



Sounding Out Reconstruction Error-Based Evaluation of Generative Models of Expressive Performance

Silvan David Peter
silvan.peter@jku.at
Johannes Kepler University
Linz, Austria

Carlos Eduardo Cancino-Chacón
carlos_eduardo.cancino_chacon@jku.at
Johannes Kepler University
Linz, Austria

Emmanouil Karystinaios
emmanouil.karystinaios@jku.at
Johannes Kepler University
Linz, Austria

Gerhard Widmer
gerhard.widmer@jku.at
Johannes Kepler University
Linz, Austria

ABSTRACT

Generative models of expressive piano performance are usually assessed by comparing their predictions to a reference human performance. A generative algorithm is taken to be better than competing ones if it produces performances that are closer to a human reference performance. However, expert human performers can (and do) interpret music in different ways, making for different possible references, and quantitative closeness is not necessarily aligned with perceptual similarity, raising concerns about the validity of this evaluation approach. In this work, we present a number of experiments that shed light on this problem. Using precisely measured high-quality performances of classical piano music, we carry out a listening test indicating that listeners can sometimes perceive subtle performance difference that go unnoticed under quantitative evaluation. We further present tests that indicate that such evaluation frameworks show a lot of variability in reliability and validity across different reference performances and pieces. We discuss these results and their implications for quantitative evaluation, and hope to foster a critical appreciation of the uncertainties involved in quantitative assessments of such performances within the wider music information retrieval (MIR) community.

2023), November 10, 2023, Milan, Italy. ACM, New York, NY, USA, 9 pages.
<https://doi.org/10.1145/3625135.3625141>

1 INTRODUCTION

The recent years have seen the creation and publication of several corpora of precisely measured and score-aligned piano performances within MIR and digital musicology communities [17, 23, 29]. This renewed interest in computational models of expressive piano performance, in particular the data-driven kind. Yet it also rekindled concerns surrounding the direct applicability of large scale data processing and machine learning techniques to this type of data.

This paper addresses one such concern, namely issues of quantitatively evaluating generative models of expressive piano performance (GMEPP) at scale. Quantitative evaluation in itself is nothing new, GMEPPs are routinely evaluated in terms of how close their predictions are to actual human expert performances. This closeness is generally estimated with figures of merit such as reconstruction errors [7, 17] or likelihood functions [10, 15, 18].

One issue with this type of evaluation arises from the fact that having a model produce a performance that is *numerically close* (in some aspect yet to be clarified) to an expert piano performance – i.e., a model that does well according to generally accepted figures of merit – possibly misses the mark of GMEPP; the goal of producing convincing, musical, and consistent performances for human listeners. Evaluating by asking such listeners is, however, only an option in a minority of situations, and most of the time the training, development, and evaluation of GMEPP requires scalable, automated metrics. This potential goal misalignment raises at least two problems: is a measured distance to a human reference performance related to the perceptual similarity of performances? And is the choice of an arbitrary human reference performance immaterial for evaluative outcome?

These questions tap into profound epistemic, perceptual, and axiological issues beyond the scope this article. What we can and do address in the following, are three smaller, but nevertheless operatively useful questions about current reconstruction error-based evaluation (REE) techniques:

- Can listeners discern performances that are indistinguishable under REE?
- To what extent does REE reliably favor [performances by] the same model under different reference and piece conditions?

CCS CONCEPTS

• **Applied computing** → **Sound and music computing; Performing arts**; • **Information systems** → **Evaluation of retrieval results**.

KEYWORDS

Performance, Expression, Evaluation, Validity, Listening Study

ACM Reference Format:

Silvan David Peter, Carlos Eduardo Cancino-Chacón, Emmanouil Karystinaios, and Gerhard Widmer. 2023. Sounding Out Reconstruction Error-Based Evaluation of Generative Models of Expressive Performance. In *10th International Conference on Digital Libraries for Musicology (DLfM)*



This work is licensed under a Creative Commons Attribution International 4.0 License.

DLfM 2023, November 10, 2023, Milan, Italy
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0833-6/23/11.
<https://doi.org/10.1145/3625135.3625141>

- To what extent does REE validly identify the [performances by] expert pianists under different reference and piece conditions?

To assess these issues, we set up two experiments. First, a listening test asking participants to identify expert performances in pairs of expert and artificially generated performances. Second, we investigate the reliability and validity of REE evaluation, using the previously assessed artificially generated performances as negatives. We discuss the results of these experiments in the context of the literature on and perception of expressive performance and we identify potential steps to improve quantitative evaluation of GMEPP. With more, larger, and ecologically valid (i.e., stemming from realistic performance scenarios) datasets of expressive piano performance becoming publicly available and used by the wider MIR community, we hope this discussion to foster a critical appreciation of the uncertainties involved in quantitative assessments of such performances.

The rest of this paper is structured as follows: Section 2 details the framework of quantitative GMEPP evaluation as investigated in this article. Section 3 describes how we extract and preprocess expressive parameters from recordings of expressive expert performances. Section 4 describes the performance discernment listening test and section 5 details the reliability and validity experiments. Finally, section 6 discusses these results for evaluation of GMEPP and concludes this article. The audio files, code, and data are available at: https://github.com/CPJKU/performance_similarity_dlfn23.

2 A FRAMEWORK OF QUANTITATIVE EVALUATION

To aid the description of our experiments, we formalize reconstruction error-based evaluation as the following framework, as shown schematically in Figure 1. We consider a *two-model evaluation framework* which asks the question: is performance P1 produced by Model1 "better" than P2 produced by Model2? Concretely, the framework takes a triplet of two performances P1 and P2 of the same piece, one generated by each model, and computes their reconstruction error with respect to (wrt) a third expert RP of the same piece.

The standard evaluative argument of GMEPP is as follows: the model which produced the performance with smaller reconstruction error is favored and its performance is taken to be more musical.

This seemingly overly formal description of a simple and widely used evaluation technique allows us to formulate experiments *about* evaluation by controlling specific elements. Specifically, we use the framework to evaluate *models with controlled ground truth* wrt evaluation, i.e., with models which are known to better or worse. Getting performances of good, musical models is straightforward, any human expert performance can be taken as such. However, finding unmusical performances faces the performance research version of the Anna Karenina principle: all musical performance are (potentially) alike, but all unmusical performances are unmusical in their own way. To mitigate the complexity, we opt for a type of randomization to create unmusical performances.

Before we describe our process to create (and validate our choice of) unmusical performances in Section 2.2, we briefly introduce the numerical representation of expressive performances in Section 2.1.

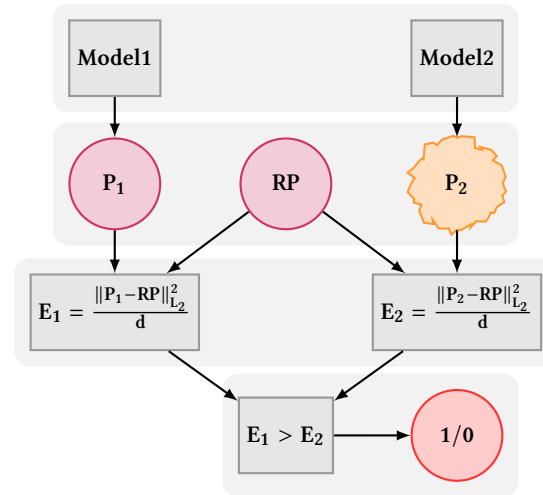


Figure 1: Schematic representation of our framework for two model evaluation. These frameworks are commonly used for the comparison of two or more candidate models of expressive performance. In our experiments, however, the models are specifically designed for their known ground truth wrt evaluation (in the sense discussed in 2): Model 1 only produces expert performances (purple), model 2 only randomly sampled performances (orange), i.e. model 1 is the musically valid one. The two models produce a performance each (P_1 and P_2). The MSE of the performances with respect to an expert reference performance (RP) is measured (E_1 and E_2 , row 3). The comparison of error terms (row 4) outputs a Boolean decision value (red).

We then connect our two main experiments to the framework in Section 2.3.

2.1 Numerical Representation of Performances

To capture nuances and deviation from the score in performances we use numerical features. Every performance yields sequences of measurements encoding an expressively relevant attribute, e.g., tempo. The sequence contains values (e.g., the current beat period) for each note or score onsets (from now on broadly referred to as dimensions), i.e., performances can be different from others in each of these dimensions and distance metrics aggregate differences in each of these dimensions into a single value. An example of the numerical sequence representation of performances in terms of beat period is illustrated in Figure 2. Performances differ from each other (vertically) in each dimension, i.e., at each score onset on the horizontal axis.

2.2 Randomization within the Ball of Expert Performances

Given a number of expert performances of the same piece, anyone can be chosen as a reference performance. This means that any other expert performance sits at some distance from the reference (in a high-dimensional space), some closer, some further away. If we are able to synthesize performances with an expected distance

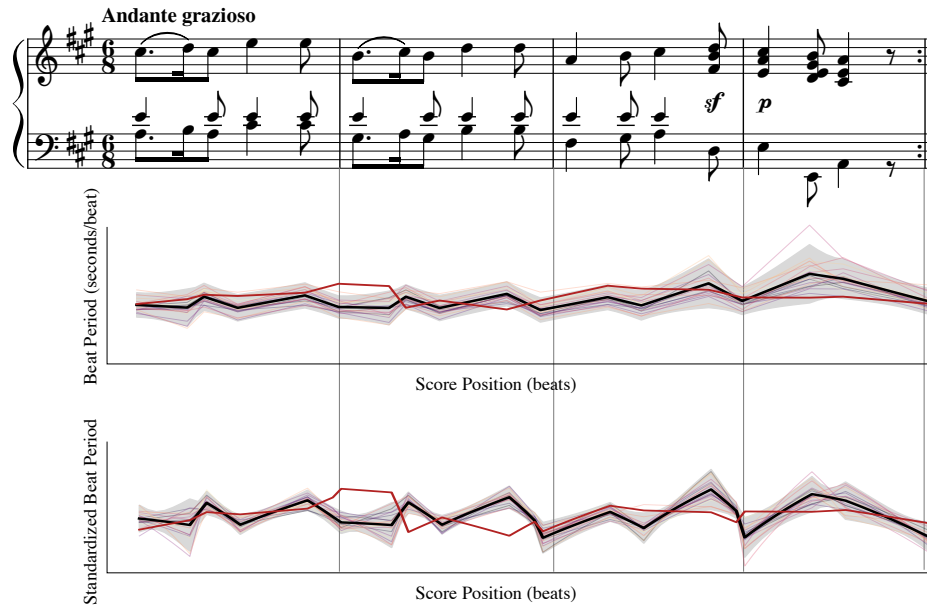


Figure 2: At top: excerpt of Mozart’s Piano Sonata K 331 Mv. 1 (Bars 5–8). Middle: non-standardized tempo curves. Bottom: mean-log standardized tempo curves (see Section 3). Colored lines represent tempo curves from the Vienna 4x22 dataset for the Mozart excerpt; black curves represent averaged tempo curves; red lines are randomly generated (non-musical) performances. The gray shaded area indicates one standard deviation above and below the average curve.

no greater than this expert performance’s distance from the reference, our performances would—in expectation and according to the framework—seem as good as an expert performance. In other words, any performance in a (high-dimensional) ball around the average expert performance looks musical to this evaluation framework, given that the ball diameter is the average distance between pairs of expert performances.

Our aim is thus to randomize performances that stay within this ball in expectation. We approximate this using mixture of Gaussian random variables, set at the mean of quantiles of the average expert performance. Figure 3 illustrates this process from top to bottom. We start by computing the average expression feature curve across performers of a chosen excerpt. We then split the dimensions (horizontal axis) according to quantiles of the expression feature (vertical axis). Finally, we define a Gaussian random variable for each quantile, defined by the mean of the expression curves within the quantile and a configurable standard deviation, which we refer to as noise level. Setting the noise level allows for (probabilistic) control over the (expected) distance to an average expert performance. Generally, the higher the noise level, the further the performance.

A possible result of randomization is shown in Figure 2 using an excerpt of a Mozart Piano Sonata. Expert performances are shown in gray and a generated randomized performance in red. Note that the shaded area where many expert performances come to lie (mean performance and one standard deviation above and below) is not an illustration of the high-dimensional ball defined by curves that do not exceed an average reconstruction error wrt the average performance. Such high-dimensional balls are difficult to visualize,

but intuitively large deviations in few dimensions are possible if the values in a majority of other dimensions fall very close to the reference.

Furthermore, note that this Gaussian mixture is not guaranteed to stay within this ball for general sequences or even any possible expression features sequences. The performances of one excerpt might lie very closely, narrowing down the possibilities such that even noise level zero, i.e., a quantile-wise deadpan performance, is beyond the ball. However, we never see this happen on our data.

2.3 Experiments

In the first experiment, we are interested in the capacity of listeners to **discern slightly randomized performances** that look similar to expert performances to the quantitative framework. Do listeners perceive these randomizations or are they too fine? In a listening test, we present participants with several pairs of performance excerpts, each pair consisting of one expert and one randomized, and ask them to identify the expert performance among each pair. The randomizations are created with precisely controlled error rates of the framework for each excerpt.

In the second experiment, we use the same randomized and expert performance pairs as in the previous one, however, with increased randomization strength and no excerpt-wise configuration of the randomization. In this scenario, a listener should be overwhelmingly likely to identify the randomization, at the cost of the randomization also being more visible to the evaluation framework, i.e., the framework should identify more than 50 % of human performances correctly.

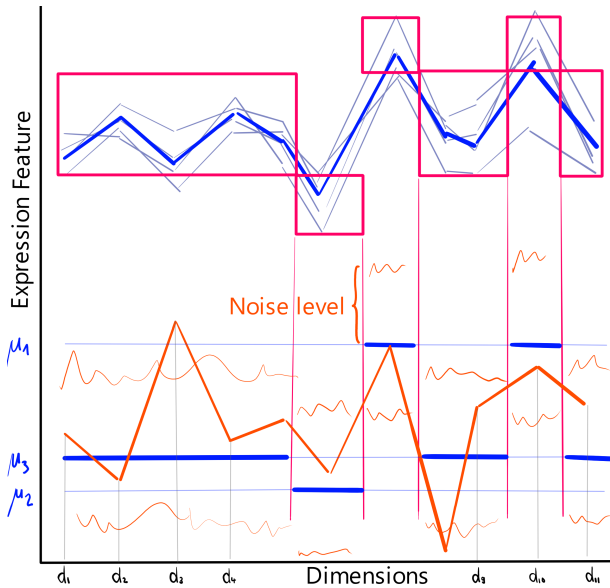


Figure 3: Illustration of the sampling process approximating the ball of expert performances with a mixture of three Gaussian random variables. The average performance (opaque blue, top) is computed from expert performances (translucent blue, top) and segmented into quantiles (red boxes). A randomized performance (orange, bottom) is then sampled from Gaussian distribution for each quantile, with a standard deviation controlled as noise level parameter.

This experiment addresses the second and third of our guiding questions: the **reliability**, i.e., the evaluative consistency, and the **validity**, i.e., the evaluative correctness, of the framework under various reference performances and validity of the quantitative evaluation framework.

3 METHODS

The previous discussion of the framework remained abstract, in this sections we discuss concrete dataset, expression features, standardizations, metrics, and randomizations used in the two experiments.

3.1 Datasets

For our analysis we use excerpts of MIDI or MIDI-like recordings with performed notes matched to their corresponding score notes extracted from two datasets:

Vienna 4x22: This dataset was originally compiled by Goebel [12] and consists of 4 excerpts of solo piano pieces, each performed by 22 pianists. The excerpts are the first 21 bars of Chopin’s Etude Op. 10 No. 3, the first 45 bars of Chopin’s Ballade in F Op. 38, the first 36 bars (i.e., the theme) of Mozart’s Piano Sonata in A K 331, and all 32 bars of Schubert D783 No. 15 (with repeats played). All performances were recorded on a Bösendorfer 290 SE Grand Piano as MIDI-like data and subsequently each played note matched to its respective score note.

KAIST / International Piano-e-Competition: This dataset consists of MIDI recordings of performances of several editions of

the International Piano-e-Competition¹ for a number of which researchers at KAIST [17] collected and corrected scores in MusicXML format. All performances were recorded on Yamaha Disklavier instruments. The scores and performances have been aligned by KAIST² using Nakamura et al.’s HMM-based alignment tool [21]. We converted these alignments to Matchfile format [11], extracted the pieces for which more than eight - or more than five in the case of Bach’s well-tempered clavier - performances exist and cleaned up the alignments.

Taken together, this yields 33 pieces or excerpts thereof – 16 by Frédéric Chopin, 8 by Johann Sebastian Bach, 5 by Ludwig Van Beethoven, 2 by Franz Liszt, 1 by Wolfgang Amadeus Mozart, and 1 by Franz Schubert – with 40786 unique score onsets, each played at least 6 and 34 times for a total of 476 performances.

3.2 Expression Features

In order to compare expressive parameters, we do not work directly with note-wise onsets, dynamics etc, but we compute four expression features: two onset-wise features (tempo and velocity), and two note-wise features (timing and articulation). These features are defined as follows:

- A **tempo** curve is derived by dividing the performed inter-onset interval (IOI) by the score IOI for every score onset, where the performed onsets are first averaged across note sharing a score onset. Tempo curves encode a measure of the rate of change measured in seconds per beat (aka beat period).
- Likewise, **dynamics** curves are computed as the average MIDI velocity of individual notes at each score onset.
- We define **timing** as the note-wise deviation from an average onset time of notes at a common score onset (as used in the tempo computation above) in milliseconds. The timing of notes sharing a score onset sums thus to zero, the timing of notes unique at their onset is also zero.
- We define **articulation** as the base-two logarithm of the played duration divided by the notated duration times the beat period.

These definitions are by no means universal, however, these or equivalent expression features are commonly used (see section 4.1 in [5]).

3.3 Standardization and Metric

The literature provides many examples of standardization, factoring and smoothing of expressive parameter curves (e.g. [8, 9, 20, 27]). Li et al. [20] proposed a number of standardization techniques which they compared as parameters in a model selection test. We evaluate four standardization techniques: mean standardization, mean-log standardization, mean/variance standardization (aka sampling standard score), and no standardization, under mean squared error (MSE). In the middle plot of Figure 2 non-standardized tempo curves are shown, and in the bottom we show the same curves but mean-log standardized. Note that the MSE of two series of data points - i.e. performances - x_1 and x_2 under mean variance

¹ formerly known as the Yamaha Piano-e-Competition <http://piano-e-competition.com/default.asp>

² https://github.com/mac-marg-pianist/chopin_cleaned

standardization is equivalent to $2 - 2 \times \rho(x_1, x_2)$, where ρ is the Pearson correlation coefficient. The given test results hence also imply evaluation under correlation, another commonly used metric.

3.4 Randomization

For our experiments we use the following quantile and noise level settings. The listening test uses quartiles (Q_1, Q_2, Q_3, Q_4), the noise level σ is set for each excerpt individually to control the evaluative validity of the framework. Formally, the randomizations are sampled from:

$$x_t \sim \begin{cases} \mathcal{N}(\mu_1, \sigma^2) & \text{if } t \in \{o \mid y_o \in Q_1\} \\ \mathcal{N}(\mu_2, \sigma^2) & \text{if } t \in \{o \mid y_o \in Q_2\} \\ \mathcal{N}(\mu_3, \sigma^2) & \text{if } t \in \{o \mid y_o \in Q_3\} \\ \mathcal{N}(\mu_4, \sigma^2) & \text{if } t \in \{o \mid y_o \in Q_4\} \end{cases}$$

The second experiment uses unequal quantiles (lowest 5%, center 90%, and highest 5%), the noise level is set to the average standard deviation across performances $\bar{\sigma}$. Formally, the randomizations are sampled from:

$$x_t \sim \begin{cases} \mathcal{N}(\mu_1, \bar{\sigma}^2) & \text{if } t \in \{o \mid P(y > y_o) \leq 0.05\} \\ \mathcal{N}(\mu_2, \bar{\sigma}^2) & \text{if } t \in \{o \mid P(y < y_o) \leq 0.05\} \\ \mathcal{N}(\mu_3, \bar{\sigma}^2) & \text{else} \end{cases}$$

where the quantiles of dimensions t are shown as sets constrained by the probabilities P of curve values y at these dimensions.

4 LISTENER DISCERNMENT EXPERIMENT

Using a listening test we estimate the degree to which listeners are capable of discerning differences in performance expression features that look similar under the quantitative evaluation framework.

4.1 Data

For the listening test we extract excerpts of pieces of the Vienna 4x22 dataset. We use two excerpts per expression feature, with all four expression features (tempo, timing, articulation, and velocity) being investigated, making for a total of eight excerpts. The excerpts are chosen based on two considerations: First, they need to cover enough musical material to be able to judge phrasing and timbre, but not be too long for the listeners. We opt for 8 - 10 measures. Secondly, we extract all excerpts fulfilling the length criteria and measure their inter-performance correlation. For each expression feature, we choose the excerpts with the highest and the lowest correlations, respectively. For high correlation excerpts, performances are very consistent across performers, we thus expect the randomization ball to be small, and identification of randomized performances correspondingly harder. We further double the number of test pairs by using two noise levels. Noise level 50 refers to standard deviations set in the randomization such that the framework identifies 50 % of the pairs correctly, i.e., the framework evaluates at chance level, the randomization is indistinguishable to the framework. At noise level 90, the framework identifies 90 % of randomizations. We expect listeners to be able to identify the stronger randomizations (noise level 90) with greater ease. Each of

the eight excerpts is matched with 44 randomized performances, 22 at noise level 50, 22 at noise level 90.

4.2 Listening Test

Participants are provided with an online questionnaire of 16 pairs of performances, one for each test case, randomly sampled from the 22×22 possible (random \times expert) pairings. On the first page, listeners are instructed to the task – listening to the two audio files and identifying the expert performance among them – and presented the five items of the short Musical Training subsection of the Goldsmiths Music Sophistication self-assessment Index (GMSI). Of the participants that completed the GMSI questions, 56% engaged in regular practice of a musical instrument for 4 or more years and 69% reported practicing their primary instrument for at least 2 hours per day. Listeners can start, pause, stop, or rewind the audio excerpts at their leisure. The possible answers include: performance 1 is the expert performance, performance 2 is the expert performance, and undecided.

4.3 Results

More than 250 listeners participate in the online study, with usable (unskipped) answers per (noise level \times feature)-configuration ranging from 185 to 240. Table 1 presents the results of the listening test. The table breaks down the answers hierarchically, with the top row identifying the four expression features studied. The next five rows from the top divide each feature into two noise levels and report from top to bottom: the noise level used, the total number of answers, the number of correct expert performer identification, the ratio of correct identification as percentage, and finally the probability (as percentage) of this outcome under a binomial distribution with success probability of 0.5, the distribution corresponding to the null hypothesis; listeners can't discern the expert performances. The next six rows report the same values again, albeit further broken down by excerpt. For each excerpt we further note the starting point and duration in measures.

Most apparent from the results is that the inconsistency of listener discernment across features. They largely fail to perform better than chance for timing and velocity, yet show clear (and statistically significant at $p=0.01$) discernment for articulation and tempo. Furthermore, the noise level influences the results as assumed for tempo and articulation, but fails to influence the judgment of the other two in a significant way. Addressing our first guiding question, listeners discerned randomization in both articulation and tempo which are indistinguishable under the evaluative framework (noise level 50). However, the framework readily identifies stronger randomizations (noise level 90) in velocity and timing, which escape the listeners.

5 VALIDITY AND RELIABILITY EXPERIMENT

This experiment addresses our guiding questions two and three, concerning the reliability and validity of reconstruction error-based evaluations under different reference performances, respectively. All experiments are carried out with respect to two of the performers' expressive parameters, namely onset-wise tempo and dynamics curves.

Feature	Articulation		Timing		Tempo		Velocity	
	50	90	50	90	50	90	50	90
Noise level	50	90	50	90	50	90	50	90
Number of answers	240	185	215	235	215	234	238	213
Number correct	148	139	126	129	147	183	125	109
Percentage correct	61.67	75.14	58.60	54.89	68.37	78.21	52.52	51.17
Percentage probability	0.01	0.00	0.22	1.69	0.00	0.00	3.82	5.15
Piece	Mozart k331	Chopin op38	Mozart k331	Chopin op38	Chopin op38	Chopin op38	Chopin op10 no3	Chopin op10 no3
Excerpt start	8	0	8	0	33	17	33	17
Excerpt duration	8	9	8	9	12	8	13	9
Number of answers	113	127	105	80	110	105	119	116
Number correct	78	70	75	64	66	60	70	59
Percentage correct	69.0	55.1	71.4	80.0	60.0	57.1	58.8	50.9
Percentage probability	0.00	3.65	0.00	0.00	0.85	2.68	1.15	7.27

Table 1: Results of the listening test broken down hierarchically, with the top row identifying the four expression features. The next five rows divide each feature into two noise levels and report from top to bottom: the noise level used, the total number of answers, the number of correct expert performer identification, the ratio of correct identification as percentage, and finally the probability (as percentage) of this outcome under the null hypothesis. The next six rows report the same values again, albeit split down by excerpt.

See Figure 1 for a schematic representation of the frameworks. We use this framework to evaluate expert performances against randomized ones. For each piece in the combined datasets described above (see Section 3.1), we create 64 randomized ones. The randomization starts from the average expert performance and follows the process described in section 2.2 and used in the listening test, albeit with one major difference: the randomization follows a mixture of three Gaussians corresponding to the top 5 %, bottom 5 %, and center 90 % quantiles, the noise level is set to the overall average standard deviation of the expression features for each piece. Given the results of the listening test, we assume a listener to be overwhelmingly likely to identify the randomization, at least for the tempo curves.

5.1 Reliability and Validity

Using the described ground truth models, we compute validity and reliability values for the given evaluation frameworks. We define **reliability** as the consistency of the evaluation framework under changes of references and across a variety of pieces, independent of the correctness of this result. Given a human expert performance (produced by the ‘musical’ model) and a random sequence (produced by the ‘unmusical’ model), does the framework consistently favor the same model wrt different reference performances? This consistency is quantified as average correlation of the binary output of the two-model evaluation (0 = model 1 has smaller MSE, 1 = model 2 has smaller MSE) wrt different targets. This is interpretable as inter-reference-performance correlation of evaluation framework. A perfectly reliable evaluation always favors the same model independent of RP.

We define the **validity** of the frameworks as the extent to which they accurately recover the ground truth. A perfectly valid evaluation will always favor the expert performance and reject the randomized sequence. Numerically, validity is estimated by the ratio of tests that erroneously recover the randomized performance over all possible reference, test and random performance combinations. As for reliability, we compute and compare this number across a variety of pieces.

All in all, then, both tempo and dynamics under four standardizations are evaluated in two tests over 33 pieces. This amounts to a total of $2 \times 4 \times 2 \times 33 = 792$ experiments. Every test is carried out for the n reference performances, $n - 1$ test expert performances, and 64 randomly sampled performances, where n is the number of expert performances available for the respective piece.

5.2 Results

In this section, we present the results of selected tests. Results are reported in Table 2, one part for dynamics curves, and another for tempo curves. MSE between expression features under mean variance standardization, i.e. the standard score per performance excerpt, proved most beneficial for the framework’s discernment capacity and is hence used throughout the experiments. Each row in Table 2 represents a piece, the values given in the first four columns are as follows. The name and opus number of the piece and its composer. The number of expert performances, the number of their shared onsets.

The following four columns are given once for tempo curves and once for dynamics curves. The mean of three MSE distributions:

global values			values wrt dynamics curves				values wrt tempo curves							
Piece	Composer	No Onsets	reliability	validity	\overline{MSE} e. - e.	\overline{MSE} e. - r.	\overline{MSE} r. - r.	reliability	validity	\overline{MSE} e. - e.	\overline{MSE} e. - r.	\overline{MSE} r. - r.	reliability	validity
Etude Op. 10 No 1	Chopin	26	1187	1.16	1.33	0.88	0.98	10.2 %	0.94	1.11	1.19	0.75	0.94	30.6 %
Etude Op. 10 No 2	Chopin	11	729	1.23	1.39	0.95	0.58	16.6 %	0.86	1.09	1.43	1.12	0.86	3.7 %
Etude Op. 10 No 4	Chopin	26	1145	0.96	1.24	0.74	0.94	1.8 %	0.18	1.28	1.28	0.82	0.18	49.0 %
Etude Op. 10 No 5	Chopin	11	894	0.96	1.22	0.67	0.87	3.6 %	0.32	0.94	1.03	0.48	0.32	28.6 %
Etude Op. 10 No 8	Chopin	28	1380	1.03	1.28	0.78	0.9	3.1 %	0.49	1.14	1.25	0.86	0.49	22.6 %
Etude Op. 10 No 12	Chopin	13	1264	0.76	1.04	0.42	0.98	0.5 %	0.2	0.85	0.88	0.28	0.2	36.9 %
Etude Op. 25 No 10	Chopin	12	910	0.39	1.06	0.28	1.0	0.0 %	1.0	0.45	0.81	0.2	1.0	0.0 %
Etude Op. 25 No 11	Chopin	25	1878	0.94	1.17	0.69	0.86	4.3 %	0.72	1.13	1.33	0.99	0.72	8.9 %
Barcarolle Op. 60	Chopin	9	1597	0.53	1.06	0.35	1.0	0.0 %	0.45	0.66	0.77	0.16	0.45	18.6 %
Sonata No. 3 Op. 58 2nd	Chopin	9	879	0.75	1.03	0.42	0.99	0.2 %	1.0	0.54	0.87	0.25	1.0	0.0 %
Sonata No. 3 Op. 58 4th	Chopin	9	2506	0.92	1.17	0.55	1.0	0.0 %	0.3	1.03	1.12	0.62	0.3	22.6 %
Scherzo No. 2 Op. 31	Chopin	12	2874	0.49	1.05	0.35	1.0	0.0 %	0.97	0.57	0.79	0.22	0.97	0.8 %
Ballade No 1 Op. 23	Chopin	18	2174	0.49	0.97	0.31	1.0	0.0 %	1.0	0.35	0.66	0.11	1.0	0.0 %
Ballade No 4 Op. 52	Chopin	12	2583	0.45	1.03	0.31	1.0	0.0 %	1.0	0.51	0.83	0.21	1.0	0.0 %
WTC BWV 848 F.	Bach	9	831	1.15	1.36	0.84	0.47	15.4 %	1.0	0.99	1.37	1.04	1.0	0.0 %
WTC BWV 857 F.	Bach	6	753	0.93	1.17	0.56	0.79	5.6 %	0.36	0.8	1.11	0.44	0.36	20.3 %
WTC BWV 860 F.	Bach	6	988	1.03	1.29	0.82	0.78	7.4 %	1.0	0.53	1.09	0.55	1.0	0.0 %
WTC BWV 889 F.	Bach	6	567	1.34	1.4	1.02	0.1	39.9 %	1.0	0.79	1.27	0.83	1.0	0.1 %
WTC BWV 848 P.	Bach	9	610	1.04	1.24	0.67	0.57	14.6 %	0.99	1.01	1.52	1.3	0.99	0.1 %
WTC BWV 857 P.	Bach	6	317	1.04	1.24	0.66	0.46	16.6 %	1.0	0.68	1.15	0.84	1.0	0.0 %
WTC BWV 860 P.	Bach	6	427	0.96	1.17	0.6	0.65	9.9 %	0.83	0.99	1.37	0.86	0.83	5.8 %
WTC BWV 889 P.	Bach	6	595	0.88	1.21	0.81	1.0	0.0 %	0.47	0.98	1.15	0.81	0.47	15.5 %
T. Etude S. 139 No 10	Liszt	16	1842	0.66	1.09	0.44	1.0	0.0 %	0.99	0.64	0.92	0.31	0.99	0.3 %
P. Etude S.141 No 1	Liszt	34	439	1.19	1.36	0.94	0.47	21.2 %	29.3 %	1.37	1.58	1.35	0.3	29.3 %
Sonata No 3 Op. 2/3 1st	Beethoven	9	2190	0.45	0.99	0.27	1.0	0.0 %	1.0	0.46	1.09	0.37	1.0	0.0 %
Sonata No 4 Op. 7 1st	Beethoven	9	2046	0.76	1.06	0.42	1.0	0.0 %	61.1 %	1.22	1.19	0.65	0.13	61.1 %
Sonata No 17 Op. 31/2 1st	Beethoven	11	1376	0.6	1.02	0.38	1.0	0.0 %	0.4 %	0.18	0.38	0.04	0.99	0.4 %
Sonata No 18 Op. 31/3 1st	Beethoven	9	1707	0.64	1.0	0.38	1.0	0.0 %	88.3 %	1.0	0.81	0.17	0.6	88.3 %
Sonata No 21 Op. 53 1st	Beethoven	25	3452	0.5	1.09	0.4	1.0	0.0 %	7.0 %	0.75	1.13	0.56	0.82	7.0 %
Etude Op. 10 No. 3	Chopin	22	162	0.34	0.97	0.29	1.0	0.0 %	0.8 %	0.43	0.83	0.44	0.97	0.8 %
Ballade No 2 Op. 38	Chopin	22	197	0.64	1.05	0.57	0.99	0.3 %	0.3 %	0.14	0.86	0.27	0.99	0.3 %
Sonata No. 11 K. 331	Mozart	22	178	0.66	1.14	0.6	0.96	1.0 %	2.2 %	0.47	0.9	0.43	0.92	2.2 %
G. Dance D. 783 No. 15	Schubert	22	109	0.6	1.19	0.56	1.0	0.1 %	6.8 %	0.66	1.14	0.65	0.79	6.8 %
Dataset		476	40786	0.8	1.15	0.57	0.85	5.2 %	14.0 %	0.78	1.06	0.58	0.73	14.0 %

Table 2: The results of the validity and reliability tests by piece. The abbreviations used are: e. = expert; perf. = performance; r. = random; corr. = Pearson correlation coefficient; mod. = model evaluation; No = number; Op. = Opus; F. = Fugue; P. = Prelude; WTC = Well-Tempered Clavier; BWV = Bach-Werke-Verzeichnis; G. = German; D. = Deutsch catalogue; K. = Köchelverzeichnis; S. = Searle catalogue

the inter expert performance MSEs, the MSEs between expert performances and randomized performances, and the MSEs among randomized performances. The next column reports the reliability of the two-model evaluation as the mean of correlations among the two model tests over different target performances. Lastly, the validity of the framework is given as the percentage of randomly sampled performances with lower MSE than a given expert performance.

5.2.1 Reliability. Values in Table 2 relating to important aspects of reliability are colored in **red**. The average correlation of all two-model evaluations wrt dynamics curves is 0.85, the highest value being 1.0 and the lowest 0.09 (Table 2, col. 8). The average correlation of all two-model evaluations wrt tempo curves is 0.73, the highest value being 1.0 and the lowest 0.13 (col. 14). Generally, there is agreement in a majority of pieces and less reliability in a minority. For 13 pieces, the correlation of evaluations drops below 0.5 wrt tempo or dynamics, highlighting high variation across pieces.

The pieces exhibiting low reliability differ between tempo curve and dynamics curves tests. Only one piece (Grande Etude de Paganini S.141 No 1) shows correlation below 0.5 in both tempo and dynamics curves.

5.2.2 Validity. Important values in Table 2 relating to the validity of the two-model evaluation are colored in **blue**. The two-model evaluation validity tests show an average of 5.3 % of comparisons wrt dynamics, and an average of 14.0 % of comparisons wrt tempo, favoring the randomly sampled performance (col. 9/14). The average for all pieces is not weighted by the number of expert performances or tests. The percentages vary greatly from perfect recovery of all expert performances to 88.3 % of evaluations favoring random performances. Again, valid evaluation wrt tempo does not imply valid evaluation wrt dynamics and vice versa. 15 pieces exhibit good performance of the framework with rejection of random performances in more than 90 % of cases for both tempo and dynamics.

6 DISCUSSION AND CONCLUSIONS

Performance data is complex and sometimes more opaque than apparent at first glance. Not without reason have researchers interested in performance practice and computational performance modelling spent decades dissecting the minutiae of phrasing, melody lead, pedalling, to name just a few aspects. To better appreciate the breadth of issues, we briefly discuss several research directions.

Directly implied in our investigations are computational models of expressive performance, we refer to [5] and [19] for a comprehensive overview. For an overview of methods for evaluating computational models of expressive performance, we refer the reader to [4].

Other cues come from performance research related to listener judgments, e.g., the seminal work by Repp [25] which presents evidence suggesting that listeners prefer average performances. Wesolowski et al. [28] present a critical view of listeners' aesthetic judgments as a methodological tool for evaluating the differences in Jazz ensemble performances by analyzing their ratings' variability. The music psychology literature provides evidence showing that the assessment of the (aesthetic) quality of a performance depends not only on the auditory component of a performance (e.g., [24]).

Performance practice research is also interested in an entirely different type of perceptual classification of performance, namely semantic descriptor of expressive performance, or, more commonly, instrumental timbre. By means of example, we refer to the sequence of studies undertaken by Bernays et al. [1–3], or more recently and from within the MIR community [6]. Besides verbal descriptions, quantitative performance research often takes the form of detailed analyses of expression features in specific contexts. Exemplary work was carried out by Goebl et al. [12, 14], e.g., their work on the sources of melody lead [13].

From a music education perspective, Gururani et al. [16] investigate quantitative descriptors for assessing the quality of performance. Pati et al. [22] present a deep learning based approach to assess student music performance.

Our tests add some bits to the knowledge surrounding measured expressive performances and their generative models. They indicate that MSE based model evaluation is not necessarily reliably favoring the same performance wrt different targets. Furthermore, MSE based model evaluation is not dependably capable of discerning expert performances from randomized performances. The pieces under examination show great variability both wrt to the tests, as well as wrt closeness of expert performances. Listeners perceive randomizations in articulation and tempo that escape the evaluation framework, but they do not notice randomizations in velocity and microtiming with the same acuity. Reasons for this can be sought both in the perception as well as in the production of expressive performances.

How then can automatic, quantitative evaluation be improved? Our experiments and experience allow only for tentative answers, but answers they still are: Most settings seem to benefit from more fine-grained evaluations. Shorter excerpts tend to give more reliable and valid results and are better suited to localize errors. If multiple performances are available, test excerpts can be chosen which have high internal consistency, i.e., high inter-performer correlation or low inter-performer MSE, respectively. Ideally these excerpts can relate to specific and discussed performance issues like phrasing, clear voices, specific timbre, etc. Formulated in the negative, researchers should avoid resting their evaluative arguments on aggregated absolute errors across large, undocumented test dataset splits. Such numbers carry too little information about the models under scrutiny.

Even better evaluation could plausibly be achievable with *distributional metrics*, e.g. the probability of generated performances under a Gaussian process (GP) regressor fitted with expert performances or inversely the likelihood of a generative GP model, like the model Teramura et al. [26] proposed, for test performances. In a similar vein, trained neural network (NN) discriminators seem a promising avenue for future research. However, neither tractable (GP) nor untractable (NN) approaches are a priori connected to listener judgment.

This is by no means an exhaustive discussion of issues surrounding the perception, characterization, and quantification of expressive performance, but we hope it serves to gain an appreciation of the intricacies of this data. Prospective as well as seasoned researchers in the field of GMEPP do well in reminding themselves of these facts: piano performance are aesthetically, culturally and axiologically rich, dynamic, and complex musical objects.

ACKNOWLEDGMENTS

This work is supported by the European Research Council (ERC) under the EU's Horizon 2020 research & innovation programme, grant agreement No. 101019375 ("Whither Music?"), and the Federal State of Upper Austria (LIT AI Lab).

REFERENCES

- [1] Michel Bernays and Caroline Traube. 2010. Expression of piano timbre: gestural control, perception and verbalization. In *Proceedings of CIM09: The 5th Conference on Interdisciplinary Musicology*.
- [2] Michel Bernays and Caroline Traube. 2013. Expressive production of piano timbre: touch and playing techniques for timbre control in piano performance. In *Proceedings of the 10th Sound and Music Computing Conference (SMC2013)*. KTH Royal Institute of Technology Stockholm, Sweden, 341–346.
- [3] Michel Bernays and Caroline Traube. 2014. Investigating pianists' individuality in the performance of five timbral nuances through patterns of articulation, touch, dynamics, and pedaling. *Frontiers in Psychology* 5 (2014), 157.
- [4] Roberto Bresin and Anders Friberg. 2013. Evaluation of Computer Systems for Expressive Music Performance. In *Guide to Computing for Expressive Music Performance*, Alexis Kirke and Eduardo R Miranda (Eds.). Springer-Verlag, London, UK, 181–203.
- [5] Carlos Cancino-Chacón, Maarten Grachten, Werner Goebel, and Gerhard Widmer. 2018. Computational Models of Expressive Music Performance: A Comprehensive and Critical Review. *Frontiers in Digital Humanities* 5 (2018), 25. <https://doi.org/10.3389/fdigh.2018.00025>
- [6] Carlos Cancino-Chacón, Silvan Peter, Shreyan Chowdhury, Anna Aljanaki, and Gerhard Widmer. 2020. On the characterization of expressive performance in classical music: First results of the con espressione game. (2020).
- [7] Carlos Eduardo Cancino-Chacón, Thassilo Gadermaier, Gerhard Widmer, and Maarten Grachten. 2017. An Evaluation of Linear and Non-linear Models of Expressive Dynamics in Classical Piano and Symphonic Music. *Machine Learning* 106, 6 (2017), 887–909.
- [8] Elaine Chew. 2012. About Time: Strategies of Performance Revealed in Graphs. *Visions of Research in Music Education* 20 (2012).
- [9] Simon Dixon, Werner Goebel, and Emiliios Cambouropoulos. 2006. Perceptual Smoothness of Tempo in Expressively Performed Music. *Music Perception* 23, 3 (2006), 195–214.
- [10] Sebastian Flossmann, Maarten Grachten, and Gerhard Widmer. 2013. Expressive Performance Rendering with Probabilistic Models. In *Guide to Computing for Expressive Music Performance*, Alexis Kirke and Eduardo R Miranda (Eds.). Springer, London, UK, 75–98.
- [11] Francesco Foscarin, Emmanouil Karystinaios, Silvan David Peter, Carlos Cancino-Chacón, Maarten Grachten, and Gerhard Widmer. 2022. The match file format: Encoding Alignments between Scores and Performances. In *Proceedings of the Music Encoding Conference (MEC)*. Halifax, Canada.
- [12] Werner Goebel. 1999. The Vienna 4x22 Piano Corpus. <https://doi.org/10.21939/4X22>
- [13] Werner Goebel. 2001. Melody lead in piano performance: Expressive device or artifact? *The Journal of the Acoustical Society of America* 110, 1 (2001), 563–572. <https://doi.org/10.1121/1.1376133> arXiv:<https://doi.org/10.1121/1.1376133>
- [14] Werner Goebel and Caroline Palmer. 2009. Synchronization of timing and motion among performing musicians. *Music Perception* 26, 5 (2009), 427–438.
- [15] Graham Grindlay and David Helmbold. 2006. Modeling, analyzing, and synthesizing expressive piano performance with graphical models. *Machine Learning* 65, 2-3 (June 2006), 361–387.
- [16] Siddharth Gururani, Kumar Ashis Pati, Cih-Wei Wu, and Alexander Lerch. 2018. Analysis of Objective Descriptors for Music Performance Assessment. In *Proceedings of the International Conference on Music Perception and Cognition (ICMPC)* (2018-01-01). Montreal, Canada. http://www.musicinformatics.gatech.edu/wp-content_nondefault/uploads/2018/06/Gururani-et-al.-2018-Analysis-of-Objective-Descriptors-for-Music-Perfor.pdf
- [17] Jeong, Dasaem and Kwon, Taegyun and Kim, Yoojin and Lee, Kyogu and Nam, Juhan. 2019. VirtuosoNet: A Hierarchical RNN-based System for Modeling Expressive Piano Performance. In *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR 2019)*. Delft, The Netherlands, 908–9015.
- [18] Tae Hun Kim, Satoru Fukuyama, Takuya Nishimoto, and Shigeki Sagayama. 2013. Statistical Approach to Automatic Expressive Rendition of Polyphonic Piano Music. In *Guide to Computing for Expressive Music Performance*, Alexis Kirke and Eduardo R Miranda (Eds.). Springer, London, UK, 145–179.
- [19] Alexander Lerch, Claire Arthur, Ashis Pati, and Siddharth Gururani. 2019. Music Performance Analysis: A Survey. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)* (2019-01-01). Delft. http://www.musicinformatics.gatech.edu/wp-content_nondefault/uploads/2019/06/Lerch-et-al.-2019-Music-Performance-Analysis-A-Survey.pdf
- [20] Shengchen Li, Dawn AA Black, and Mark D Plumbley. 2015. The clustering of expressive timing within a phrase in classical piano performances by Gaussian Mixture Models. In *Proceedings of the 11th International Symposium on Computer Music Multidisciplinary Research (CMMR 2015)*. Plymouth, UK, 322–345.
- [21] Eita Nakamura, Kazuyoshi Yoshii, and Haruhiro Katayose. 2017. 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 2018)*. Suzhou, China, 347–353.
- [22] Kumar Ashis Pati, Siddharth Gururani, and Alexander Lerch. 2018. Assessment of Student Music Performances Using Deep Neural Networks. *Applied Sciences* 8, 4 (2018), 507. <https://doi.org/10.3390/app8040507>
- [23] Silvan David Peter, Carlos Eduardo Cancino-Chacón, Francesco Foscarin, Andrew Philip McLeod, Florian Henkel, Emmanouil Karystinaios, and Gerhard Widmer. 2023. Automatic Note-Level Score-to-Performance Alignments in the ASAP Dataset. *Transactions of the International Society for Music Information Retrieval (TISMIR)* (2023). <https://doi.org/10.5334/tismir.149>
- [24] Friederich Platz and Reinhard Kopiez. 2012. When the Eye Listens: A Meta-analysis of How Audio-visual Presentation Enhances the Appreciation of Music Performance. *Music Perception* 30, 1 (2012), 71–83.
- [25] B. Repp. 1997. The aesthetic quality of a quantitatively average music performance: Two preliminary experiments. *Music Perception* 14, 4 (1997), 419–444.
- [26] Keiko Teramura, Hideharu Okuma, Yuusaku Taniguchi, Shimpei Makimoto, and Shin Maeda. 2008. Gaussian process regression for rendering music performance. In *Proceedings of the 10th International Conference on Music Perception and Cognition (ICMPC 10)*. Sapporo, Japan.
- [27] A Tobudic and G Widmer. 2003. Playing Mozart phrase by phrase. In *Proceedings of the Fifth International Conference on Case-Based Reasoning (ICCB-03)*. Springer-Verlag, 552–566.
- [28] B C Wesolowski, S A Wind, and G Engelhard. 2016. Examining Rater Precision in Music Performance Assessment: An Analysis of Rating Scale Structure Using the Multifaceted Rasch Partial Credit Model. *Music Perception* 33, 5 (June 2016), 662–678.
- [29] Huan Zhang, Jingjing Tang, Syed RM Rafee, Simon Dixon, George Fazekas, and Geraint A Wiggins. 2022. ATEPP: A Dataset of Automatically Transcribed Expressive Piano Performance. In *Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR)*.