# SOUNDS OUT OF PLÄCE? SCORE-INDEPENDENT DETECTION OF CONSPICUOUS MISTAKES IN PIANO PERFORMANCES

**Alia Morsi**[1] **Kana Tatsumi**[2] **Akira Maezawa**[3] **Takuya Fujishima**[3] **Xavier Serra**[1]

[1] Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain

[2] Nagoya Institute of Technology, Nagoya, Japan

[3] Yamaha Corporation, Hamamatsu, Japan

## ABSTRACT

In piano performance, some mistakes stand out to listeners, whereas others may go unnoticed. Former research concluded that the salience of mistakes depended on factors including their contextual appropriateness and a listener's degree of familiarity to what is being performed. A *conspicuous* error is considered to be an area where there is something *obviously* wrong with the performance, which a listener can detect regardless of their degree of knowledge of what is being performed. Analogously, this paper attempts to build a score-independent conspicuous error detector for standard piano repertoire of beginner to intermediate students. We gather three qualitatively different piano playing MIDI data: (1) 103 sight-reading sessions for beginning and intermediate adult pianists with formal music training, (2) 245 performances by presumably late-beginner to early-advanced pianists on a digital piano, and (3) 50 etude performances by an advanced pianist. The data was annotated at the regions considered to contain conspicuous mistakes. Then, we use a Temporal Convolutional Network to detect the sites of such mistakes from the piano roll. We investigate the use of two pre-training methods to overcome data scarcity: (1) synthetic data with procedurally-generated mistakes, and (2) training a part of the model as a piano roll auto-encoder. Experimental evaluation shows that the TCN performs at an F-measure of 0.78 without pretraining for sight-reading data, but the proposed pretraining steps improve the F-measure on performance and etude data, approaching the agreement between human raters on conspicuous error labels. Importantly, we report on the lessons learned from this pilot study, and what should be addressed to continue this research direction.

## 1. INTRODUCTION

A commonly held notion in automatic music performance analysis (MPA) research is that deviations of music performances from their underlying music score can be regarded as performance mistakes. But previous music pedagogy research suggests that some of such deviations are more apparent to a listener than others [1, 2]. For example, a chord that is voiced differently from that written in the score might be overlooked, but missing a note in a characteristic motif or playing a note that clashes with the underlying harmony would stand out. Repp [1] referred to errors of the former category as *perceptually inconspicuous*. Accordingly, we consider a **conspicuous error** to be "a performance error that can be detected by the majority of listeners with a formal music training, regardless of their degree of knowledge about the underlying music score of a performed piece."

This paper explores the potential of building score-independent models that detect regions of *conspicuous errors* in MIDI piano performances of piano solo pieces based on Western music theory, as shown conceptually in Figure 1. Based on the intuition that a listener is capable of detecting obvious mistakes in piano performances by listening to the surrounding context, we use a non-causal variant of the Temporal Convolutional Network (TCN) [3] We gather datasets for our task, since despite the plethora of work in automatic MPA that has spanned both the score-dependent (or reference-dependent) [4–7] and score-independent paradigms [8–13], there is no data available to support our desired goal.

More specifically we: (1) gather three datasets of conspicuous errors in various performance situations, reporting on the dataset creation process and annotation procedure, (2) study the properties of the annotated data through (i) observing the annotated data for sources of inconsistencies, (ii) analyzing the relationship between inconspicuous and conspicuous errors and (ii) analyzing the ambiguity of the task through listening experiments, (3) present a model based on TCN to identify conspicuous errors from piano MIDI performance and discuss its effectiveness through experimental evaluation, and (4) present and evaluate two pre-training strategies, depending on the nature of the unlabeled data that can be acquired. A subset of the gathered data and listening examples can be found on the companion page [1]

## 2. RELATED WORK

We distinguish between locally and globally-based automatic MPA. In local approaches (such as the majority

© A. Morsi, K. Tatsumi, A. Maezawa, T. Fujishima, and X. Serra. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** 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 *Proc. of the 24rd Int. Society for Music Information Retrieval Conf.,* Milano, Italy, 2023.

[1] https://bit.ly/3UCCiea

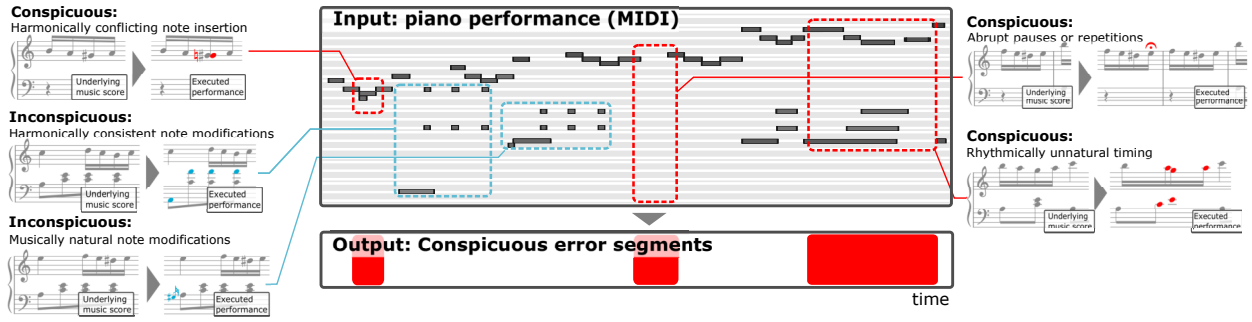



**Figure 1**: Illustration of our problem definition. Some errors stand out more than others in performance. Our goal is to identify segments containing conspicuous errors to the listeners, without the need for music score data.

of score-dependent performance assessment), the analysis is conducted at a note (or equivalent) level. Global approaches learn from data mapping large performance snippets (often entire performances) to overall evaluations.

Local approaches include score-based performance mistake identification, which tends to cover note-level (or equivalent) errors such as pitch [1, 2, 4, 7] and rhythm mistakes [2]. Pitch mistakes are essentially categorized as ***pitch intrusions*** (extra note) and ***pitch omissions*** (missing note), and occasionally ***pitch substitutions*** (wrong note in-place of a correct one), although the latter can be treated the joint occurrence of the former two [1]. Alignment/score comparison-based approaches for detecting deviations are locally-based by definition. Piano assessment examples of such include [4, 7, 14], which cover pitch mistakes. Not all local approaches are score-dependent, such as those which capture note-level aspects relating to the articulation or sound quality. Examples are [15] and [12], for piano (3-point scale for quality of legato or staccato) and trumpet (7-point scale) respectively.

Global approaches to performance assessment have usually been score-free, with the exception of [5] which utilizes the score as input. Usually, such approaches are based on regression models mapping features to performance-wide ratings [9, 11, 16, 17], or end-to-end approaches which learn correspondences between whole or parts of performances to performance wide ratings [5, 10, 13]. Such ratings can be discrete or continuous and can span several performance dimensions. Although the connection has not been explicitly made, we speculate that most likely they would excel in capturing conspicuous performance mistakes that manifest as consistent errors/error patterns across a performance.

Accordingly, we frame our approach as a score-independent locally based one since our goal is to return binary labels for each time point in a piano MIDI roll reflecting the presence or absence of an obvious performance mistake. Therefore, we need similarly annotated data for piano MIDI performances to train our models. Despite score deviations not necessarily indicating conspicuous errors, our desired output is closest to that of score-based performance mistake identification systems because their output can be interpreted as a binary sequence indicating the presence or absence of a score deviation albeit without perceptual relevance. However, their methods are not applicable for our problem formulation.

## 3. DATA

We obtain 3 sources of non-commercial, piano MIDI performance data for different playing situations:
**Sight-Reading Data (SR)**: 103 sight-reading performances comprising mostly of piano reductions of popular classical pieces, arranged for beginner to intermediate difficulty. They are played by seven beginning to intermediate adult pianists with formal music training.
**Performance Data (PF)**: 245 performances of approximately 3 minutes each, collected from a digital piano recording app. Not all performed pieces are known, but most of them are pop and classical, that are either read from a score, or semi-improvised. While user attributes are unknown, the performance data suggests that the skill levels range between late-beginner and early-advanced.
**Burgmüller Data (BM)**: 50 performances from Burgmüller's 25 Etudes, Op. 100 recorded twice on a digital piano. They are played by an advanced pianist who had previously played the etudes. The pianist practiced each etude briefly before recording two takes.

The total time for the **SR**, **PF**, and **BM** are 379, 723, and 60 minutes respectively, of which 128, 176, and 3 minutes were annotated as conspicuous errors. Non-overlapping splits of **SR** and **PF** are used for training, validation, and testing, whereas **BM** is kept exclusively for testing. The annotation procedure is described in 3.1. **SR** and **PF** subsets cannot be shared, but short excerpts of them, and the full **BM** set can be found in the companion page.

### 3.1 Annotation Procedure

We had 2 annotators: ***Annotator 1***, who has experience as a classical piano teacher, and ***Annotator 2***, has training in music production and is also an intermediate-level pianist. We asked Annotator 1 to label the **SR** and **BM** data, and asked Annotator 2 to label the **PF** data, and to indicate (yes/no) whether they know the piece being performed. For the **SR** and **PF** subsets, annotators were given instructions to annotate *obvious* performance mistakes that can be recognized even without checking the score, and it was left open to them to decide what that entails. The an-

notation was done with Cubase [2], and they were asked to add an annotation at MIDI note 0 covering the span of the time window which they judge as pertaining to an error. Despite the potential label ambiguity due to the openness of the instructions, we wanted to observe the judgments of different people in this pilot study so that we can improve the data annotation protocol for future experiments.

The **BM** subset was treated differently because it has been played off of known music score data. First, the performances were automatically annotated with sites of score deviations using a score alignment system. Then, the annotator manually reviewed the labels by listening to the performance while looking at the corresponding sheet music, and added missing deviations from the score or removed those which do not reflect errors. The annotator simultaneously manually labeled each error as conspicuous or not.

### 3.2 Annotation Examples and Pitfalls

Some types of errors were labeled more consistently than others. The more common error modes, as shown in Figure 2, include insertions and deletions of notes that do not fit in musical context, abrupt pauses, and unstable rhythm coming from hesitations during playing. Annotators have shown reasonable consistency in terms of label location and span when mistakes are relatively short after which the player recovers into their playing flow, such as those of Figure 2. However, more compound deviations were labelled ambiguously. For example, sometimes after an error a player would 'sneak-in' some practice before resuming the flow of the piece. In such examples, if the short phrase being practiced sounds out of context, but in itself is coherent, an open question is where the label should be, and whether it should be one continuous label or an intermittent one.

Moreover, we also observe the presence of non-annotated conspicuous mistakes in the data, but there is an inherent ambiguity in how one would assess a "bad but acceptable" and "erroneous" performance". In a discussion with Annotator 1 after the annotations, they indicated that their mental model for deciding whether a segment should be labelled was dependent on every performance. If a region contrasts with their expectation of the music given how that performer is playing, then it was annotated. This opens the possibility that annotators have calibrated what should count as a mistake based on individual performance. Silence regions are one of the main sources of ambiguity, since silences between correct portions are non-annotated regardless of their length, but silences within or surrounding mistake portions often receive a mistake label.

### 3.3 Analysis of the dataset

#### *3.3.1 Conspicuous to total label ratio in **BM***

Although the ratio of annotated regions to total performance time is very small in the **BM** data, its annotation approach of allows us to investigate the relationship between the set of errors obtained by comparing with a score

[2] https://www.steinberg.net/cubase/

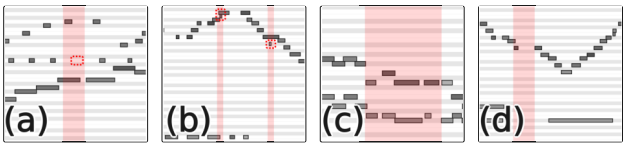



**Figure 2**: Examples of musical attributes that seemed to be consistently annotated as conspicuous errors (in red). (a) missed note that breaks a pattern, (b) harmonically unnatural note insertions, (c) repetition, (d) abrupt pauses.

(presumably all errors) to conspicuous errors. We found that 59% of all identified errors were perceived as conspicuous. Note that this is a very subset-specific result, because it depends on the ratio between subtle and obvious errors in the performances themselves as much as the qualities of the performer and the annotation.

#### *3.3.2 Listening test of conspicuous errors*

Through a listening test of some performance portions labeled as conspicuous errors and unlabeled areas for **PF**, we assess how different subjects agree with the annotations and among themselves. We chose **PF** because we expect it to contain a nice balance between famous and unknown pieces for each subject.

**Conditions**: We recruited 31 subjects, not necessarily trained musicians. 84% of the subjects had experience playing a musical instrument, and 97% of the subjects had experienced either reading or notating music scores. Each subject is asked to first listen with headphones to a snippet from the **PF** dataset, ranging from 4 to 12 seconds. The snippet is either (1) a randomly chosen conspicuous mistake segment, with 2 seconds of padding on either end, or (2) a segment that contains no error label, whose duration is the average duration of the conspicuous error segments within the piece, plus two seconds of padding. The subjects were allowed to skip questions and no constraints were given on the number of times the snippet may be listened to. The subject is then asked to choose if they hear an obvious mistake or not, along with the subject's knowledge of the piece. This procedure was repeated 15 times. Then, we scale the counts obtained when presenting non-conspicuous snippets, to provide a sensible assessment of the dataset itself. That is, the ratio of snippets containing the inconspicuous error to the conspicuous ones, $\rho_0$, should match the ratio between the total duration of the inconspicuous error labels to that of the conspicuous labels in the dataset, $\rho_1$. Thus, we scale the count of the responses obtained when presenting the inconspicuous error by $\rho_1/\rho_0$.

**Results and discussion**: A total of 462 responses were obtained (30-31 responses per snippet). The precision, recall, and the F-measure of how correctly the subjects identified the mistakes were 0.37, 0.50, and 0.43, respectively. The result suggests that the notion of conspicuous error is not so clear-cut when only presenting a short snippet surrounding an error, without providing a longer musical context. We also found that famous pieces tend to get more consistent responses. To check this, we computed for each

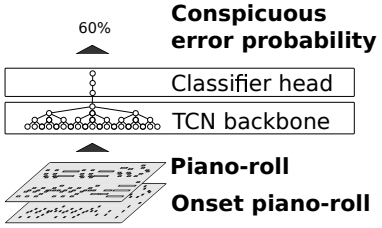



**Figure 3**: Our method reads a piano roll and outputs the probability of the center of a segment being a conspicuous error. It is comprised of a TCN backbone and a 1d convolution classifier head.

snippet (1) the probability that a song is unknown and (2) the entropy of the probability that a subject would identify that snippet to contain an error. The correlation between (1) and (2) was 0.63, indicating a moderate correlation between how well the piece is known among the subjects and how consistent are the labels.

## 4. METHODOLOGY

Given a sequence of piano note events, the goal is to infer a time sequence of binary labels that indicates the presence of conspicuous errors at a given time.

### 4.1 Model

Our model is a TCN-based network that receives a piano roll $\mathbf{X}$ as input and emits a binary label of conspicuous error $e$ at each time frame of the piano roll. As shown in Figure 3, it is comprised of a feature extraction backbone followed by a classification head. We choose to assign a label at *frame-level* instead of note-level, since not only the note itself but its absence can indicate errors.

#### *4.1.1 Piano Roll Input*

Two piano rolls are extracted for a given sequence of piano note events, one for the note onset and another for the sustained portion according to the key depression. Specifically, suppose a set of $I$ MIDI note events (start time, end time, pitch, velocity) given as $\{(s_i, e_i, p_i, v_i)\}_i^I$, and a sampling rate of $R$ are given. Then, a 256-dimensional piano roll $X \in \mathbb{R}^{256\times T}$ is computed, such that $X(p_i, \text{round}(Rs_i)) = v_i$, and $X(128 + p_i, \text{round}(Rs)) = v_i$ for $s \in [s_i, e_i]$. Partitura [18] is used for the computation, and $R$ is set to 16 Hz.

Notice that the sustain pedal information is ignored in the computation of the piano roll. This is necessary to prevent the piano roll of the sustained portion from smearing since a beginning pianist has a tendency to keep the pedal depressed which causes and excessive elongation of the computed note durations.

#### *4.1.2 Conspicuous mistake detector*

We model the mistake detector as a simple TCN comprising of a feature extraction backbone followed by a classification head, based on preliminary experiments exploring model architectures and inspired by the approach in [13].

**Feature extraction backbone**: Given the piano roll $X$, the feature extraction backbone computes a feature $\phi \in \mathbb{R}^{D\times T}$. We set $D = 256$ in this paper. This is realized as a 5-layer noncausal TCN with dilation of [1,2,4,8,16], and for all layers, has an output channel size of 256, kernel size of 3, uses ELU nonlinearity and has a residual connection, similar in spirit to [3].

**Classification head**: Given the feature $\phi$, a network comprising of three layers of 1x1 convolution with output channel sizes [256,64,1] with residual connections and ELU nonlinearity followed by a sigmoid function is used to arrive at the conspicuous error posterior probability $e$.

### 4.2 Training strategies

The model is trained using RAdam with a learning rate of $10^{-3}$, as to minimize the cross-entropy between the conspicuous error probability $e$ and the posterior distribution computed from the ground-truth label. We augment the data by randomly transposing the entire MIDI file in the training data. Furthermore, when computing the cross-entropy loss, we smooth the ground-truth label to account for annotation inconsistencies in the start and end times of the conspicuous error segment. Furthermore, since it is difficult to obtain annotations of conspicuous errors, we pre-train the model as well, using the following two strategies.

#### *4.2.1 Pretraining the feature extractor as an autoencoder*

The feature extractor can be trained in an unsupervised manner, by training it as an autoencoder for a much larger collection of piano performances in the wild. Specifically, we train an auto-encoder using the feature extraction TCN introduced earlier as the encoder and a TCN with transposed 1d convolutions instead of a 1d convolution as the decoder. This way, the space of $\phi$ is pre-trained as to model the space of piano performances within a given receptive field of a TCN. This method could be useful if a large dataset of performances of unknown performance qualities are obtainable.

#### *4.2.2 Pretraining the model with synthetic mistake labels*

The model can also be pre-trained on performance data onto which mistakes are simulated and corresponding mistake labels are inserted to match the expected format of data in Section 3.1. Specifically, we apply systematic adjustments to a set of mistake-free performances and modify the note events, in a manner inspired by performance mistakes made by beginning adult pianists [19]. For each note event, with probability $p_c$ we modify the note in one of the following ways:

1. With probability $p_o$ omit a note with a probability $p_o$
2. With probability $p_r$ replace a note, to the same note transposed $n$ semitones, to simulate hitting the wrong key.
3. With probability $p_i$ insert a note that is transposed by $n$ semitones.

| Method | Precision | Recall | F-measure |
|---|---|---|---|
| Baseline | **0.79** | **0.80** | **0.78** |
| SYNTH | 0.65 | 0.76 | 0.69 |
| SYNTH(FT) | 0.61 | 0.69 | 0.62 |
| AE | 0.55 | 0.59 | 0.55 |
| AE+SYNTH | 0.44 | 0.65 | 0.51 |

(a) **SR** Data

| Method | Precision | Recall | F-measure |
|---|---|---|---|
| Baseline | 0.28 | 0.46 | 0.33 |
| SYNTH | 0.27 | 0.54 | 0.34 |
| SYNTH(FT) | **0.30** | 0.61 | **0.38** |
| AE | 0.28 | 0.52 | 0.34 |
| AE+SYNTH | 0.27 | **0.63** | 0.36 |

(b) **PF** Data

| Method | Precision | Recall | F-measure |
|---|---|---|---|
| Baseline | 0.26 | 0.36 | 0.26 |
| SYNTH | 0.26 | **0.69** | 0.35 |
| SYNTH(FT) | 0.26 | 0.49 | 0.32 |
| AE | 0.27 | 0.46 | 0.31 |
| AE+SYNTH | **0.28** | 0.52 | **0.35** |

(c) **BM** Data

**Table 1**: Results for different training strategies

4. With probability $p_p$ pause the performance by a small amount distributed uniformly between 0.3 and 0.8 seconds. With probability $p_{pr}$, repeat the last played note.
5. With probability $p_s$ pause the performance by a large amount distributed uniformly between 2 and 4 seconds. Repeat the last played note.

In this paper, we set $p_c = 5\%$, $p_o = 10\%$, $p_i = 39\%$, $p_r = 39\%$, $p_s = 2\%$, and $p_p = 10\%$. Furthermore, for note replacement and insertion, $n$ is chosen so that $n = 1, 2$ are chosen with probabilities of 22% and $n = 4, 6$ by 2%. For a set of mistake-free performances, we obtained 260 hours of mostly jazz and classical MIDI piano performances. The quality and repertoire are comparable to those available from Yamaha PianoSoft [3] .

This method is useful if many performances that are known to be relatively error-free are obtainable Furthermore, this idea may possibly be used for data augmentation, at the risk of increasing false positives, since not all synthetic errors sound conspicuous, as also hinted by [1,2].

### 4.3 Experiment: Model Evaluation

We evaluate our model using different training strategies.

#### *4.3.1 Experimental conditions*

Our model has been trained with the following strategies:

1. Baseline - The model is trained on **SR** and **PF** data.
2. SYNTH - Same as Baseline, in addition to the inclusion of a subset of the synthetic data introduced in Section 4.2.2 during training and validation.
3. SYNTH(FT) - The model is pretrained on the synthetic data, then fine-tuned using **SR** and **PF**. This simulates a situation where a new annotated dataset becomes available after traing a model solely trained on a synthetic data.
4. AE - Train TCN autoencoder introduced in Section 4.2.1 as a pretraining step for the backbone TCN, using approximately 100,000 MIDI performances played by various users. The set of performances does not contain **SR PF** or **BM**, although it is obtained from the same source as **PF**. The model is fine-tuned on **SR** and **PF**.
5. AE+SYNTH - Use the pretrained autoencoder backbone and fine-tune using **SR**, **PF** and the synthetic data.

The trained models have been validated on **SR** and **PF**, and tested on a test split of **SR**, **PF**, and the entire **BM**.

As the metric, we have evaluated the transcription precision/recall/F1-measure using `mir_eval` [20], treating the estimated and the ground-truth annotations as note events occurring at a predefined pitch. When computing the transcription metrics, the note onset and offset tolerances have been set to 2 seconds. Furthermore, based on the validation set, the ends of the estimated segments have been padded by 0.2 seconds and overlapping segments have been merged.

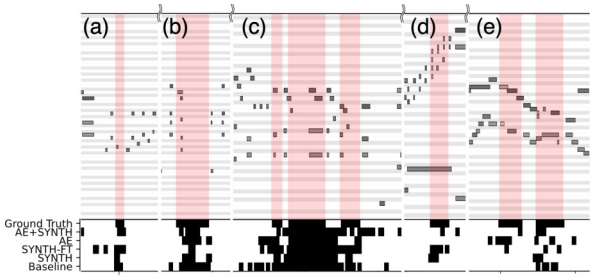
#### *4.3.2 Results and discussion*

The results are shown in Table 1. For **PF** and **BM** datasets, the augmentation strategies offer some improvements. The two strategies proposed, i.e., the use of synthetic data and autoencoder, also result in improvements. In general, both strategies tend to improve the recall rate, suggesting that they provide similar qualitative improvements, and either one can be used depending on the data available.

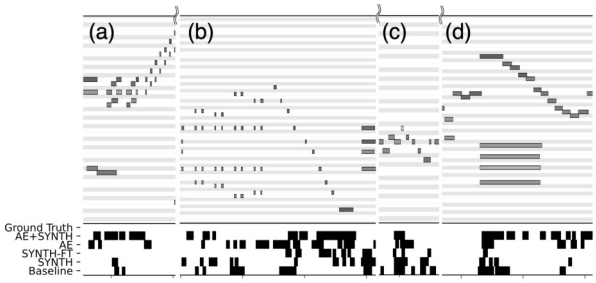
Despite the augmentation strategies, the F-measures for **PF** and **BM** data suggest future room for improvement, even taking into account the ambiguity of conspicuous errors. The **PF** and **BM** data are difficult to infer, as seen by the differences in the F-measure between the **SR** dataset and the two. As another example, the validation F-measure of the models on the synthetic dataset is about 0.60. This suggests that the model is moderately capable of pin-pointing the ground-truth labels if they are easy to classify, or generated stochastically but systematically. At the same time, however, the model has room for improvement, as the best-performing F-measure of 0.38 on the **PF** dataset falls somewhat short of the oracle F-measure of 0.43, as discussed in Section 3.3.

The method performs well for the **SR** data, perhaps because most of the mistakes are quite conspicuous in a sight-reading situation, especially compared to **PF** and **BM**, both of which contain mostly beginner-intermediate performances with occasional mistakes. The performance tends to drop as more pretraining steps are added, presumably because the pretraining data mostly contain data of the same type as the **PF** set, increasing the disparity between the training data and the test data. In sight-reading situations, the results suggest it is sufficient simply to train on a

[3] `https://shop.usa.yamaha.com/`




(a) True positives. The black band indicates the detected conspicuous error with different training strategies. The model presumably responds to (a) repetition, (b) silences, (c) slight hesitations in playing, (d) note insertions, and (e) lack of synchrony voices.




(b) False positives. The model presumably confuses (a) the repeated motives as an error, (b) rhythm with rest as abrupt pauses, (c) an audible but weak note with a note deletion, and (d) a long chord after a fast passage with hesitations.

**Figure 4**: Examples of typical operation and failure modes.

dataset that solely contains data from the same set, instead of pretraining or augmenting the dataset with typical amateur performances containing some conspicuous errors.

#### *4.3.3 Qualitative insights of the estimates*

Figure 4 shows some examples of true positives that are consistent across different strategies and consistent false positives. The proposed method tends to capture repetition, pauses, hesitations, and note insertions that occur in narrow pitch intervals as mistakes. At the same time, however, the very same properties arising from musical expression or composition are detected as false positives, such as repeated motifs, ornaments, and grand pauses. Even though such musical aspects are superficially performed similarly to the aforementioned mistakes, humans are capable of differentiating between genuine performance mistakes and those within musical contexts. This suggests that the model has room to improve by modeling the underlying composition better. The readers are invited to check the companion page for examples.

## 5. LIMITATIONS AND IMPROVEMENTS

Our work opens door to many open problems that need to be solved, some more fundamental than others.

**Problem definition and annotation protocol**: More work is needed to define the concept of conspicuous errors, and how the task should be evaluated from a music technology perspective. Accordingly, a more comprehensive protocol for data collection should be developed. Although we had kept the annotation instructions open to also develop an understanding of annotator behavior, it became evident that our data collection approach does not guarantee that the labels we have are for solely conspicuous errors. In [1], conspicuous errors were identified in a music performance by finding the subset of agreed-upon mistake labels between multiple listening subjects.

To define manifestations of conspicuous errors, a midpoint should be found between a rule-based approach and one learned from empirical labels. The outcome should be a set of error descriptions, some of which happen at particular time instants and some over longer windows, whether continuous windows or a longer span of intermittent labels. However, since the conspicuousness of errors is inspired by a perceptual idea, we think these errors should be defined through an empirical process albeit better defined than the one in this study to avoid the same pitfalls.

**Synthetic mistakes**: Synthetic data is important for improving performance, but current synthesized mistakes sound unnatural. A simple example was a case of induced pitch insertions, where it seemed impossible that someone can perform with such confidence and tempo despite the extent of out-of-context pitch insertions. We observe that beginners make mistakes and employ recovery strategies in a manner that is more complex than the presented method, so a better understanding of beginning pianists' behavior is necessary to create more natural-sounding mistakes.

**Listener, player, expression, and style**: Conspicuous errors are dependent on the listener's knowledge of the piece and the proficiency of the performer. Furthermore, conspicuous error and expression are two sides of the same coin. For example, hitting an adjacent key can either come across as an expressive ornament or a conspicuous error. This suggests that conspicuous error detection should inherently be conditioned on the style, the level of the listener, and the player's proficiency.

**Connecting with pedagogy and edu-tainment**: The impact of music education software which provides analysis solely founded on rigid note-level rhythmic and pitch correctness has been challenged [21] on the basis that users might end up too focused on playing too correctly (almost robotically) to attain the highest scores. There are many pedagogical considerations for designing useful automatic assessments [22].

## 6. CONCLUSION

This paper presented a study on detecting conspicuous performance mistakes for a piano solo performance of beginning to intermediate players. We (1) clarified the idea of a *conspicuous* error in line with previous research, (2) gathered locally annotated piano MIDI performance data, and (3) discussed sources of inconsistencies in our data through analysis of the annotation procedure and subjective tests. Although some of our models show an acceptable performance on the test split of the **SR** data subset, we find that the our pre-training suggestions do not provide remarkable performance improvements.

## 7. REFERENCES


[1] B. H. Repp, "The art of inaccuracy: Why pianists' errors are difficult to hear," *Music Perception: An Interdisciplinary Journal*, vol. 14, p. 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] M. E. P. Davies and S. Böck, "Temporal convolutional networks for musical audio beat tracking," in *Proc. European Signal Processing Conference (EUSIPCO)*, September 2019.

[4] E. Nakamura, K. Yoshi, and H. Katayose, "Performance error detection and post-processing for fast and accurate symbolic music alignment," in *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, October 2017, pp. 347–353.

[5] J. Huang, Y. N. Hung, K. A. Pati, S. Gururani, and A. Lerch, "Score-informed networks for music performance assessment," in *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, 2020.

[6] H. Zhang and Y. a. Jiang, "Learn by referencing: Towards deep metric learning for singing assessment," in *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, 2021.

[7] E. Benetos, A. Klapuri, and S. Dixon, "Score-informed transcription for automatic piano tutoring," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2012, pp. 2153–2157.

[8] T. Nakano, M. Goto, and Y. Hiraga, "An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features," in *Proc. International Conference on Spoken Language Processing (ICSLP)*, September 2006, p. 1706–1709.

[9] C. W. Wu, S. Gururani, C. Laguna, A. Pati, A. Vidwans, and A. Lerch, "Towards the objective assessment of music performances," in *Proc. International Conference on Music Perception and Cognition (ICMPC)*, July 2016.

[10] K. A. Pati, S. Gururani, and A. Lerch, "Assessment of student music performances using deep neural networks," *Journal of Applied Sciences*, vol. 8, no. 4, 2018. [Online]. Available: https://www.mdpi.com/2076-3417/8/4/507

[11] J. Abeßer, J. Hasselhorn, S. Grollmisch, C. Dittmar, and A. Lehmann, "Automatic competency assessment of rhythm performances of ninth-grade and tenth-grade pupils," in *Joint Proc. International Computer Music Conference (ICMC), and Sound and Music Computing Conference (SMC)*, September 2014, pp. 1252–1256.

[12] T. Knight, F. Uphamm, and I. Fujinaga, "The potential for automatic assessment of trumpet tone quality," in *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, 2011.

[13] P. Seshadri and A. Lerch, "Improving music performance assessment with contrastive learning," in *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, November 2021.

[14] T. Fukuda, Y. Ikemiya, K. Itoyama, and K. Yoshii, "A score-informed piano tutoring system with mistake detection and score simplification." Proc. Sound and Music Computing Conference (SMC), Jul 2015.

[15] V. Phanichraksaphong and W.-H. Tsai, "Automatic evaluation of piano performances for steam education," *Applied Sciences*, vol. 11, no. 24, 2021. [Online]. Available: https://www.mdpi.com/2076-3417/11/24/11783

[16] J. Abeßer, J. Hasselhorn, S. Grollmisch, C. Dittmar, and A. Lehmann, "Automatic quality assessment of vocal and instrumental performances of ninth-grade and tenth-grade pupils," in *Proc. International Symposium on Computer Music Multidisciplinary Research (CMMR)*, 2013.

[17] J. Pan, M. Li, Z. Song, X. Li, X. Liu, H. Yi, and M. Zhu, "An Audio Based Piano Performance Evaluation Method Using Deep Neural Network Based Acoustic Modeling," in *Proc. Interspeech 2017*, 2017, pp. 3088–3092.

[18] C. E. Cancino-Chacón, S. D. Peter, E. Karystinaios, F. Foscarin, M. Grachten, and G. Widmer, "Partitura: A Python Package for Symbolic Music Processing," in *Proc. Music Encoding Conference (MEC2022)*, 2022.

[19] Y. Morijiri, S. Obata, A. Maezawa, and T. Fujishima, "Understanding the challenges for adult beginners at piano practice from an analysis of errors," in *Proc. Asia-Pacific Symposium for Music Education Research (APSMER2021)*, 2021.

[20] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, and D. P. W. Ellis, "MIR_EVAL: A Transparent Implementation of Common MIR Metrics," in *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, 2014, pp. 367–372. [Online]. Available: http://dblp.uni-trier.de/db/conf/ismir/ismir2014.html#RaffelMHSNLE14

[21] A. Acquilino and G. Scavone, "Current state and future directions of technologies for music instrument pedagogy," *Frontiers in Psychology*, vol. 13:835609, 2022.

[22] V. Eremenko, A. Morsi, J. Narang, and X. Serra, "Performance assessment technologies for the support of musical instrument learning," in *Proc. International Conference on Computer Supported Education (CSEDU)*, May 2020, pp. 629–640.