SIMULATING PIANO PERFORMANCE MISTAKES FOR MUSIC LEARNING

Alia MORSI (alia.morsi@upf.edu) [1], **Huan ZHANG** (huan.zhang@qmul.ac.uk) [2], **Akira MAEZAWA** [3], **Simon DIXON** [2], and **Xavier SERRA** [1]

[1] *Music Technology Group*, **Universitat Pompeu Fabra**, Barcelona, Spain
[2] *Centre for Digital Music*, **Queen Mary University of London**, London, United Kingdom
[3] **Yamaha Corporation**, Hamamatsu, Japan

ABSTRACT

The development of machine-learning based technologies to support music instrument learning needs large-scale datasets that capture the different stages of learning in a manner that is both realistic and computation-friendly. We are interested in modeling the mistakes of beginner-intermediate piano performances in practice or work-in-progress settings. In the absence of large-scale data representing our target case, our approach is to start by understanding such mistakes from real data and then provide a methodology for their simulation, thus creating synthetic data to support the training of performance assessment models. The main goals of this paper are: a) to propose a taxonomy of performance mistakes, specifically apt for simulating or reproducing/recreating them on mistake-free MIDI performances, and b) to provide a pipeline for creating synthetic datasets based on the former. We incorporate prior research in related contexts to facilitate the understanding of common mistake behaviours. Then, we design a hierarchical mistake taxonomy to categorize two real-world datasets capturing relevant piano performance contexts. Finally, we discuss our approach with 3 music teachers through a listening test and subsequent discussions.

1. INTRODUCTION

To build music education systems that can detect different types of performance mistakes we must devise frameworks that represent such mistake patterns effectively. Analyzing beginner-intermediate piano performance data from practice settings to understand and model learner behaviours would be useful, but to our knowledge there are no published datasets with annotations specifically capturing this context. Furthermore, it remains an open question what type of annotation would be both realistic and computationally friendly, and what its relationship would be to the learning behaviours we should seek to detect. Our goal is to develop an understanding of such mistakes and provide a methodology for their realistic simulation, to provide relevant synthetic data for model (pre-)training.

In computational research on piano performance mistakes, there is a tendency to discuss mistakes in terms of literal deviations from the underlying music score, such as pitch insertions and deletions [1–3], and rhythmic deviations [2], with the latter being less frequently examined because pitch is relatively fixed by the compositional notation of Western tonal music [4] while rhythmic errors are more difficult to observe under score-performance alignment algorithms [5]. While this paradigm permits interesting analysis of performance mistakes, especially when coupled with musical and psychological perspectives, it alone is not ideal for recreating them in a realistic manner.

We propose a hierarchical, computation-friendly taxonomy of piano performance errors based on observations from two piano performance datasets related to our target case, where mistake behaviours are higher level units formed from literal score deviations on the axes of pitch, time, velocity, and structure (which we consider lower-level). Then, we create a performance mistake simulation model based on this taxonomy and use it to apply mistakes to MIDI piano performances belonging to a suitable repertoire with respect to the targeted proficiency levels. Finally, we present results from a questionnaire and interview study with music teachers, who provided feedback on the realism of our synthesized examples, taxonomy, and methodology.

The paper is organized as follows: section 2 builds a foundation to understand piano performance mistakes by connecting with related research regarding their causes, manifestations, and categorizations. In section 3 we demonstrate our main observations from real performance data, after which we define our hierarchical framework for mistake analysis and simulation in section 4. Sections 5 and 6 concern the implementation of our mistake simulator, and the preparation of MIDI performances for use as input. The results and insights of the teacher interviews are shown in section 7, followed by our conclusion and future work (section 8). All code and materials are available on the companion page [1].

2. PIANO PERFORMANCE MISTAKES

Technical difficulties, lack of concentration, and poor memorization are some of the many factors which contribute to performance errors [1, 6]. Some research further examines patterns of performance mistake production with respect to their underlying psychological processes [4, 7, 8], where

[1] https://github.com/Alia-morsi/piano-synmist

insights on a performer's mental representations of a music score can be provided through analyzing the distribution and relative frequencies of performance mistakes [2].

Since the main purpose of this work is to simulate performance mistakes based on a realistic taxonomy, this section reviews the appearance of performance mistakes (and hence how they are detected, and potentially reproduced), some of their prominent patterns, and how different researchers have categorized them.

2.1 Appearance of Performance Mistakes

In most research on piano performance mistakes, determining their presence or their extent is based the faithfulness of a performance to its corresponding music score [1]. One limitation of this approach is that it treats all deviations from the music notation as performance errors, despite such deviations (ornamentation, rubato, etc) being expected as part of a performer's artistic license in Western tonal music [4]. Furthermore, it does not account for the fact that not all score deviations are perceived as errors by listeners [1], nor for the differences between editions of scores.

2.1.1 Pitch Insertions, Deletions, and Substitutions

In research relying on such deviations for analysis, whether in music technology contexts [3, 5, 6, 9, 10] or psychological studies [1, 4], performance mistakes are often reported in terms of the edit distance between score and performed note sequences, i.e. as pitch insertions, deletions, and substitutions, although often the latter is just treated as a simultaneous pitch deletion and insertion. In some cases, percentages of these note-edit categories are reported [1, 2, 6]. Such statistics can be useful to observe quantitative relationships between performances of music pieces with similar characteristics (e.g. style, technical difficulty). Flossman et al. [6] investigated whether the ratios found by Repp [1] are preserved across other data (ballade, polonaise, nocturne, etc.). Though without context those statistics might not be very informative on their own which is why researchers often propose further categories as shown in 2.2. Moreover, for the purpose of simulating mistakes, it is unlikely that applying these same percentages even on the same pieces would result ones that sound natural.

Despite the importance of timing and velocity in the perception of mistakes, they have not received as much attention as pitches. We believe that this could be due to the relative ease of pitch quantization and categorization compared to timing, as there could exist confusions between expressive timing and mistake behaviors (fermata vs hesitation, and rubato vs tempo instability, etc). Another reason could be due to variability in note velocity measures. Both should receive more attention in subsequent performance mistake studies with recent advances in Automatic Music Transcription for piano with respect to temporal resolution [11] and velocity estimation [12].

2.1.2 Reproducing Piano Performance Mistakes

To the best of our knowledge, apart from the work of Morsi et al. [13] where the simulation of performance mistakes was conducted as data augmentation, there has not been work specifically addressing the reproduction or simulation of performance mistakes. Despite differences in naming, the score deviations applied in their system were: 1) pitch omission, 2) pitch substitution, 3) pitch insertion, 4) short pause (between 0.3 and 0.8 seconds) and repeat last note, and 5) long pause (between 2 to 4 seconds) and repeat last note. Insertions and substitutions were constrained to be n semitones (specified by the user) around the correct note.

2.1.3 Common Patterns and Behaviours

Several researchers affirm the impact of musical context on the type of errors [1, 2, 4, 7]. For example, harmonically related errors more associated with homophonic than polyphonic passages [4], and pitch errors mostly occur in non-melody voices or inside chords [1, 2]. Furthermore, less errors are observed in notes of recurring musical motifs than for other passages [2], and higher error frequencies are observed with higher note densities in a piece, reflecting the increased technical demands the more the notes that should be played per time unit.

Other patterns could be related to the context of a performance (i.e sight reading vs 'quick study' vs recital) and performer characteristics such as age [8] or performance level [14]. For example beginners are more likely to repeat errors, either because they do not realize there was an error, or because they are unable to correct it in repeat performances [14]. Children with more musical training are better able to detect and correct performance mistakes with less perseveration behavior [15]. Morijiri et al. [16] note that beat interruptions were the most common type of errors among adult beginners. Weber and Parncutt [14] propose a theory for error management in music performance, where a whole error undergoes *pre*, *during*, and *post error* stages. A performer's *pre-error* behavior is a form of risk management, while the *during* and *post-error* stages are forms of error management. This connects with the observation that the majority of insertion errors are of low note velocity compared to their immediate neighbourhood [17], as a performer's awareness that they are about to make a mistake causes them to play the inserted note at a lower velocity as a form of risk management. It also suggests that it might be useful to consider the connection between consecutive mistake events so that the behaviours of the three stages can be effectively demonstrated.

2.2 Categorizations of Performance Mistakes

Researchers have proposed different categories to contextualize the deviations between the notes of a performance and its score (described in section 2.1.1) into more descriptive categories based on the performance context, to better identify performance mistakes and compare between error-trends. Palmer and van de Sande [4] have categorized errors on the dimensions of size (note, chord, or a combination), source (contextual/non-contextual), type (insertion, deletion, substitution, shift), and movement (anticipatory/perseveratory). In another example on children piano sightreading, Gudmundsdottir [8] classified the pitch insertions observed as erroneous pitches (those which do not match target pitches in the score) and redundant pitches

(repetitions of correct pitches, which could be due to hesitation).

In further analysis of the Magaloff Corpus [6, 18], the following error categories were proposed, which are mostly pitch related by virtue of their context for mistake analysis given that this corpus represents on-stage behavior by a professional pianist: **1) Forward-Related Error**: an 'out-of-score' choice due to the influence of an upcoming note; **2) Backward-Related Error**: the same, but due to the influence of a previous note; **3) Repeated Note**: an example of a backward-related error where a note is played twice, with the second note often played weaker than the first; **4) Nonharmonic Error**: a note insertion that introduces a harmonic clash; **5) Harmonic Error**: an insertion that does not introduce a harmonic clash, which usually signals a memorization problem, although occasionally it could be deliberate, for harmonic emphasis; **6) Tied Note**: Could take the form of a technical simplification applied when a repeated note is held like a tied one (a deletion), or the form of a repeated note instead of a tied one whether due to a memorization problem or intentionally; **7) Systematic Error**: an error that occurs in the same context more than 60% of the time it is observed; **8) Note Order Error**: occurs when switching the playing order of two or more successive notes; **9) Omitted Inner Voice**: deletions that involve inner voices. It is worth noting that errors can belong to more than one category.

3. EMPIRICAL OBSERVATIONS FROM PERFORMANCE DATA

We use two datasets to investigate performance errors. The *Burgmüller* set [13] consists of 50 recordings (25 pieces recorded twice) from the *Burgmüller Etudes, Op.100*, performed by an advanced pianist who only studied the Etudes briefly before the recording. The context of the recordings allows for natural mistake behaviours such as repetitions and occasional pauses, unlike well-practiced performance settings. The dataset presented by Jiang [19], which we call the *Expert-Novice* set, includes 83 piano performances from 21 adult beginner players, with a repertoire covering 7 easy pieces of folk and pop songs in the Western tonal tradition, with some rhythmic syncopation. It includes beginner-level mistakes in a recital context.

We analyzed both the *Burgmüller* and *Expert-Novice* datasets to better describe, represent, and simulate performance mistakes. Both sets contain natural mistakes since none of the performers were instructed to make mistakes deliberately. The result of this process is the basis for the proposed framework in section 4.

3.1 Burgmüller Dataset

Although the mistakes in *Burgmüller* are not beginner-intermediate level due to the experience of the performer (as noted in the feedback from music teachers in section 7), we believe that practice-like behaviours can be observed given its collection context (non-recital setting with limited time given to the performer to learn the pieces).

The performances are provided in MIDI format with annotations of the sites of mistakes. We manually listen to the mistakes and inspect their corresponding music scores. We analyze the first recordings of Etudes 9, 10, and 17-20 (each of the 25 Etudes was played twice), and provide our raw observations on the paper's companion page.[1]

The most commonly observed mistake is what we refer to as a **note mistouch**, which is an extra note inserted concurrently with a score note a tone or semitone above or below it. According to the categories explained in section 2.2, this would most often be of a *nonharmonic* nature, but also could arise as a *forward-related* or *backward-related* error, or a even a *systematic* one if it occurs multiple times over the course of the performance. A less frequent error type is a **note substitution**, where an incorrect note replaces a correct one. The substitution could a *harmonic error*, or a *nonharmonic* one (often due to confusion of accidentals). In some cases the substituted note is played strongly, indicating confidence of the player, and in other cases the substitutions demonstrate the *pre-error* behaviour described in section 2.1.3.

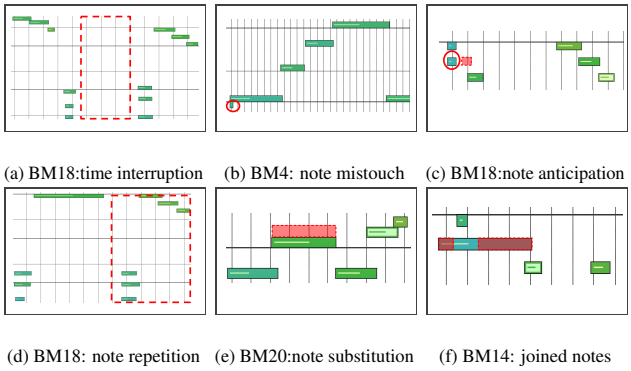
From a temporal perspective, we observe **time interruptions** which can manifest as long or short pauses in unlikely locations that interrupt the musical flow, which can sometimes occur after another mistake. Another pattern is that of holding a note or chord longer than expected, and causing a delay in the onsets of the upcoming score notes. We refer to this pattern as a **dragging** note. Sometimes it is observed on a correct score note, hence manifesting as irregularity in note times and velocities, and sometimes the dragging note behaviour happens with an incorrect note.

Note repetitions are also quite common. Often they are either a repeat of the last played note, or a repeat from a reasonable prior point in the score (e.g. the previous beat). Such repetitions often occur as a reaction to making a mistake, whether it is a time interruption, note substitution, a mistouch, or another. We also observe **joined notes**, where a player would tie a note instead of playing it twice (referred to as *tied note* in section 2.2, and occasional **note deletions**. Some examples are shown in Figure 1.

3.2 Expert-Novice Dataset

Since the performances are provided as audio, we transcribe [11] and manually correct the performances (pipeline available [1]) to obtain MIDI versions for easier inspection. We observe similar mistake categories as those found in *Burgmüller*.

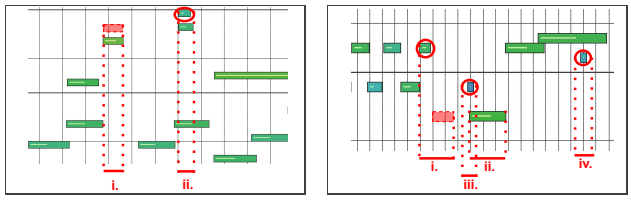
Note mistouches are observed around running scales or rapid chord changes. **Note substitutions** mostly occur with melody notes, often around accidentals, but also less frequently within the context of chords or harmonic intervals (2 note chords). The substitutions could be affected by the previous or next harmony, making them *backward-related* or *forward-related* errors respectively. An additional type of extra note was specifically observed in *Expert-Novice* with the addition of a note horizontally on an extra time step. It it is not clear if the mistake is a note substitution and a repeat, or due to mis-memorization and incorrect sense of the rhythmic grid. Both could be likely given the performances are beginner-level. Finally, **note deletions** take place in a less subtle manner than in *Burgmüller*, with



(a) BM18:time interruption (b) BM4: note mistouch (c) BM18:note anticipation

(d) BM18: note repetition (e) BM20:note substitution (f) BM14: joined notes

Figure 1: Illustrations of some examples from analysis of the Burgmüller data. Each subfigure is captioned with Etude number and error type. Red highlights the mistake.



(a) BM20: (i) a note substitution followed by (ii) a mistouch.

(b) BM16: (i) a note substitution followed by (ii) a mistouch, (iii) a time interruption and (iv) a note mistouch.

Figure 2: Note substitutions leading to subsequent mistakes.

either a melody note or an entire chord missing from the performance.

Time interruptions (or silences) and **dragging** notes are also observed, where in some cases the interruptions are followed by **note repetitions**. **Rhythmic errors** are different from dragging, occurring when one musical event gets displaced, or is performed with a wrong duration, but still fits into the rhythmic grid. This type of error usually occurs consistently within the piece, likely originating from practice. The more jazzy and syncopated repertoire shows more timing mistakes.

4. HIERARCHICAL TAXONOMY FOR MISTAKE ANALYSIS AND SIMULATION

From the analysis in section 3 we conclude that it is infeasible to isolate the detected mistake behaviours solely as either pitch, time, velocity or structure related mistakes, as most of the encountered error categories would manifest in terms of the same granular score deviations with slightly different parameters. For example, *time interruptions* (Fig. 1a) can be implemented as the insertion of a time gap through a shift t in the time of the performance MIDI roll, and a repetition error would also be formed of the same time shift operation with additional note insertions (Fig. 1d). The parameters of the inserted notes could be determined based on the behavioural profile of the player, how they are expected to repeat, and the effect of making a mistake on further mistakes until the player recovers, as shown in Fig. 2.

Taking into account the previous conclusions, we propose a multi-level taxonomy for understanding and simulating performance mistakes. It separates between **low-level** operations (granular deviations between a performance and its music score), and conceptual mistake categories such as those highlighted in section 2.2 which we refer to as **mid-level**. Hence, mistakes are mid-level concepts comprised of one or more lower level operations. Mid-level mistake categories should be chainable to form meaningful behaviours, to allow the construction of higher level concepts would permit the rendering of realistic student performances based on descriptions of student profiles or emotional states.

This taxonomy would be relevant for both the automatic analysis of performance files with mistakes or for the simulation of errors on mistake-free MIDI performances, though in this paper we only do the latter. For analysis, the goal would be to use computational approaches like audio-to-score alignment to detect the low-level deviations and connect them to the proposed mid-level categories. For mistake simulation, implementations of the desired mid-level mistakes should be provided using low-level operations.

Although we do not surpass the mid-level, we envision that higher levels of this taxonomy would allow the simulation of performances that reflect the different traits of learners using combinations of the proposed mid-level mistake concepts. For example, students who demonstrate impatience through pursuing tempi that they cannot really manage would run into frequent mistouch, substitution, and rollback errors, or less confident students would be likely to produce more ghost notes and joint notes caused by their nervous touch of the keys. Simulations at this level are analogous to training a model that generates text in the style of a specific writer, or to compose music in the style of a particular composer. As discussed in section 7, the piano teachers have encountered examples where student attitudes affect the patterns of mistakes they commit. This direction would require obtaining and comparing performance data of different students, which we hope to conduct in the future.

4.1 Low Level: Granular Score Deviations

They are operations directly applicable on a MIDI file with each affecting a single dimension or axis operations. **a) pitch insertion** and **b) pitch deletion** are the most basic, which are adding or removing pitches with respect to the reference. We treat substitutions as higher level operations that constitute a deletion and an insertion. **c) delay note** and **d) anticipate note** are adjustments to the note onset time. **e) extend note** and **f) shorten note** adjust the duration via the note offset time. Further temporal operations are **g) shift time**: insert a gap in the score; **h) go back**: move the 'playhead' back to a past point in the reference; and **i) skip**: skip a portion of the score.

4.2 Mid Level: Mistake Categories and Behaviours

The categories proposed at this level are what we would use to describe the mistake behaviours in section 3. Each is composed of one or more of the low level deviations described in section 4.1, and is connected to musical context/texture. We define 2 classes of mistake behaviours: **recovery operations** and **core mistakes**. Recovery operations encompass how players would regain the flow in

their performance, whether it is due to a loss of concentration, a memorization error, or a response to another kind of error until recovery. Core mistakes are the main mistake events, such as accidental note mistouch, or incorrect note asynchrony.

4.2.1 Recovery Operations

Rollback is when a player repeats a section that has already been played. It consists of a *shift time* and several *pitch insertion* operations. Sometimes rollback affects only the hand with a mistake, or sometimes both hands. The velocity and time of the repeated notes can vary depending on the target behaviour. Rollback usually follows a *re-orientation*, which in some cases follows a *core mistake*.

Re-orientation is the behaviour expressed when a player is in the process of regaining their playing-flow, usually in response to committing an error, but could also be due to a need for concentration due to the demands of the piece performed. It consists of a *shift-time* (reflecting either a short or long pause), and potentially one or more *extend-notes* until the flow is resumed, and could be followed by a *rollback* before the flow is resumed.

4.2.2 Core Mistakes

Mistouches are extra notes inserted concurrently with a score note a tone or semitone above or below it. On their own, they do not interrupt flow (due to their vertical co-occurence with a correct note). However, they can affect subsequent notes (resulting in error chaining), or could be followed by one or more recovery operations. From our observations, they commonly occur after jumps (or leaps), especially with staccato, or during runs (such as octave runs).

Ghost Notes are notes that are hit too lightly, which results in a short note with an almost inaudible volume. In the teacher feedback (section 7), they identified this error as common with students who tend to lack confidence, and mostly happens with notes played with the 5th (little) finger in an arpeggio or a chord. One also noted that this error is associated with a high reorientation time with beginner students (as they are trying to find the source of the sound discrepancy), and is likely followed by a rollback.

Incorrect Asynchrony is when a set of notes played together have onsets starting at slightly different times. This translates into several small **delay note** or **anticipate note** operations. Asynchrony is often used as an expressive tool, and we do not have a formal definition of when it is considered an error or not.

Substitution Error is when a player hits the wrong note/s with confidence, whether due to a mis-memorization or a wrong hand movement. Such substitution could be *harmonic* (e.g. a musically appropriate interval above or below the score note) or *nonharmonic* (the substitute note does not fit the musical context, like a misread accidental). Such an error could affect one note, two notes, or the whole chord.

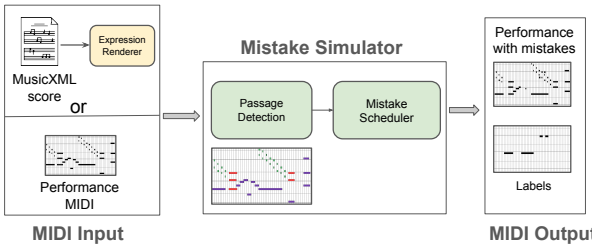


Figure 3: Mistake simulation pipeline. In the passage detection illustration, green and red stand for double notes and block chords, respectively.

Anticipated Note is when a correct note is played earlier than expected. This could be due to mismemorization of the rhythm, or simply due to accidentally touching the next note when playing a sequence. Again, this could also be an expressive tool, so not every anticipation should be considered wrong. In some cases, anticipated notes arise from a mistouch.

Join Notes occur when two consecutive strokes of the same key become one longer press. It fulfills both the *forward-related* error and the *tied notes* categories from [6], and can be executed as a combination of **extend note** and **delete note** lower level operations.

5. MISTAKE SIMULATOR IMPLEMENTATION

We develop a pipeline (Fig. 3) for applying the mid-level mistakes on mistake-free MIDI performances based on 2 steps: **1) Musical Element Detection**, which analyzes the score to identify relevant musical elements or passages for the application of mistakes, and **2) Mistake Scheduling**, which applies mid-level mistake operations on the MIDI performance (see section 5.2.1). The mistake scheduling phase is responsible for determining which mid-level mistakes to apply and their parameter values based on the output of step 1. The supported mistake operations are explained in section 5.2.1. They are a subset of the mid-level behaviours (section 4.2) and are built on an implementation of the low-level score operations described in section 4.1. Our code relies on the Partitura package [20]. As input, the Mistake Simulator expects a mistake-free MIDI performance, either a rendered performance from a music score or a real performance. We apply VirtuosoNet [21] to render a curated set of music scores within the skill level of our target user group (see section 6). The real performances we use are from the Vienna 4x22 [22] and SMD [23] datasets. The output of this pipeline are 2 MIDI tracks for each piece: one with the modified performance MIDI with mistakes, and another with labels corresponding to the applied mistake operations (both low and mid level), as shown in Fig 3.

5.1 Musical Element Detection

We implemented a basic rule-based system to detect regions of notes in a MIDI performance as **harmonic intervals** (2

note chords), **scales** and **block chords**. We chose these elements as simple examples of locations where it could be interesting to apply mistakes, although it is not by any means a thorough list.

Harmonic Intervals: For each note n, we obtain its set of onset neighbours with a diameter of d semitones $N_{n,d,\theta} = \{|Onset(k) - Onset(n)| < \theta, |Pitch(k) - Pitch(n)| <= d, k \in K\}$ where θ is the timing threshold which is set as 50ms. If $\|N_{n,d,\theta}\| = 2$ (only two notes within this onset window), then we label n as part of a harmonic interval.

Block Chords: Similarly, for the block chords we search for the neighbours with approximately the same onset and offset window where $\|N_{n,d,\theta}\| > 2$.

Scales: We look for the consecutive neighbours of n: $N_{n,d,\theta} = \{Onset(k) - Offset(n) < \theta, |Pitch(k) - Pitch(n)| <= d, k \in K\}$. We used a diameter d value of 2, although the value should be adjusted if this is to account for scales that include wider intervals.

5.2 Mistake Scheduling

In addition to the result of the mistake element detection step, the mistake scheduler expects a probability table (e.g. $P_{\text{Mistouch}}(n|n \in \text{BlockChords}) = 0.4$) which assigns a heuristic distribution for each mistake per musical element. Given a target number of mistakes per element class, the scheduler samples the notes on which mistakes will be applied.

5.2.1 Mistake Operations

Of the mid-level operations of section 4, we implement one recovery operation and four core mistake operations. **Rollback**(n, t): Starting from the onset of n for a subsequent (user-defined) window of t seconds, all note events are repeated, while all subsequent notes outside the window are shifted by t. **Forward-backward insertion**(m, n): A forward or backward neighbour m of n will be inserted on top of note n with similar onset and offset. **Mistouch**(m, n): A mistouched note m will be inserted on top of n with similar onset and offset times, where $|Pitch(m) - Pitch(n)| < 2$. **Note Substitution**$(m, n)$: n will be replaced by a pitch alteration m. In the case of chords, the whole chord group is replaced with the new pitches. **Dragging Note**(n, t): Starting from the onset of n, all note events are shifted by t seconds. At the end of the pipeline two parallel MIDI tracks can be exported: a) the modified performance and b) labels reflecting all the operations applied, in terms of low level operations and mid-level mistakes.

6. PERFORMANCE MIDI PREPARATION

6.1 Rendered Music Scores

We filter a set of piano-learning repertoire and download corresponding score renditions from Musescore [2] . To obtain expressive performances of the MIDI files, we applied the open-source performance rendering model VirtuosoNet [21] to the scores. Overall, we gathered 153 music scores, with a rendered duration of 208.16 minutes. Further details are provided on the companion page.[1] Our scores included the following: 1) Notable piano education works *Alfred's Basic Adult All-in-One Course (Book 1-2)* [24] and *Piano Adventures by Nancy Faber (Book 1-2)* [25]. 2) *Burgmüller Op.100* for comparison with the Burgmüller performance set (section 3.1 **3)** Samples from three Czerny etudes (*Op.599, Op.849, Op.299*). **4)** A compilation from catalogues of easy-to-intermediate level classical masterpieces, compiled by Schirmer Library [3] and ABRSM examination board [4] , to form the *Easy Classical* collection. Note that some of the pieces are custom reductions or rearrangements of famous pieces. All public-sourced MusicXML scores have been manually inspected. For the custom reduction / rearrangement scores from the easy classical selections, we validate them by playing to ensure they are within a beginner-intermediate level range.

6.2 MIDI Performance Data

We use subsets of the ASAP [26] and SMD [23] datasets filtered by heuristics to eliminate performances that would certainly be out of reach for the beginner-intermediate range: Average note density (notes per second) ≤ 10; Performance length ≤ 150 seconds; Number of polyphonic voices ≤ 3 (this is verified manually, by removing all the fugues from ASAP, and checking the orchestration for SMD).

Since ASAP contains multiple interpretations of the same piece, we only count one for each composition to avoid redundancy. Overall, this brings 32 pieces from ASAP and 11 pieces from SMD to constitute our score collection. We also use the Vienna 4×22 corpus [22] that features four excerpts with slower passages from the classical repertoire.

7. FEEDBACK FROM MUSIC TEACHERS

We conducted a 90 minute interview with three music teachers, each consisting of a listening survey to rate the realism mistake excerpts and an open-ended discussion about common student mistake patterns and our proposed framework.

7.1 Part 1: Mistake Excerpt Ratings

The teachers rated 12 short mistake excerpts on a scale of 1–5, where 1 denotes a mistake that sounds artificial, and 5 denotes one that sounds like a real student mistake. They were asked to justify each choice, thus providing insights into the potential shortcomings of our approach. The samples included two real mistake excerpts from the Burgmüller dataset, two synthetic mistakes applied on real performances (sec. 6.2), and eight synthetic mistakes on rendered performances (sec. 6.1). The full study as well as the teacher responses can be found in the project repository.[1]

The rating results demonstrated a degree of agreement between the teachers (standard deviation for all samples averaged to 0.65), suggesting a common perception of student mistake patterns. The highest average rating for an excerpt was 4.33, with the perceived naturalness owing to a confusion before the mistake which they thought reflected real student behaviour. The worst excerpt received a 1.33 average rating, and the reason given by one of the teachers was that the high pitched notes are difficult to get wrong when

[2] musescore.org

[3] https://www.halleonard.com/series/SCHLIB/schirmers-library-of-musical-classics

[4] https://www.abrsm.org/en-es

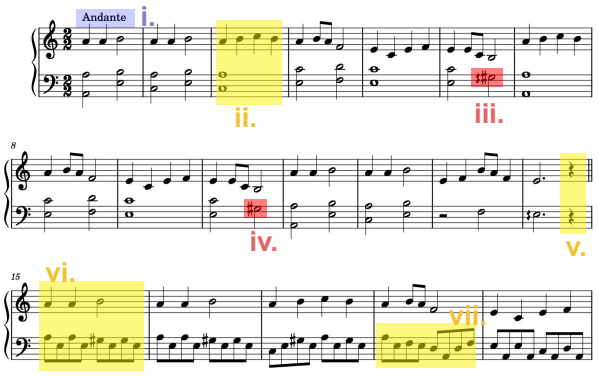


Figure 4: Teacher labeling of common mistakes on *Sakura*, from Alfred's [24] book 2: i) Tempo misread; ii) LH rhythm might erroneously double right hand; iii) and iv) Accidental misread; v) Rest duration not being properly executed; vi) Change of LH duration leads to wrong rhythm; vii) Complex structure, thumb crossing may lead to delays.

in playing in the mid registers. The teachers demonstrated most discrepancy in an excerpt from *Clair de Lune*, which was rated 5 by a teacher and 1 by the others. The low raters thought the mistakes were too drastic such that a natural behaviour would have been to rollback rather than continue. On the other hand, the teacher that rated it highly found the confident recovery sensible given that it happened with the motif, something the player should have internalized well.

Interestingly, they commented that real Burgmüller excerpts did not always sound like natural student mistakes (with the lower rated example receiving an average rating of 2.67), mainly due to the confidence demonstrated by the performer even when committing and recovering from mistake. To them, this behaviour rather sounded as if an advanced performer was deliberately inducing mistakes in their performance.

During the test, the teachers expressed a difficulty of decoupling between evaluating the naturalness of a mistake from the unnaturalness of the expression and phrasing compared to their expectation of a student performance, worrying that a suspicion that an excerpt is synthetically rendered might affect their judgement. However, the two examples of human-recorded performances with synthetic mistakes do not have consistently higher ratings than the rendered ones,

7.2 Part 2: Open Discussion on Mistake Taxonomy and Common Student Error Patterns

In general, the teachers found the proposed mistake categories sensible. As highlighted by their responses to the questionnaire, the recovery from a mistake significantly affected whether an excerpt was perceived as natural or artificial, which validated our dedication of a separate *Recovery* category for the mid-level mistakes. However, they noted that it was challenging for them to think in terms the exact deviations constituting a mistake, because their teaching typically focuses on demonstrating how to improve an incorrectly performed region through highlighting correct techniques, articulations, and reinforcing higher-level musical concepts like phrasing and dynamic contrasts.

Additionally, they emphasized the importance of score agreement in early learning stages where students might be unaware of their reading errors. We asked them to mark mistake-prone regions on beginner scores, and they highlighted areas that included altered accidentals, changes of tempo markings or note durations and rests, as shown in Fig 7.2. They also mentioned that the playing context influences mistake expectations, with a higher presence of rollback behavior during practice sessions due to repeated attempts to correct fragments, as opposed to other situations where the focus is on continuity, like run-throughs or recitals.

Finally, they provided examples where student attitudes or personalities affected their mistakes and performance habits. Some students might demonstrate an insistent character (hammer the keys really hard as they try to to fix the mistakes, reinforce the correct playing with high-energy, show few hesitations); or be 'unsure' (very insecure about the note, produce ghost note that hits the key very slightly); or 'rushing' (struggles to keep the tempo or rhythmic flow, frequently resulting in mistakes up when playing with two hands). Furthermore, they noted the effect of player age (adults vs children) on rollback behaviour, and the difference between the mistakes made by advanced players on familiar pieces and by students on newly-learned pieces. These observations connect with those in Section 2.1.3, both suggesting that for our pipeline to create coherent full simulated performances it would require connecting between the target behaviour, its effect on the mistake parameter values, and the connections between consecutive mistakes.

8. CONCLUSION AND FUTURE WORK

In this paper, we introduced a hierarchical taxonomy for categorizing performance mistakes based on the analysis of the *Burgmüller* and *Expert-Novice* datasets, and developed a pipeline for simulating these mistakes on error-free piano MIDI files. We validated our taxonomy with music teachers and gained their support. Looking ahead, we aim to analyze more performances to better understand the correlations between different types of mistakes and their parameters, and explore how various musical elements influence these errors. This research marks an initial step towards creating a comprehensive dataset for developing machine-learning applications in music education.

Acknowledgments

We would like to thank our participating music teachers: Alexander von Oppenbach, Davide Ugalde, and Pablo Puentes for their invaluable feedback and enthusiasm. This research was partially supported by the project Musical AI - PID2019-111403GB-I00/AEI/10.13039/501100011033, funded by the Spanish Ministerio de Ciencia e Innovación and the Agencia Estatal de Investigación, and the UKRI Centre for Doctoral Training in Artificial Intelligence and Music (grant number EP/S022694/1).

9. REFERENCES

[1] B. H. Repp, "The Art of Inaccuracy: Why Pianists' Errors Are Difficult to Hear," *Music Perception: An*

Interdisciplinary Journal, vol. 14, no. 2, pp. 161–183, 1996.

[2] B. Gingras, C. Palmer, P. N. Schubert, and S. McAdams, "Influence of Melodic Emphasis, Texture, Salience, and Performer Individuality on Performance Errors," *Psychology of Music*, vol. 44, p. 847–863, 2016.

[3] E. Benetos, A. Klapuri, and S. Dixon, "Score-informed Transcription for Automatic Piano Tutoring," in *European Signal Processing Conference*, 2012.

[4] C. Palmer and C. van de Sande, "Units of Knowledge in Music Performance," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 19, no. 2, pp. 457–470, 1993.

[5] E. Nakamura, K. Yoshii, and H. Katayose, "Performance Error Detection and Post-Processing for Fast and Accurate Symbolic Music Alignment," in *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, 2017.

[6] S. Flossmann, W. Goebl, and G. Widmer, "The Magaloff Corpus: An Empirical Error Study," in *Proceedings of the 11th International Conference on Music Perception and Cognition (ICMPC)*, 2010.

[7] C. Palmer and C. van de Sande, "Range of Planning in Music Performance," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 21, p. 947–962, 1995.

[8] H. Gudmundsdottir, "Pitch Error Analysis of Young Piano Students' Music Reading Performances," *International Journal of Music Education*, vol. 28, pp. 61–70, 2010.

[9] Y. Jiang, F. Ryan, D. Cartledge, and C. Raphael, "Offline Score Alignment for Realistic Music Practice," in *Proceedings of the 16th Sound and Music Computing Conference (SMC)*, 2019, pp. 387–393.

[10] T. Fukuda, Y. Ikemiya, K. Itoyama, and K. Yoshii, "A Score-Informed Piano Tutoring System with Mistake Detection and Score Simplification," in *Proceedings of the 12th International Conference in Sound and Music Computing (SMC)*, 2015.

[11] Q. Kong, B. Li, X. Song, Y. Wan, and Y. Wang, "High-Resolution Piano Transcription with Pedals by Regressing Onset and Offset Times," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 29, pp. 3707–3717, 2021.

[12] H. Kim and X. Serra, "DiffVel : Note-Level MIDI Velocity Estimation for Piano Performance by A Double Conditioned Diffusion Model," in *16th International Symposium on Computer Music Multidisciplinary Research (CMMR)*, Tokyo, Japan, 2023.

[13] A. Morsi, K. Tatsumi, A. Maezawa, T. Fujishima, and X. Serra, "Sounds Out of Pläce? Score-Independent Detection of Conspicuous Mistakes in Piano Performances," in *Proceeding of the 24th International Society on Music Information Retrieval (ISMIR)*, 2023.

[14] S. Kruse-Weber and R. Parncutt, "Error Management for Musicians: an Interdisciplinary Conceptual Framework," *Frontiers in Psychology*, vol. 5, 2014.

[15] C. Palmer, "Music Performance," Tech. Rep., 1997. [Online]. Available: www.annualreviews.org

[16] Y. Morijiri, O. Satoshi, A. Maezawa, and T. Fujishima, "Understanding the challenges for adult beginners at piano practice from an analysis of errors," *Proceedings of the 13th Asia-Pacific Symposium for Music Education Research, Virtual Conference*, vol. 28, 2021.

[17] 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.

[18] S. Flossmann and G. Widmer, "Toward a Model of Performance Errors: A Qualitative Review of Magaloff's Chopin," in *International Symposium on Performance Science*, 2011.

[19] Y. Jiang, "Expert and Novice Evaluations of Piano Performances : Criteria for Computer-Aided Feedback," in *Proceeding of the 24th International Society on Music Information Retrieval (ISMIR)*, 2023.

[20] C. Cancino-Chacón, S. D. Peter, E. Karystinaios, F. Foscarin, M. Grachten, and G. Widmer, "Partitura: A Python Package for Symbolic Music Processing," in *Proceedings of the Music Encoding Conference (MEC)*, Halifax, Canada, 2022.

[21] D. Jeong, T. Kwon, Y. Kim, K. Lee, and J. Nam, "VirtuosoNet: A Hierarchical RNN-based System for Modeling Expressive Piano Performance," in *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, Delft, Netherlands, 2019.

[22] W. Goebl. (1999) The Vienna 4x22 Piano Corpus. http://dx.doi.org/10.21939/4X22.

[23] V. Konz, W. Bogler, and V. Arifi-M, "Saarland Music Data," *Late-Breaking and Demo Session of the International Society on Music Information Retrieval (ISMIR)*, 2011.

[24] M. Manus, A. V. Lethco, and W. A. Palmer, *Alfred's Basic Adult All-in-One Course*. Alfred Music, 1994.

[25] N. Faber and R. Faber, *Faber Piano Adventures*. Faber Piano Adventures, 1996.

[26] F. Foscarin, A. Mcleod, P. Rigaux, F. Jacquemard, and M. Sakai, "ASAP : A Dataset of Aligned Scores and Performances for Piano Transcription," in *Proceedings of the 21st International Society for Music Information Retrieval Conference (ISMIR)*, 2020.