EXPERT AND NOVICE EVALUATIONS OF PIANO PERFORMANCES: CRITERIA FOR COMPUTER-AIDED FEEDBACK

Yucong Jiang
University of Richmond
yjiang3@richmond.edu

ABSTRACT

Learning an instrument can be rewarding, but is unavoidably a huge undertaking. Receiving constructive feedback on one's playing is crucial for improvement. However, personal feedback from an expert instructor is seldom available on demand. The goal motivating this project is to build software that will provide comparably useful feedback to beginners, in order to supplement feedback from human instructors. To lay the groundwork for that, in this paper we investigate performance assessment criteria from both quantitative and qualitative perspectives. We gathered 83 piano performances from 21 players. Each recording was evaluated by both expert piano instructors and novice players. This dataset is unique in that the novice evaluators are also players, and that both quantitative and qualitative evaluations are collected. Our analysis of the evaluations indicates that the kind of specific, concrete piano techniques that are most elusive to novice evaluators are precisely the kind of characteristics that can be detected, measured, and visualized for learners by a well-designed software tool.

1. INTRODUCTION

Learning to play a musical instrument can be rewarding, but is also unavoidably a huge undertaking. Receiving feedback on one's playing is crucial for improvement. However, personal feedback from an expert instructor is seldom available on demand; it is typically available (if at all) only in weekly music lessons. Our long-term goal in this project is to build software that will provide comparably useful feedback to beginners, as needed, in order to supplement insights from human instructors. The modes of computer-generated feedback could involve textual or visual indicators, or a mix of both. However, determining what kinds of feedback are especially helpful for beginners (among those that are feasible for computers to generate) is not trivial and should not be based on assumptions. To lay the groundwork for meaningful computer-aided feedback, therefore, in this paper we gather information on how experts and novices assess piano performances and what criteria they tend to rely on in such assessments.

Recent years have seen rapid growth in commercial products for computer-assisted instrumental learning. Unfortunately, most applications cannot deal with performances involving expressive timing: they expect users to play at a fixed tempo throughout a piece, even though such performances in real life are often perceived as boring and far short of the full expressive potential of music. For example, Yousician [1] and Simply Piano [2] color correctly played notes as the user progresses through a song at a preset tempo. While platforms like this have their own purposes and values, such oversimplified music playing experiences can mislead some learners to think that making music is all about playing the correct notes (rather than better sounding notes). Moreover, real-time feedback could distract players from listening to themselves, and as Percival et al. [3] point out, "computer analysis and interaction should occur *after* a student has finished playing".

Therefore, for our purposes, it makes more sense to envision software that can analyze a complete performance recording before providing feedback. Given such a recording, we would like to investigate what additional evaluation criteria (beyond note accuracy) should be incorporated into the feedback. In fact, even beginner-level players can usually tell when they've hit wrong notes, as the music won't sound right, but they often lack the ability to make more sophisticated judgments about the quality of their playing: articulation, tempo control, dynamics, and interpretation or expressiveness. Therefore, in this paper we focus on analyzing performances that are relatively "correct" in terms of wrong notes, so that they are ready for more nuanced aspects to be evaluated.

We have gathered 83 such piano performances from 21 players, each of whom chose from among seven beginner pieces. Each recording was evaluated by four expert piano instructors, and also by 17 peers from among the players themselves, with both numerical ratings and written comments. In this paper we examine (1) whether instructors and players evaluate performances differently, (2) whether better players are also better evaluators, and (3) what objective indicators can be detected and measured by computers that would reflect comparable evaluation criteria. This dataset is unique in that the peer evaluators are also players, and both quantitative and qualitative evaluations are collected. Each performance has also been aligned to its score, making it possible in the future to derive addi-

tional objective measurements (e.g., tempo variations), and to support further analyses relating performances to their scores (e.g, inter-song performance analysis).

2. RELATED WORK

A recent review paper [4] offers a comprehensive discussion of computer-aided instrument learning. The authors emphasize the differences between systems that are designed to measure competence and those designed to enhance learning, as the former only need to provide a rating, but the latter need to provide descriptive evidence justifying the evaluation. Two other review papers [5] [6] discuss the potentials of utilizing MIR techniques in music education. Example work on piano music tutor systems include [7] [8] [9] [10] [11].

Music performance analysis (MPA) is a broader topic, encompassing other purposes and uses beyond assisting learners, but a recent review paper [12] does discuss its application potential and challenges in regard to music education. Another closely related topic is modeling expressive music performance [13] [14] , which focuses on more abstract and higher-level aspects of a performance.

Related to examining performance evaluations, [15] discusses subjectivity in music performance assessment, [16] investigates how individual raters differ in their rating scale structure, and [17] provides insights on the benefit of peer assessment of music performance.

3. DATASET DESCRIPTION

Our dataset includes three components: 83 piano performance recordings in the WAV format, spanning seven different musical pieces; 803 evaluations of these performances, with players' metadata; and 83 audio-to-score alignments (with seven MusicXML score files and 83 alignment text files) indicating the starting time in the audio of each musical note in each score. This dataset is publicly available at facultystaff.richmond.edu/∼yjiang3/papers/ismir23/.

3.1 Performance Recordings

We recruited 21 participants from a local college, using flyers and campus-wide email announcements. These participants represent a range of piano experience, from a low of three months to a high of 16 years. Each participant completed a short questionnaire before recording a performance for the project. Except for one music major and one music minor, the participants play piano as a hobby. More than one participant recounted the story that they took piano lessons growing up, played on-and-off throughout the years, and recently came back to practicing it in college. When asked to self-identify their piano skill levels, nine of the 21 described their skills as "advanced", eight as "intermediate", and four as "beginner". (None chose the "professional" category from our prompt.)

We selected seven pieces from a popular score book for adult group piano classes [18], and asked each player to play however many pieces they felt like from these. (This flexibility helped recruit lower-level players who might otherwise be intimidated by this task.) The sheet music was shared with them weeks in advance to allow time for practice and preparation. Table 1 provides the names of these pieces and the number of performances of each. The players were advised to warm up before a recording session, and when recording, were offered the option either to be left alone in the piano room (to decrease nervousness) or to have the researcher present. They were allowed to re-record multiple times until satisfied with their own playing (e.g., with the preponderance of notes played correctly).

Piece Name	#Measures	#Recordings
Careless Love	16	11
Cielito Lindo	16	6
Lavender's Blue	16	17
Over the Waves	32	11
She Wore a Yellow Ribbon	34	13
The Blues	16	17
The Entertainer	40	8

Table 1. Summary of performance recordings.

3.2 Performance Evaluations

To evaluate the quality of these recordings, we recruited four professional piano instructors and 17 out of the 21 players (the other four were unfortunately not available for this stage). The instructors all have doctoral degrees and at least two decades of teaching experience. Each performance was evaluated by all four instructors and at least five (sometimes six) randomly chosen peers, resulting in 803 evaluations in total. The evaluators were asked to provide a numerical rating from one (poor) to five (excellent) for each recording, and also to briefly explain the basis for their rating, describing what criteria they considered. We collected the evaluations through a web-based form where users can play the recordings (grouped by score), enter evaluations, and save their progress. The sheet music is also linked from the form. It took between two to three hours for each instructor to evaluate all 83 performances, and 50 minutes on average for each peer player to evaluate 27 or 28 performances.

3.3 Audio-to-Score Alignment and Web Interface

Based on the sheet music, we created seven digital scores in the MusicXML format, and aligned each performance to its score. The alignment was achieved by the hidden Markov model proposed in [19], with occasional manual corrections. Each of the 83 alignment files contains two columns of values: a musical time in the score, and its played time in the recording. The alignment can also support future analyses relating performance attributes to elements in scores. For example, one could easily investigate whether players tend to slow down at a particular measure.

To make it more convenient to explore this dataset, we have built a demonstration web-based interface where a user can select and play any of these performances, while looking at the sheet music with the currently

played notes highlighted. The evaluations of the selected performance are also shown on the same page. This interface can help anyone interested in this dataset to find connections among performances, scores, and evaluations. This website is publicly available at facultystaff.richmond.edu/∼yjiang3/papers/ismir23/.

4. QUANTITATIVE ANALYSIS

4.1 Self-Identified Piano Levels

To verify the accuracy of the performers' piano skill levels, we separate the performances into three groups according to their self-reported levels, and compare the evaluators' ratings of those performances in each group. Figure 1 compares three box plots, one for each group of performance ratings, where each performance rating in a group represents the average among the four instructors. Although the median rating increases with the skill levels, the three distributions overlap with each other. To test whether the difference between the advanced group's ratings and the intermediate group's is statistically significant, we conduct a one-sided Welch's t-test and get a p-value of 0.1826 (α = .05)—so we cannot say the former played better than the latter. (The beginner group's n is too small for statistical tests.) Although "beginner", "intermediate", and "advanced" categories are common labels applied to self-study courses and musical scores for amateur musicians, this result indicates that the ambiguity and subjectivity inherent in defining these categories make self-identified skill levels unreliable.

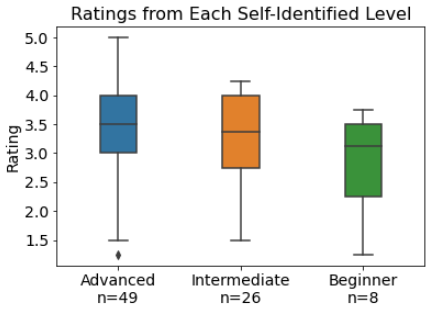

Figure 1. Comparing (averaged instructor) ratings among three self-identified groups.

4.2 Expert Evaluations

To examine how each instructor distributes their rating levels, we count the ratings at each level and compare their frequencies, as shown in Figure 2. It is clear that Instructors 1 and 2 tend to give high ratings more often than Instructors 3 and 4; the former also avoid giving the lowest rating almost completely. This indicates that the absolute rating values may be subjective and skewed.

Therefore, to measure how similarly these instructors rate, it makes more sense to compare ratings according to relative rather than absolute values. For this purpose we use Kendall's τ coefficient, which focuses on the rank correlation and can handle ties (with the tau-b version). Table

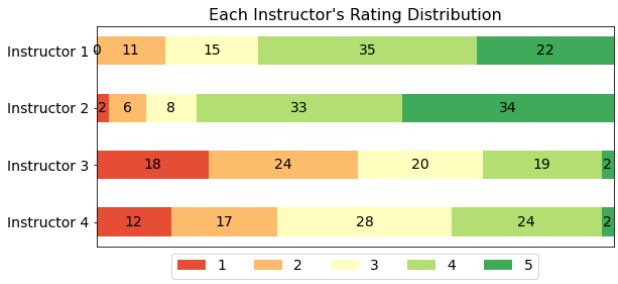

Figure 2. Comparing instructor rating distributions.

2 shows how each pair of instructors' ratings are associated with each other, with correlations sorted in descending order; all p-values are close to zero, statistically significant at the α = .01 level. These instructors show strong correlations (> .5) with one another, especially Instructor 3 and Instructor 4.

I3 & I4	I1 & I2	I2 & I4	I1 & I4	I1 & I3	I2 & I3
.806	.595	.563	.521	.514	.508

Table 2. Kendall's τ correlations (I=Instructor).

4.3 Peer Evaluations

As described in Section 3.2, 17 of the players also provided peer evaluations of the performances. To measure how the players' ratings compared to the instructors', we calculate the Kendall's τ correlation between ratings provided by each player and the average rating for the same recording subset provided by the instructors. Let's define k_p as the correlation for the pth player, where $p = 1, 2, \ldots, 17$. All k_p end up ranging between .401 and .741 (p-values < .01), with a mean of .542.

If we use k_p to represent the degree of "accuracy" of the pth player 's ratings, we can investigate the question of *whether better players are also better evaluators*. Let's define r_p as the average rating received by the pth player from all four instructors (for all pieces by this player). We use Spearman's ρ to measure how r_p and k_p are monotonically related, and the result is:

$$\rho_{r,k} = 0.152$$
$$p\text{-value} = 0.56$$

Although the correlation is positive, the large p-value prevents us from rejecting the null hypothesis that no relationship exists between how well individuals play and how accurately they rate performances.

5. QUALITATIVE ANALYSIS

5.1 Content Analysis and Annotation

To understand the evaluation criteria used by the instructors and the (novice) peer evaluators, we conduct a content analysis of their written comments [20]. The process involves first building an annotation model representing various evaluation criteria that appear in the text, and then

using this model to annotate each evaluation comment. (These two steps are iterative, as described later, in keeping with best practices for textual analysis [21].) For example, one of the comments—"The player didn't play the staccato notes in the left hand. No dynamic changes. A wrong note was played."—is annotated with *staccato*, *dynamic contrast*, and *wrong note* (terms that are then categorized under Articulation, Dynamics, and Note Accuracy respectively).

We use specialized text analytics software (QDA Miner, from Provalis Research) to construct the annotation model and to annotate each comment. To define appropriate annotations, we find recurring words and phrases (frequency ≥ 3), and look at each original comment in context to understand the intended meaning. For example, one of the most frequent phrases is "left hand", and one of its recurring contexts is that the left hand notes were played too loudly; therefore we create an annotation called *left hand loudness*. Other key words often associated with this aspect include "bass", as in "... I would like the bass [to] sound softer". By searching for related key words (e.g., "balance"), we have found similar contexts describing *right hand loudness* or just the *balance in general*. We group these conceptually related annotations under the same category called *Balance*. As we examine the contexts of these frequent words and phrases one-by-one, we create new annotations (and categories), and use them to annotate evaluator comments.

Building annotations and annotating comments is an iterative process: while examining the comments, we have discovered infrequent but useful key words like "8va" and "cresc" that we should search for. We sometimes carve out a new annotation from existing ones when observing enough cases to form a pattern (e.g., we have created a separate *tempo steadiness* annotation from *good tempo* and *inaccurate tempo*.) We have also spot-checked individual comments to make sure all evaluation criteria are sufficiently represented in our annotation model.

5.2 The Annotation Model

Figure 3 shows the annotation model developed from our dataset, containing 47 annotation terms arranged in 11 categories (and two subcategories). Many of these annotations can represent both positive and negative aspects of a performance: for example, *tempo steadiness* can be used to annotate both steady tempo and unsteady tempo. This is harmless, as our goal is to identify evaluation criteria, not the valence of the evaluations *per se*. A small handful of annotation terms exist only in the instructors' comments or only in the peers' comments, and these are mostly annotations in the Styles category: four styles are mentioned only by the instructors and five styles only by the peers. In addition, *dynamic shaping* (20 instances), *melodic shaping* (9 instances), and *rubato* (11 instances) only exist in the instructors' comments.

The annotation model derived from this dataset represents a diverse set of criteria, and it serves as a pool from which computers can select and generate measurements.

Tempo and timing
- inaccurate tempo
- good tempo
- tempo steadiness
- tempo contrast
- ritardando
- rubato
- pause

Note accuracy
- correct note
- wrong note
- missed note
- wrong octave

Phrasings
- phrasing
- melodic shaping

Pedal
- inaccurate pedal
- good pedal

Dynamics
- accurate dynamics
- inaccurate dynamics
- dynamic contrast
- dynamic shaping

Balance (between hands)
- balance in general
- left hand loudness
- right hand loudness

Articulation
- articulation in general
- legato
- staccato
- accent

Confidence
- confident or hesitant

Note connection
- choppy
- connectedness

Styles
- smooth
- heavy
- light
- abrupt
- crisp
- character
- lively
- flow
- lyrical
- mechanical
- style
- bland
- with emotion

Rhythm
- correct rhythm
- incorrect rhythm
- ***Notes too short***
  - shortened note
  - left hand too short
- ***Notes too long***
  - note too long

Figure 3. The annotation model. Lower case: annotations; bold: categories; italic: subcategories.

Many of the criteria are objective in nature: e.g., tempo change, note accuracy, and rhythm. These have low ambiguity and thus computational methods can detect them in a fairly straightforward manner; in fact, many traditional MIR techniques can be used for measuring these criteria. For example, we can easily track tempo changes based on audio-to-score alignment results (although deriving *perceived* tempo involves a few more parameters [22]). At the other end of the spectrum, however, criteria like confidence and style are very abstract, and thus are extremely hard for computers to detect. The rest of the criteria fall in the middle. Dynamics, phrasing, and articulation, for example, are directly linked to measurable features of the audio signal, but they involve many other parameters and can be subjective. Some literature addresses this duality (e.g., [23] on articulation and [24] on dynamic shaping), but there is no consensus on how to model such features, and attempts are scarcer than the more traditional MIR work mentioned above. Such aspects are almost never considered in computer-aided instrument learning applications.

5.3 Frequency of Evaluation Criteria

In the end, we have a total of 885 annotation instances for the instructors' comments, and 1015 for the peers' comments, averaging 2.7 and 2.2 annotations per comment respectively. We count the number of annotation instances under each annotation category separately for the instructors and for the peers, and calculate the frequency with which each annotation category was used by the two groups respectively. These percentages are shown in Figure 4. For both the instructors and the peers, Tempo, Note Accuracy, and Rhythm are the top three categories, accounting for just over 60% of the total annotations (although the peers describe Tempo more frequently and Rhythm less frequently than the instructors). For the remaining categories, Balance, Styles, and Confidence show the most difference between the two groups.

Balance is the fourth most common evaluation criterion found among the instructors' comments, accounting for 8.4% of the annotations, versus only 2.4% of the peers'. Balance between hands is a well-known challenge for piano beginners, and it can be difficult to notice on one's own; the percentage discrepancy between experts and novices suggests a promising opportunity for computer assistance. Meanwhile, Styles is the fourth most common criterion for the peer evaluators (9.4%), but it is only the eighth for the instructors (4.4%). Confidence accounts for 7.3% of peer annotations, but only 2.7% of instructor annotations. Styles and Confidence are both abstract concepts. For the instructors, these two are both ranked after Balance, Dynamics, Articulation, and Pedal, which represent concrete piano techniques, and they occupy 28.2% in total. In contrast, all these four categories have lower percentages for the peers, and they occupy only 18.2% in total. This discrepancy implies that as compared to experts, novices might be more likely to judge a performance using abstract concepts, while experts tend to point out specific piano techniques.

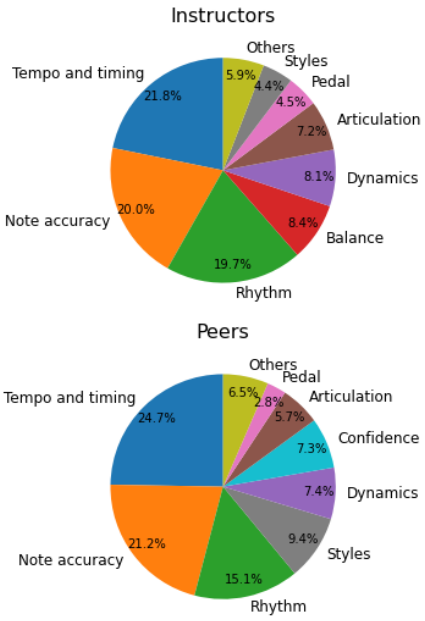

Figure 4. Comparing annotation category percentages.

We further investigate the usage of the four "technique" criteria and the two "abstract" criteria just described by comparing three groups among the peer evaluators: the four peer players whose performances received the lowest average ratings, the four peer players whose performances received the highest average ratings, and the four ranked in the middle. For each group, we calculate the usage percentages as above, and focus on comparing the six criteria. Figure 5 shows the comparison among the three groups, as well as how they compare to the instructors. The left bars indicate a consistent positive association between piano skill levels and the usage of piano technique criteria. Although the (opposite) trend of the right bars is less consistent, as the middle group used abstract criteria less frequently than the higher-skilled group, the lower-skilled group indeed used a significantly higher percentage of abstract criteria than the average of all 17 peer evaluators (16.7%). This suggests that lower-skilled piano players lack the ability to pin down specific piano techniques involved in a performance, and their evaluation criteria tend to be correspondingly more general and abstract, e.g., "There is some hesitancy in the chords. The bass clef chords are a bit abrupt".

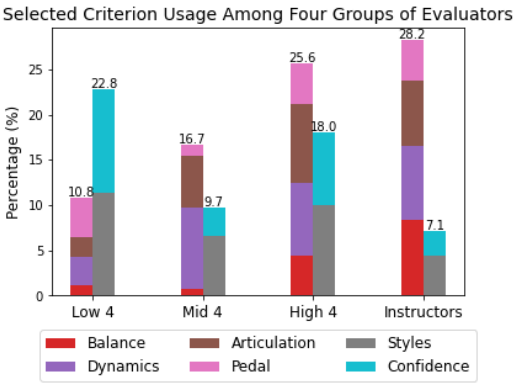

Figure 5. Comparing piano technique criteria (left) and abstract criteria (right) usage among three groups of peers and the instructors.

6. DISCUSSION

6.1 Do Instructors and Players Evaluate Performances Differently?

In terms of numerical ratings, the consistent high correlations between the peer evaluations and the average of the instructor evaluations (Section 4.3) suggest that even novices have a reliable sense of what good or poor performance is like. In terms of evaluation criteria, both the peers and the instructors use Tempo, Note Accuracy, and Rhythm the most, accounting for a little over 60% of the total comment annotations from both groups. However, beyond these top three criteria, the two groups exhibit different patterns: the peers tend to use more abstract and general criteria like Confidence while the instructors use more concrete and specific piano techniques like Balance (between hands). It is unsurprising that instructors would use more technical criteria, given their own training and teaching experience, and computer-generated feedback based on these criteria could be particularly illuminating to individuals seeking to improve their playing.

6.2 Are Better Players Also Better Evaluators?

For the sake of this discussion, we define good evaluations as evaluations similar to the ones done by the instructors. In terms of *ratings*, we do not find enough evidence confirming better player are also better evaluators—even players of poor performances can provide accurate ratings. However, in terms of evaluation *criteria*, we have found evidence that higher-skilled players tend to provide better

evaluations. Specifically, they are more capable of judging a performance based on piano techniques, which are in line with piano instructors' evaluations.

6.3 Computer-Aided Feedback

These results suggest that there is a consistent standard of good and poor performances—at least for beginner pieces. However, numerical ratings have limited value for helping students learn. In fact, students take private music lessons not to be given a rating, but to seek specific formative feedback for improvement. Building a machine that can provide comparable feedback would offer much greater value to end users.

Much of the terminology in the annotation model is score-dependent, verifying whether or not the performer has followed elements in the sheet music. The basic elements are notes, rhythm, and tempo/timing, which also correspond with the top three evaluation criteria. Relevant MIR tasks for detecting such errors include music transcription [25], source separation [26], and audio-to-score alignment [27]. Other typical elements in the sheet music are dynamics, articulation, and pedaling, and some attempts (such as [23] [28] [29]) have been made at modeling and detecting them.

However, not every element or aspect worth evaluating is explicitly indicated in the score. For example, pedaling and balance between hands are often only implied in the score, and can also be up to personal interpretation by the performer. In such instances, text-based feedback can be of limited utility, and what a computer may be able to do more effectively is provide visualized feedback. The value of such feedback lies in making implicit aspects of a performance explicit to the player, rather than instructing the player what to do. For example, the computer could show a tempo curve indicating (intentional or unintentional) tempo changes. Such visualizations can be especially helpful to beginners, who might not be able to notice such aspects easily.

6.4 Peer Evaluation for Education

At the end of each peer evaluation session, we asked the evaluator two open-ended questions: "How do you feel after listening to so many recordings in a row?" and "How do you feel about this process compared to how you evaluate your own playing?" Most evaluators expressed that it was a positive experience, with words like "fun", "very interesting", and "enjoyed it". A couple of them mentioned they were able to pay more attention to the elements in the sheet music when evaluating others. Four of them indicated that the process of comparing multiple recordings helped them judge their own playing better. This overall positive response suggests that there is educational potential for peer evaluation platforms where piano learners could anonymously give each other feedback.

6.5 Limitations

There are some inherent limitations in this project. First, the pieces we focus on are all at the beginner level, meaning they are relatively short and involve relatively few sophisticated piano techniques. It is possible that our findings might not apply to performances (or evaluations) of more advanced pieces. Second, the recording process might involve some bias against more advanced players who felt confident and could sight read, and thus did not prepare as much as the beginners. Third, the size of the dataset is relatively small. This facilitated our process of careful, manual content analysis, but imposes some limits on the statistical analysis.

7. CONCLUSIONS AND FUTURE WORK

In this paper, we contribute a unique dataset of amateur piano performance recordings and corresponding expert and peer evaluations. This dataset allows for interesting multifaceted analysis of nuances in the peer evaluations, because the peer evaluators are also players, and both quantitative and qualitative evaluations are recorded. Through the initial analyses presented in this paper, we find that even novices exhibit reliable judgement at distinguishing good performances from poor ones, but higher-skilled novices tend to base their judgement on piano techniques (as experts do), while lower-skilled novices rely on more subjective and/or abstract impressions. Most evaluation criteria used by experts are concrete, and are therefore precisely the kind that can be detected and measured by software evaluating an audio signal and its relationship to the score. Visualizing these aspects could provide valuable assistance to beginners seeking constructive insights on their playing. Despite some limitations to the generalizability of its findings, this paper lays the groundwork for building more advanced computer-aided instrument learning software. In future work, we plan to combine the audio-to-score alignments in this dataset with other MIR techniques to derive specific measurements reflecting experts' evaluation criteria. Once that is achieved, we then plan to compare those computer-generated evaluations of these recordings (including measurements and/or visualizations) to the human annotations.

8. ACKNOWLEDGMENTS

We extend our sincere thanks to multiple individuals for their contributions to this project. First, we thank the piano players and the piano instructors for participating in our experiments. Second, we thank Joon Han and Liz Smith for helping record the piano performances. Third, we thank Caitlin Sales for creating the demo website. We would also like to thank the reviewers of this paper for their detailed and constructive feedback.

9. REFERENCES

[1] Yousician website. https://yousician.com/.

[2] Simply Piano website. https://www.hellosimply.com/simply-piano.

[3] G. Percival, Y. Wang, and G. Tzanetakis, "Effective use of multimedia for computer-assisted musical instrument tutoring," in *Proceedings of the International Workshop on Educational Multimedia and Multimedia Education*, 2007, pp. 67–76.

[4] V. Eremenko, A. Morsi, J. Narang, and X. Serra, "Performance assessment technologies for the support of musical instrument learning," in *Proceedings of the 12th International Conference on Computer Supported Education (CSEDU)*, 2020, pp. 629–640.

[5] C. Dittmar, E. Cano, J. Abeßer, and S. Grollmisch, "Music information retrieval meets music education," in *Dagstuhl Follow-Ups*, vol. 3. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2012.

[6] P. Kasák, R. Jarina, and M. Chmulík, "Music information retrieval for educational purposes-an overview," in *2020 18th International Conference on Emerging eLearning Technologies and Applications (ICETA)*. IEEE, 2020, pp. 296–304.

[7] R. B. Dannenberg, M. Sanchez, A. Joseph, P. Capell, R. Joseph, and R. Saul, "A computer-based multimedia tutor for beginning piano students," *Interface*, vol. 19, no. 2-3, pp. 155–173, 1990.

[8] A. Arzt, S. Böck, S. Flossmann, H. Frostel, M. Gasser, and G. Widmer, "The complete classical music companion v0. 9," in *Proceedings of the AES International Conference on Semantic Audio, London, UK*, 2014, pp. 18–20.

[9] F. Tsubasa, Y. Ikemiya, K. Itoyama, and K. Yoshii, "A score-informed piano tutoring system with mistake detection and score simplification," in *Sound and Music Computing Conference*, 2015.

[10] S. Ewert, S. Wang, M. Müller, and M. Sandler, "Score-informed identification of missing and extra notes in piano recordings," in *Proceedings of the 17th International Society for Music Information Retrieval (ISMIR) Conference*, 2016.

[11] B. F. Yuksel, K. B. Oleson, L. Harrison, E. M. Peck, D. Afergan, R. Chang, and R. J. Jacob, "Learn piano with bach: An adaptive learning interface that adjusts task difficulty based on brain state," in *Proceedings of the CHI conference on human factors in computing systems*, 2016, pp. 5372–5384.

[12] A. Lerch, C. Arthur, A. Pati, and S. Gururani, "An interdisciplinary review of music performance analysis," *Transactions of the International Society for Music Information Retrieval*, vol. 3, no. 1, 2020.

[13] C. E. Cancino-Chacón, M. Grachten, W. Goebl, and G. Widmer, "Computational models of expressive music performance: A comprehensive and critical review," *Frontiers in Digital Humanities*, vol. 5, p. 25, 2018.

[14] C. Cancino-Chacón, S. Peter, S. Chowdhury, A. Aljanaki, and G. Widmer, "On the characterization of expressive performance in classical music: First results of the con espressione game," *arXiv preprint arXiv:2008.02194*, 2020.

[15] S. Thompson and A. Williamon, "Evaluating evaluation: Musical performance assessment as a research tool," *Music Perception*, vol. 21, no. 1, pp. 21–41, 2003.

[16] B. C. Wesolowski, S. A. Wind, and G. Engelhard Jr, "Examining rater precision in music performance assessment: An analysis of rating scale structure using the multifaceted rasch partial credit model," *Music Perception: An Interdisciplinary Journal*, vol. 33, no. 5, pp. 662–678, 2016.

[17] D. Blom and K. Poole, "Peer assessment of tertiary music performance: Opportunities for understanding performance assessment and performing through experience and self-reflection," *British Journal of Music Education*, vol. 21, no. 1, pp. 111–125, 2004.

[18] J. W. Bastien, *The older beginner piano course*. Kjos West, 1977.

[19] C. Raphael, "Automatic segmentation of acoustic musical signals using hidden Markov models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 21, no. 4, pp. 360–370, 1999.

[20] T. G. Harwood and T. Garry, "An overview of content analysis," *The marketing review*, vol. 3, no. 4, pp. 479–498, 2003.

[21] Q. Deng, M. J. Hine, S. Ji, and S. Sur, "Inside the black box of dictionary building for text analytics: a design science approach," *Journal of international technology and information management*, vol. 27, no. 3, pp. 119–159, 2019.

[22] K. Seyerlehner, G. Widmer, and D. Schnitzer, "From rhythm patterns to perceived tempo," in *Proceedings of the 8th International Society for Music Information Retrieval (ISMIR) Conference*, 2007.

[23] R. Bresin and G. Umberto Battel, "Articulation strategies in expressive piano performance analysis of legato, staccato, and repeated notes in performances of the andante movement of Mozart's sonata in g major (k 545)," *Journal of New Music Research*, vol. 29, no. 3, pp. 211–224, 2000.

[24] G. Widmer and A. Tobudic, "Playing Mozart by analogy: Learning multi-level timing and dynamics strategies," *Journal of New Music Research*, vol. 32, no. 3, pp. 259–268, 2003.

[25] E. Benetos, A. Klapuri, and S. Dixon, "Score-informed transcription for automatic piano tutoring," in *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*. IEEE, 2012, pp. 2153–2157.

[26] S. Ewert and M. Müller, "Score-informed source separation for music signals," in *Dagstuhl Follow-Ups*, vol. 3. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2012.

[27] R. Agrawal and S. Dixon, "A hybrid approach to audio-to-score alignment," *arXiv preprint arXiv:2007.14333*, 2020.

[28] K. Kosta, "Computational modelling and quantitative analysis of dynamics in performed music," Ph.D. dissertation, Queen Mary University of London, 2017.

[29] B. Liang, G. Fazekas, and M. B. Sandler, "Detection of piano pedaling techniques on the sustain pedal," in *Audio Engineering Society Convention 143*. Audio Engineering Society, 2017.