Computational Analysis of Audio Recordings of Piano Performance for Automatic Evaluation *

Norihiro Kato[1], Eita Nakamura[2], Kyoko Mine[3], Orie Doeda[1], and Masanao Yamada[1]

[1] National Institute of Technology, Kushiro College, Japan
ym@kushiro-ct.ac.jp
[2] Kyoto University, Japan
[3] Osaka Ohtani University, Japan

Abstract. We developed a computational evaluation method for piano performance with the goal of building a practice support system for beginners. We recorded students' performances as audio data and applied several recent methods for audio-to-MIDI transcription based on deep neural networks to extract the pitch, onset time, and offset time of musical notes. To determine the correctness of the performance, we aligned the extracted MIDI data with the musical score using a hidden Markov model (HMM). We compared the audio-to-MIDI transcription methods and optimized the weight on different types of performance errors to conform to teacher's assessment. Our experiments showed a strong correlation between the rate of performance errors obtained from the alignment and the evaluation by a teacher who listened to the performance. The results that indicate performance errors and tempo stability can be used in a practice support system that provides feedback to learners.

Keywords: Computational performance evaluation · Hidden Markov model · Audio-to-MIDI transcription.

1 Introduction

Effective performance evaluation plays a crucial role in music education, as it enables learners to receive valuable feedback for improving their skills [7]. However, traditional methods often fall short in providing timely and personalized feedback, particularly in large classrooms or for remote learning situations. In recent years, there has been a growing interest in applying technology to evaluate the performance of musical instruments [2, 7]. Research efforts have explored various approaches, aiming to support learners in their musical journey. One of these methods includes visualization of performances [8], which has seen advancements in incorporating augmented reality technology [1, 4].

Among the diverse range of musical instruments, the piano has many performers at varying skill levels, which results in differing performance evaluation

* This work was supported by JSPS KAKENHI Grant Numbers 21K02846, 21K12187, 22H03661.

criteria across studies — from basic elements such as pitch and rhythm in beginners, to musical expression in advanced players. In previous research [12], an overall evaluation was performed for each full piece performed by beginners using acoustic data and support vector regression (SVR) models. Despite these advancements in performance evaluation techniques, challenges still remain in leveraging technology for effective formative assessment. Formative assessment, an evaluation of learning achievement during the instructional process, allows for adjustments in learning activities and teaching methods. As each individual has different performance aspects that need to be improved, it can be difficult for teachers to provide sufficient advice to every student in a class.

In response to these challenges, our study aims to enhance formative assessment in piano performance education for beginners by contrasting score data with transcriptions of audio recordings in MIDI (Musical Instrument Digital Interface) format. Our focus is on individual elements of performance such as wrong touch, re-playing, and tempo consistency for personalized evaluation. With these considerations in mind, our central research question is: How closely can computationally analyzed performance errors align with teachers' evaluations? This investigation could pave the way for a feedback system that utilizes computationally analyzed errors for formative assessment.

2 Related work

Active research on performance evaluation of piano has explored various methods to provide feedback to learners within the music education context [5]. Fukuda et al. [3] developed a system that detects mistakes in the user's performance and simplifies challenging sections of a musical score based on the user's skill level. Wang et al. [11] proposed two audio-based piano performance evaluation systems for beginners using deep neural networks (DNN). DNN-based piano performance evaluations use end-to-end processing, eliminating the need for sheet music. However, for formative evaluation, which requires assessing each note, musical note information needs to be transformed from the audio signal to compare with sheet music.

Extracting musical note information from audio performance data is a challenging problem in the area of music information retrieval, especially for polyphonic music such as piano music. Recently, accurate methods have been realized using machine learning techniques, including PovNet [10] based on a dual DNN architecture for pitch estimation and onset and velocity (intensity) estimation, and the high-resolution MIDI transcription (HighReso) method [6] based on convolutional-recurrent neural network. While these MIDI transcription methods are usually evaluated on relatively "clean" piano recording data, their performance for low-quality recordings common in practical situations of music education has not been investigated well in the literature. In this study, we use both two methods as a part of the automatic evaluation method and compare the results.

3 Method

The data we used in this study comes from the final exam of group piano lessons conducted in the Early Childhood Education major at a certain university's Faculty of Education. In the exam, the students perform one piece each from Beyer Op. 101 and children's songs, and one instructor evaluated their performances. The criteria for this exam were (a) wrong touch and re-playing, (b) rhythm, and (c) dynamics on a scale from A to C (A+, A, A–, B+, B, B–, C).

The performances in the exam were recorded as audio data using an iPad, which was placed in a concert hall with a total area of 1,741 square meters and 217 seats, positioned 3.23 meters in front of the grand piano and at a height of 0.65 meters. To convert the audio data into MIDI, we employed two recent methods for MIDI transcription, PovNet [10] and HighReso [6].

After the transcription, we prepared the MIDI data of the sheet music manually, and an HMM was employed to align the MIDI data from both the performance and the sheet music. Benefitting from the HMM's ability to discern correctly and incorrectly played notes [9], we calculated their proportions relative to the total number of played notes. Since the wrong touches and re-playing of evaluation criterion (a) corresponds to pitch errors and extra notes, respectively, we define the evaluation score E as

$$E = w_p E_p + (1 - w_p) E_e, \tag{1}$$

where E_p denotes the pitch error rate, E_e the extra note rate, and w_p the weight for pitch errors. We measured the correspondence between the teacher's evaluation and the evaluation score E using the Spearman's rank correlation coefficients while varying the weight w_p.

For the variation in rhythm, we calculated the relative speed of each note in the performance locally by dividing the aligned sheet music's Inter-Onset Interval (IOI), by the performance IOI. IOI is the time interval between the onsets of consecutive musical notes. We took the median of this local tempo list as the reference tempo for the performance, and used the root mean square of the relative errors between the local tempo and the reference tempo as an indicator of playing speed consistency. Extra notes that did not have corresponding sheet music notes were excluded from this calculation.

4 Results and Discussion

A total of 29 students from three classes took the exam. The results of comparison between the transcription methods, PovNet and HighReso, are shown in Fig. 1. The results for PovNet consistently showed a high correlation with teacher evaluations, with the correlation coefficient reaching its maximum value of 0.86 when $w_p = 0.58$. For HighReso, the maximum value was 0.77 when $w_p = 0.905$. In the following discussions, PovNet with $w_p = 0.58$ is used.

The relationship between the evaluation score E and the teacher evaluations is shown in Fig. 2. Overall, performances with a lower score have higher teacher

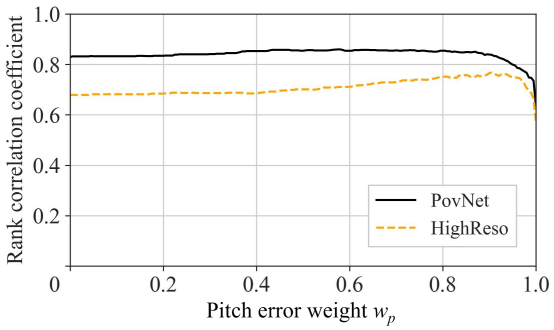

Fig. 1. Comparison of audio-to-MIDI transcription methods

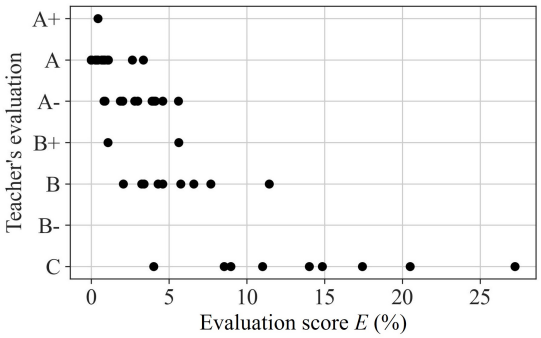

Fig. 2. Relationship between incorrect notes and teacher's evaluation

evaluations, and vice versa. Below, we discuss some of the performances that deviate from this trend.

The performance at the right end of the B evaluation (Bayer Op.101, No.78) has a higher evaluation than some of C-graded performances, despite having more performance errors indicated by the score E. To examine the cause, Fig. 3 displays the progression of the performance compared to the sheet music. Consecutive extra notes indicate re-plays, and blue plus signs scattered among black dots are due to pitch detection errors by PovNet. Extra notes that cannot be associated with a position in the sheet music are plotted at the zero point of score time. In the figure, a significant number of extra notes are detected because of re-playing at approximately 23 seconds of score time. It is likely that the teacher considered this as a single mistake, and despite the large number of detected errors, still awarded a B evaluation.

The rightmost data point in the C evaluation of Fig. 2, representing the children's song 'Crickets', is another intriguing case. Although the performance errors are comparable to the average of B-rated performances, it falls into the C rating category. Fig. 4 illustrates the progression of this performance, which

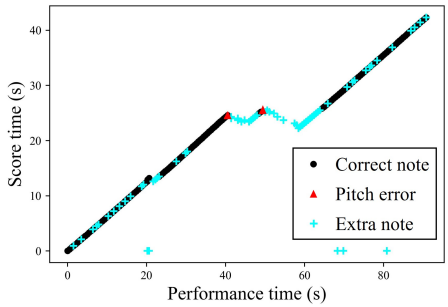

Fig. 3. Performance progression of Bayer Op.101, No.78

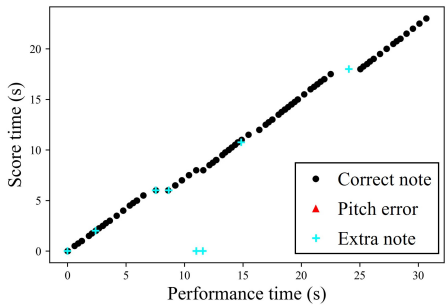

Fig. 4. Performance progression of "Crickets"

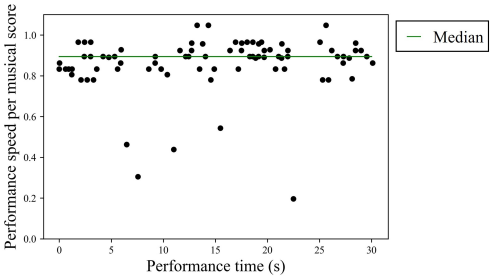

Fig. 5. Tempo deviation of "Crickets"

contains hardly any extra notes, but reveals gaps starting at approximately the 7-second and 23-second in the performance time. To illustrate this brief pause clearly, Fig. 5 shows the rhythmic variability calculated by the method described in the previous section, for the same performance. The values below 1 indicating a slower performance than the sheet music. From this figure, it is easy to discern the points where the hands have momentarily stopped.

5 Conclusion

In this research, we explored a novel approach to piano performance evaluation by utilizing advanced machine learning methods, applying PovNet and HighReso

for audio-to-MIDI transcription in real-world classroom settings. We found that PovNet demonstrated superior performance and a significant correlation with teacher evaluations. We also observed deviations through the data analysis, and were able to identify and visualize contributing factors to these deviations, such as re-plays and brief pauses. Our findings demonstrate the potential of machine learning in performance evaluation to provide personalized and timely feedback.

Future research directions include developing an interactive feedback system for music practice, aiding learners in error identification and correction. Additionally, conducting a multi-year longitudinal study can assess the long-term effects of technology-enhanced formative assessment on students' musical development, achievement, motivation, self-efficacy, and engagement.

References

1. Deja, J.A.: Piano learning and improvisation through adaptive visualisation and digital augmentation. Companion Proceedings of the 2022 Conference on Interactive Surfaces and Spaces pp. 41–45 (2022)
2. Dorfman, J.: Theory and Practice of Technology-based Music Instruction. Oxford University Press (2022)
3. Fukuda, T., Ikemiya, Y., Itoyama, K., Yoshii, K.: A score-informed piano tutoring system with mistake detection and score simplification" within the music education contexts. Proceedings of the 12th Sound and Music Computing Conference (SMC) **1**, 105–110 (2015)
4. Heyen, F., Ngo, Q.Q., Kurzhals, K., Sedlmair, M.: Data-driven visual reflection on music instrument practice. ACM CHI Conference on Human Factors in Computing Systems (2022)
5. Kim, H., Ramoneda, P., Miron, M., Serra, X.: An overview of automatic piano performance assessment within the music education contexts. Proceedings of the International Society for Music Information Retrieval **1**, 465–474 (2017)
6. Kong, Q., Li, B., Song, X., Wan, Y., Wang, Y.: High-resolution piano transcription with pedals by regressing onset and offset times. IEEE/ACM Transactions on Audio, Speech, and Language Processing **29**, 3707–3717 (2021)
7. Lerch, A., Arthur, C., Pati, A., Gururani, S.: An interdisciplinary review of music performance analysis. Transactions of the International Society for Music Information Retrieval **3**(1), 221–245 (2021)
8. Lima, H.B., Santos, C.G.R.D., Meiguins, B.S.: A survey of music visualization techniques. ACM Computing Surveys (CSUR) **57**(7), 1–29 (2022)
9. Nakamura, E., Yoshii, K., Katayose, H.: Performance error detection and post-processing for fast and accurate symbolic music alignment. Proceedings of the International Society for Music Information Retrieval pp. 347–353 (2017)
10. Shibata, K., Nakamura, E., Yoshi, K.: Non-local musical statistics as guides for audio-to-score piano transcription. Information Sciences **566**, 262–280 (2021)
11. Wang, W., Pan, J., Yi, H., Song, Z., Li, M.: Audio-based piano performance evaluation for beginners with convolutional neural network and attention mechanism. IEEE/ACM Transactions on Audio, Speech, and Language Processing **29**, 1119–1133 (2021)
12. Wu, C.W., Gururani, S., Pati, A., Vidwans, A.: Towards the objective assessment of music performances. International Conference on Music Perception and Cognition (ICMPC) pp. 99–103 (2016)