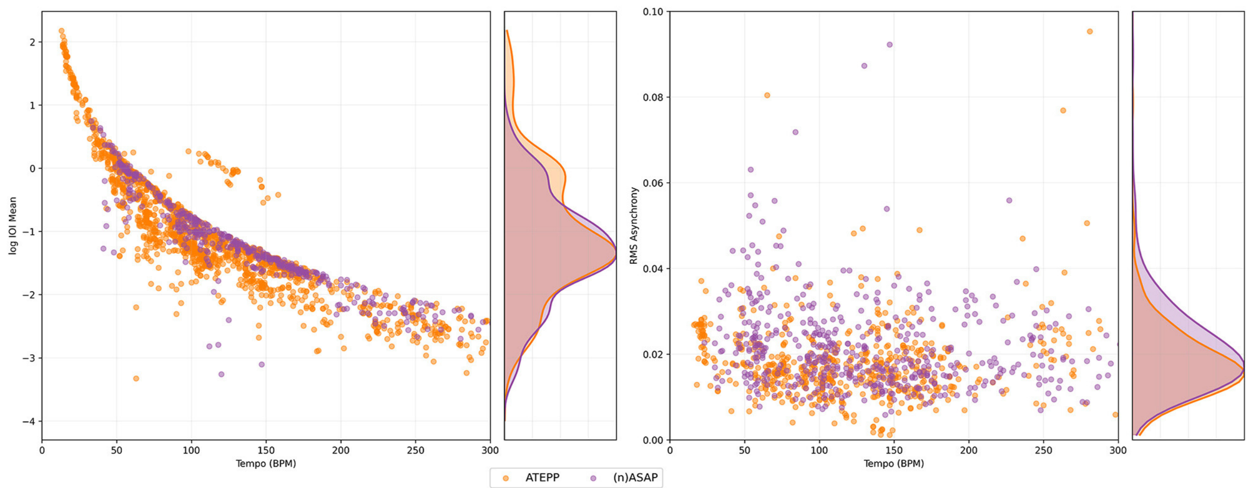

Figure 9 Comparison between ATEPP (transcribed MIDI) and (n)ASAP (recorded MIDI) in terms of timing features discussed in previous experiments, computed on a subset of 125 pieces present in both datasets: score-normalized, log-scaled inter-onset interval values (left, discussed in Section 5) and overall root mean square asynchrony (right, see Section 7), both as a function of estimated tempo (beats per minute). Each point corresponds to one performance-wise mean.

Assessing transcription quality without ground truth data remains an open methodological challenge in music information retrieval and beyond, and developing principled evaluation methods for transcribed symbolic data in such settings is a direction we leave to future work. However, to contextualize potential limitations of our findings, we present an exploratory comparison of distributional characteristics between transcribed and recorded MIDI data.

In the following, we compare ATEPP with another recorded dataset from our analyzed corpora across selected timing features, with the aim of exploring systematic differences rather than validating transcription quality, and assessed their potential impact on expressive timing analyses.

Figure 9 compares ATEPP and (n)ASAP on a subset of 125 pieces present in both datasets. The left subplot shows the distribution of score-normalized, log-scaled

IOI values as a function of estimated tempo (BPM), corresponding to the features used in Section 5. The right subplot depicts the distribution of average asynchrony at polyphonic score onsets containing three or more non-ornamental notes, computed as the RMS of absolute timing differences across all possible note pairs within each onset, again plotted against estimated BPM. The RMS measure corresponds to one of the features used in Section 7. Each point in the figure corresponds to a performance-wise mean.

The kernel density estimates reveal substantial overlap between ATEPP (transcribed) and (n)ASAP (recorded) data across both feature dimensions. For IOI values, both datasets exhibit similar central tendencies and spread, with IOI and tempo being negatively correlated (as expected) across both datasets. The RMS asynchrony distributions likewise show comparable shapes, with both datasets showing peaks in lower mean RMS asynchrony values (<20 ms) and similar spread patterns. This similarity in distributions is also reflected in our experiments (see above), where both datasets show comparable ranges and overall trends.

Note that the similarity in IOI and RMS distributions may not generalize to other expressive features not shown here (e.g., dynamics, articulation, pedaling). Transcription models vary in their ability to capture different aspects of performance, and timing (particularly in case of the piano) may be among the more reliably transcribed features compared to velocity or continuous controls.

9 CONCLUSION

This study conducted a large-scale analysis of expressive timing in classical piano performances, focusing on three core hypotheses from earlier performance research.

First, regarding the relational invariance of timing patterns to global tempo, our findings suggest that timing patterns generally do scale proportionally with tempo. In a few cases, however, scaling varied with metrical position, providing weak evidence against full relational invariance. Together, these results largely support the relational invariance hypothesis proposed by Repp (1994).

Second, we examined how articulation is influenced by score markings, interval consonance, and register. The results confirm that articulation correlates with score markings and register, while no consistent relationship was observed with interval consonance. The score marking and register findings are consistent with Bresin and Battel (2000) and Repp (1995), respectively, and our large-scale results lend additional empirical support to these hypotheses in ecologically valid performance contexts. The lack of support for the interval consonance hypothesis of Repp (1996b) is noteworthy: while this hypothesis had previously only been tested in controlled listening experiments, our experiments on real-world performance data do not replicate previously drawn conclusions, suggesting that consonance may not reliably shape articulation in the diverse repertoire and performance contexts captured by our corpora.

Third, we investigated chord-asynchrony patterns in relation to various metrical and textural aspects, along with the role of the timing of melody and bass notes in accentuating a particular stream. We also investigated the relationship between melody lead and dynamic differences, where our findings suggest that a positive relationship between dynamics and the amount of melody note lead can be observed on chords occurring on the beat level, but less so on other metrical positions. This is broadly consistent with Goebl (2001), who demonstrated that melody lead is closely tied to dynamic accentuation, and with Palmer (1996a) and Goebl et al. (2010), who found asynchrony to be most pronounced at metrically strong positions. Our large-scale results thus replicate and generalize these earlier findings across a much broader range of repertoire and performers.

Finally, we presented a high-level comparison between recorded and transcribed MIDI data, showing that the ATEPP (transcribed) dataset exhibits distributional patterns similar to that of recorded performance data, particularly mirroring (n)ASAP in its diversity and scale, which supports the reliability of automatic transcription for large-scale expressive timing analysis. While this is an encouraging result for the use of transcribed data in performance research, the comparison remains exploratory and restricted to timing-related features. Other aspects of performance may be less reliably captured by automatic transcription models, and caution is warranted when generalizing these findings beyond the features examined here.

Generally, we wish to propose and promote large-scale *distant listening* analysis as a complement to detailed, controlled studies in empirical musicology. The value of such an approach, as illustrated by this study, lies not in replacing controlled experiments but in providing a broader ecological context for evaluating their conclusions. Where our findings converge with earlier small-scale studies, e.g., as in the case of score marking and register effects on articulation, or the role of dynamic accentuation in chord asynchronies, the large-scale replication adds ecological validity and generalizability. Where they diverge or remain inconclusive, e.g., relational invariance and interval consonance, they raise important questions about the boundary conditions under which earlier hypotheses hold and motivate more targeted follow-up investigations. At the same time, we acknowledge that the increased noise, heterogeneity, and reliance on automatic annotation that characterize large-scale approaches introduce their own methodological challenges, and that the conclusions drawn from such studies should be interpreted with appropriate caution.

NOTES

1. Although the Zeilinger/Beethoven and Magaloff/Chopin datasets are proprietary, we provide all performance measurements necessary to reproduce our analyses in full.
2. While the analysis in the referenced study uses a large number of IOI levels (113 for the Theme and 223 for the Variation), we decided to group the onset factor into three levels of *metrical strength* (offbeat, beat, and downbeat) to simplify the analysis and reduce the number of potential interactions. This approach aims to make the results more interpretable while still allowing us to investigate the interaction between tempo and timing.
3. We refrain from using a clustering approach for the split, as it may result in unequal group size, which may have an undesirable effect on the subsequent ANOVA test, potentially violating the assumption of homogeneity of variances.
4. Regarding the Zeilinger dataset, the sustain-pedal information has been found to be problematic. The pedal event messages include multiple concurrent streams of control data rather than a single clean sustain signal, possibly from incorrect merging of multiple takes. While the overall temporal shape appears plausible, it is unclear which stream of pedal states corresponds to the note events, and no reliable post-hoc procedure can be validated without additional recording metadata. Since the dataset is not publicly available and no authoritative calibration information exists, we treat the Zeilinger sustain-pedal data as unreliable and exclude it from our analysis.

ACKNOWLEDGMENTS

PH, SP, and GW acknowledge support by the European Research Council (ERC) under the EU's Horizon 2020 programme (grant no. 101019375, 'Whither Music?'). The work by HZ and SD was supported by funding from UK Research and Innovation (grant no.) EP/S022694/1). CC acknowledges funding by the Austrian Science Fund (FWF) under grants PAT 8820923 ('Rach3: A Computational Approach to Study Piano Rehearsal') and PIN 1347924 ('AURA: Augmenting musical interaction via EVAs').

AUTHORS' CONTRIBUTIONS

PH and HZ conceptualized the study, conducted the experiments, and wrote and revised the manuscript. SP contributed critical analysis and discussion. SP, SD, and GW provided critical review, commentary, and editorial revision. CC reviewed code. SD, GW, and CC secured funding for the work.

COMPETING INTERESTS

SD was editor-in-chief and CC is a member of the editorial team; neither had any editorial oversight of or decision-making role regarding this article.

AUTHOR AFFILIATIONS

Patricia Hu https://orcid.org/0009-0006-8057-3993
Institute of Computational Perception, Johannes Kepler University, Linz, Austria

Huan Zhang https://orcid.org/0000-0001-6987-7554
Centre for Digital Music, Queen Mary University of London, London, United Kingdom

Silvan David Peter https://orcid.org/0009-0000-8328-291X
Institute of Computational Perception, Johannes Kepler University, Linz, Austria

Carlos Eduardo Cancino-Chacón https://orcid.org/0000-0001-5770-7005
Institute of Computational Perception, Johannes Kepler University, Linz, Austria

Simon Dixon https://orcid.org/0000-0002-6098-481X
Centre for Digital Music, Queen Mary University of London, London, United Kingdom

Gerhard Widmer https://orcid.org/0000-0003-3531-1282
Institute of Computational Perception, Johannes Kepler University, Linz, Austria

REFERENCES

Benade, A. H. (1977). *Fundamentals of Musical Acoustics*. Oxford University Press.

Bresin, R., and Battel, G. U. (2000). Articulation strategies in expressive piano performance: Analysis of legato, staccato, and repeated notes in performances of the Andante movement of Mozart's Sonata in G Major (K.545). *Journal of New Music Research*, *29*(3), 211–224. 10.1076/jnmr.29.3.211.3092.

Cancino-Chacón, C. E., Gadermaier, T., Widmer, G., and Grachten, M. (2017). An evaluation of linear and non-linear models of expressive dynamics in classical piano and symphonic music. *Machine Learning*, *106*, 887–909. 10.1007/s10994-017-5631-y.

Cayless, H. A. (2005). DocScapes: Visualizing document structures with SVG. In *Proceedings of the Digital Humanities*, British Columbia, Canada.

Desain, P., and Honing, H. (1989). The quantization of musical time: A connectionist approach. *Computer Music Journal*, *13*(3), 56–66. 10.2307/3680012.

Desain, P., and Honing, H. (1994). Does expressive timing in music performance scale proportionally with tempo? *Psychological Research*, *56*(4), 285–292. 10.1007/bf00419658.

Flossmann, S., Goebl, W., Grachten, M., Niedermayer, B., and Widmer, G. (2010). The Magaloff project: An interim report. *Journal of New Music Research*, *39*(4), 363–377. 10.1080/09298215.2010.523469.

Foscarin, F., Karystinaios, E., Peter, S. D., Cancino-Chacón, C., Grachten, M., and Widmer, G. (2022). The match file format: Encoding alignments between scores and performances. In *Proceedings of the Music Encoding Conference (MEC)*, Halifax, Canada.

Goebl, W. (1999). The Vienna 4x22 piano corpus (master's Thesis, Numerisch-klassifikatorische Interpretationsanalyse mit dem 'Bösendorfer Computerflügel'). 10.21939/4X22.

Goebl, W. (2001). Melody lead in piano performance: Expressive device or artifact? *The Journal of the Acoustical Society of America*, *110*(1), 563–572. 10.1121/1.1376133.

Goebl, W., Flossmann, S., and Widmer, G. (2010). Investigations of between-hand synchronization in

Magaloff's Chopin. *Computer Music Journal*, *34*(3), 35–44. 10.1162/comj_a_00002.

Gotham, M., Gullings, K., Hamm, C., Hughes, B., Jarvis, B., Lavengood, M., and Peterson, J. (2023). *Open Music Theory*. VIVA Pressbooks.

Hentschel, J., Neuwirth, M., and Rohrmeier, M. (2021). The annotated Mozart sonatas: Score, harmony, and cadence. *Transactions of the International Society for Music Information Retrieval*, 4(1), 67–80. 10.5334/tismir.63.

Hentschel, J., Rammos, Y., Neuwirth, M., and Rohrmeier, M. (2025). A corpus and a modular infrastructure for the empirical study of (an)notated music. *Scientific Data*, *12*(1), 685. 10.1038/s41597-025-04976-z.

Honing, H. (2007). Is expressive timing relational invariant under tempo transformation? *Psychology of Music*, *35*(2), 276–285. 10.1177/0305735607070380.

Honing, H., and Ladinig, O. (2008). The potential of the internet for music perception research: A comment on lab-based versus web-based studies. *Empirical Musicology Review*.

Honing, H., and Reips, U.-D. (2008). Web-based versus lab-based studies: A response to Kendall. *Empirical Musicology Review*, *3*(2), 73–77. 10.18061/1811/31943.

Hu, P., and Widmer, G. (2023). The Batik-plays-Mozart corpus: Linking performance to score to musicological annotations. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Milan, Italy.

Jänicke, S., Franzini, G., Cheema, M. F., and Scheuermann, G. (2015). On close and distant reading in digital humanities: A survey and future challenges. *EuroVis (STARs)*, 2015, 83–103.

Kendall, R. (2008). Commentary on "The potential of the Internet for music perception research: A comment on lab-based versus Web-based studies" by Honing & Ladinig. *Empirical Musicology Review*, *3*(1), 8–10. 10.18061/1811/31693.

Kong, Q., Li, B., Song, X., Wan, Y., and Wang, Y. (2021). High-resolution piano transcription with pedals by regressing onset and offset times. *IEEE/ACM Transactions on Audio Speech and Language Processing*, *29*, 3707–3717. 10.1109/taslp.2021.3121991.

Kosta, K., Bandtlow, O. F., and Chew, E. (2015). A change-point approach towards representing musical dynamics. In *International Conference on Mathematics and Computation in Music*, London, UK, (pp. 179–184).

Langner, J., and Goebl, W. (2003). Visualizing expressive performance in tempo-loudness space. *Computer Music Journal*, *27*(4), 69–83. 10.1162/014892603322730514.

Lerch, A., Arthur, C., Pati, A., and Gururani, S. (2021). An interdisciplinary review of music performance analysis. *Transactions of the International Society for Music Information Retrieval (TISMIR)*.

Moretti, F. (2013). *Distant Reading*. Verso Books.

Palmer, C. (1989). Mapping musical thought to musical performance. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(2), 331. 10.1037/0096-1523.15.2.331.

Palmer, C. (1996a). Anatomy of a performance: Sources of musical expression. *Music Perception*, *13*(3), 433–453. 10.2307/40286178.

Palmer, C. (1996b). On the assignment of structure in music performance. *Music Perception*, *14*(1), 23–56. 10.2307/40285708.

Peter, S. (2023). Online symbolic music alignment with offline reinforcement learning. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Milan, Italy.

Peter, S. D., Cancino-Chacón, C. E., Foscarin, F., Henkel, F., and Widmer, G. (2023). Automatic note-level alignments in the ASAP dataset. *Transactions of the International Society for Music Information Retrieval (TISMIR)*, *6*(1), 27–42.

Piez, W. (2005). SVG visualization of TEI texts. In *Proceedings of the Digital Humanities*, British Columbia, Canada.

Pugin, L. (2015). The challenge of data in digital musicology. *Frontiers in Digital Humanities*, *2*, 4. 10.3389/fdigh.2015.00004.

Repp, B. H. (1990). Patterns of expressive timing in performances of a Beethoven minuet by nineteen famous pianists. *The Journal of the Acoustical Society of America*, *88*(2), 622–641. 10.1121/1.399766.

Repp, B. H. (1992a). Diversity and commonality in music performance: An analysis of timing microstructure in Schumann's "Träumerei.". *The Journal of the Acoustical Society of America*, *92*(5), 2546–2568. 10.1121/1.404425.

Repp, B. H. (1992b). Probing the cognitive representation of musical time: Structural constraints on the perception of timing perturbations. *Cognition*, *44*(3), 241–281. 10.1016/0010-0277(92)90003-z.

Repp, B. H. (1994). Relational invariance of expressive microstructure across global tempo changes in music performance: An exploratory study. *Psychological Research*, *56*, 269–284. 10.1007/bf00419657.

Repp, B. H. (1995). Acoustics, perception, and production of legato articulation on a digital piano. *The Journal of the Acoustical Society of America*, *97*. 10.1121/1.413065.

Repp, B. H. (1996a). Patterns of note onset asynchronies in expressive piano performance. *The Journal of the Acoustical Society of America*, *100*(6), 3917–3932. 10.1121/1.417245.

Repp, B. H. (1996b). Pedal timing and tempo in expressive piano performance: A preliminary investigation. *Psychology of Music*, *24*(2), 199–221. 10.1177/0305735696242011.

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

Zhang, H., Tang, J., Rafee, S., Dixon, S., and Fazekas, G. (2022). ATEPP: A dataset of automatically transcribed expressive piano performance. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Bengaluru, India.

TO CITE THIS ARTICLE:
Hu, P., Zhang, H., Peter, S.D., Cancino-Chacón, C.E., Dixon, S., & Widmer, G. (2026). Revisiting Expressive Timing in Piano Performance at Scale: A Distant Listening, Multi-Corpus Study. *Transactions of the International Society for Music Information Retrieval, 9*(1), 329–346. DOI: https://doi.org/10.5334/tismir.317

Submitted: 1 July 2025 **Accepted:** 4 May 2026 **Published:** 17 July 2026

Transactions of the International Society for Music Information Retrieval is a peer-reviewed open access journal published by Ubiquity Press.