Designing a Music Difficulty Measure for Controllable Automatic Piano Rearrangement

Hikari Miyaji, Keito Sawada, Wen-Chin Huang and Tomoki Toda
Nagoya University, Japan
E-mail: miyaji.hikari@g.sp.m.is.nagoya-u.ac.jp

***Abstract*—Automatic piano rearrangement systems for controlling score difficulty can support players without musical rearrangement skills and help to increase their motivation to practice. While the existing systems allow difficulty adjustment in the predefined levels, more flexible control—such as local arrangement of only difficult passages and continuous or multifaceted difficulty scaling—is desirable. The conventional methods for designing difficulty measures and applying them to difficulty estimation rely on information dependent on specific score data formats and on whole score information, limiting their applicability for the data formats that can be used and making local difficulty assessment impossible. In this paper, we propose a music difficulty measure (DM) for estimating the difficulty of piano scores to address these issues. DM consists of seven features that are independent of data format and whole score information, making them applicable both as bar-level measures (DM-bar) and as whole-score measures (DM-whole). We also propose a linear regression-based music difficulty estimation method using DM: a linear regression model is trained on DM-whole to predict whole difficulty, and the same model is applied to DM-bar to estimate bar-level difficulty. In addition, statistics of the estimated bar-level difficulties over a whole score are further integrated into the model to improve the estimation accuracy. The experimental evaluation demonstrates the effectiveness of the proposed measure on the music difficulty estimation task.**

I. Introduction

Rearranging piano scores towards different difficulties to match the performance skill of piano players can increase their motivation to practice. Since rearrangement requires a certain degree of specialized musical knowledge, self-rearrangement is usually difficult for most players. Therefore, automatic rearrangement systems that can control the difficulty of existing scores are desired.

There has been developed the automatic piano rearrangement system capable of adjusting the difficulty into four levels [1]. However, the performance skill of individual players can vary drastically due to their experiences and physical conditions. For example, people with small hands may find it hard to press distant notes simultaneously, while those who are not accustomed to moving their fingers quickly may struggle to play at fast tempos. Therefore, there is a need for more flexible control, such as local arrangement of only difficult passages and continuous or multifaceted difficulty scaling.

There are also some approaches for estimating the difficulty of piano scores that use deep learning [2]–[4]. Despite the high accuracy, these methods lacked interpretability as they did not explicitly reveal factors which led to the estimation, hindering their applicability for multifaceted piano rearrangement. On the other hand, some studies attempted to analyze difficulty using manually defined features, which were based on the experience of pianists and were easy to interpret for piano players [5], [6]. However, they operated on the song level, thus failed to measure local difficulty. Also, some features were not obtainable in certain data formats.

In this paper, we propose a music difficulty measure for highly controllable automatic piano rearrangement. Specifically, we design a set of features that capture local difficulty by using only information obtained from each bar, without relying on information from the whole score. This approach enables compatibility with a wider range of score data formats. The proposed difficulty measure incorporates features defined based on practical piano performance experience, allowing for the analysis of score difficulty from both local and multifaceted perspectives. Furthermore, we propose a method for estimating both whole-level difficulty and bar-level difficulty using the proposed measure, and conduct experimental evaluation. Our results demonstrate that local (bar-level) difficulty is a crucial factor in estimating whole-level difficulty, and that the proposed measure is effective in capturing local difficulty.

II. Related Works

A. Automatic Piano Rearrangement for Changing Difficulty Levels

There are several studies on controlling the difficulty of piano scores. For example, some methods convert piano ensemble scores into solo scores [7], [8], and others detect user performance errors and simplify solo piano scores based on these errors [9]. All of these methods focus on reducing difficulty by removing notes from existing scores, and do not consider approaches for increasing the difficulty of a score.

To address this, Suzuki [1] developed an automatic rearrangement system that enables not only decreasing but also increasing the difficulty of piano solo scores by converting them into token representations and training a difficulty conversion model. In this system, in order to enable conversion across multiple difficulty levels with a single model, pairs of scores of the same piece at different difficulty levels are used for training the model to perform bidirectional difficulty conversion. However, this approach only allows conversion within predefined difficulty levels, making it difficult to achieve highly flexible and multifaceted difficulty control.

979-8-3315-7206-8/25/$31.00

B. Designing a Music Difficulty Measure for Piano Score

A specific measure that can capture the difficulty of piano scores is required to achieve highly flexible and multifaceted difficulty control. Sébastien et al. [5] proposed a difficulty measure consisting of seven features, which can be automatically extracted from MusicXML score data. They compared the results of difficulty classification using their proposed measure with those obtained by principal component analysis, and also conducted subjective evaluations with pianists. These results demonstrated that the proposed measure is effective for classifying the difficulty of piano scores. On the other hand, a limitation of this measure is its restriction to specific score data formats. Some of its features rely on information unique to MusicXML data (such as polyrhythms, where different rhythms are played simultaneously), and thus cannot be applied to other data formats. As a result, the range of applicable datasets is limited. In particular, most publicly available datasets—such as POP909, which contains 909 popular song piano arrangements [10], and PiJAMA, which includes 2,777 performances by jazz pianists [11]—use the MIDI format, which is easier to collect. Therefore, by designing features that do not depend on specific data formats, it is possible to broaden the applicability of difficulty measures. Another issue is that the conventional measure includes the total number of bars in the score, which makes it unsuitable for use as a local difficulty measure. In order to achieve highly controllable automatic piano rearrangement, it is necessary to analyze the score in finer temporal segments, which requires the design of difficulty measures that do not depend on information from the whole score.

Chiu et al. [6] estimated the difficulty of piano scores using not only note information and metadata that can be directly extracted from MIDI data, but also eight custom-designed features. Among these proposed features, some focus on the relationship between the left and right hands, as well as on fingering. However, since not all score data include explicit annotations for hand parts or fingering, such features inevitably restrict the applicability of the difficulty measure.

III. Proposed Music Difficulty Measure and Music Difficulty Estimation Method

A. A Music Difficulty Measure

To address the limitations of conventional difficulty measures—namely, their restriction to specific score data formats and their lack of applicability as local difficulty measures—we propose a new difficulty measure (DM) that enables local difficulty estimation with minimal dependence on the score data format.

The DM consists of seven features, as shown in Table I, which are determined based on input from piano instructors and experienced players, as well as by reference to the conventional difficulty measures [5], [6]. For all features, lower values indicate lower difficulty, while higher values correspond to higher difficulty, making the measure intuitively interpretable. To broaden the applicability of the DM, each feature is calculated using only basic note information, such as note duration and pitch, and metadata including tempo. By restricting the use of information extractable only from specific data formats—such as fingering or articulations—the DM can be applied not only to MusicXML scores but also to MIDI scores. Furthermore, since the DM does not include features that require information about the entire whole score (e.g., the total length of the piece), it can be used as a local difficulty measure. In this study, the score is divided into individual bars, and the DM is used to represent the difficulty of each bar (DM-bar). Additionally, to directly compare the bar-level difficulty and the whole score difficulty, we assume that the average bar difficulty corresponds to the global difficulty. Therefore, the overall difficulty measure (DM-whole) is defined as the average value of DM-bar across all bars.

TABLE I
Summary of the proposed difficulty measure.

Features	Abstract
Notes	Number of notes within the estimated range
SN	Average number of notes played simultaneously
Key	Number of sharps or flats indicating the key
Accidental	Ratio of accidentally marked notes to total number of notes
Move	Sum of (pitch difference from next note / note duration)
Spread	Pitch difference between highest and lowest notes at the same time
QPM	Playing speed in quarter notes

The definitions of each feature of the DM are shown below.

1) Notes: Notes represents the number of notes within the specified difficulty estimation range. Specifically, in DM-bar, *Notes* refers to the number of notes contained within each bar, while in DM-whole, it denotes the average number of notes per bar.

2) SN (Simultaneous Notes): SN represents the average number of notes that begin sounding simultaneously. This is given by (1):

$$\mathrm{SN} = \frac{1}{T}\sum_{i=1}^{T} c(i), \tag{1}$$

where T denotes the number of onset times, and $c(i)$ is the number of notes that begin sounding at onset i.

3) Key: Key indicates the tonality of a piece and is represented on the score by the number of sharps or flats. In the DM, *Key* simply corresponds to the number of sharps or flats in the key signature, taking an integer value between 0 and 7. As the number of *Key* increases, the use of black keys also increases, generally requiring more complex fingerings and thus resulting in higher performance difficulty.

4) Accidental: Accidentals are symbols—such as sharps, flats, and naturals—attached to notes to alter their pitch. These symbols are used to notate pitches that are not included in the key signature specified for the piece. In the DM, the set of keys normally used is determined by the key signature, and if a note's pitch does not correspond to one of these keys, the note is considered to have an accidental, regardless of whether an explicit accidental symbol appears in the score. Thus, it should be noted that some notes are possibly counted as accidentals even if no accidental symbol is present in the notation. Let

a denote the number of notes with accidentals. Using *Notes*, *Accidental* is given by (2).

$$\text{Accidental} = \frac{a}{\text{Notes}} \tag{2}$$

5) Move: Move represents the speed at which both hands move horizontally while performing the score. The specific calculation method is described as follows.

First, for each pair of temporally adjacent notes, the pitch width p is calculated using the MIDI note number. A MIDI note number is an integer value assigned to each key, with middle C (C4) of an 88 note piano-style keyboard defined as 60, and the value increases or decreases by one for each semitone. There are four possible cases for note transitions: a single note to a single note, a single note to multiple notes, multiple notes to a single note, and multiple notes to multiple notes. For a transition from a single note to a single note, p is defined as the absolute value of the difference between their respective MIDI note numbers. For transitions from multiple notes to a single note, the note among the multiple notes whose pitch is closest to the single note is selected, and the absolute value of their pitch difference is taken as p. The same approach applies to transitions from a single note to multiple notes. For transitions from multiple notes to multiple notes, p is defined as the sum of the absolute values of the pitch differences between the lowest notes and the highest notes in each group. If the time interval between the onset of one note and the next is greater than or equal to half a bar, p is set to 0. This is based on the assumption that when there is sufficient time before the next note, the impact on performance difficulty is minimal.

The value of *Move* is defined by (3):

$$\text{Move} = \sum_{i=1}^{T} \frac{p_i}{d}, \tag{3}$$

where T is the number of note onset times, p_i is the pitch width corresponding to the note sounding at onset i, and d is the duration of the preceding note on the time axis. The duration is normalized such that a value of 1 corresponds to one quarter of a bar. Dividing p by d means that shorter note durations result in higher values of *Move*, indicating increased difficulty. This reflects the fact that when the preceding note has a shorter duration, a quicker hand movement is required, thus increasing the performance difficulty.

6) Spread: Spread is defined as the average pitch difference between the highest and lowest notes played simultaneously. Specifically, it is given by (4):

$$\text{Spread} = \frac{1}{T} \sum_{i=1}^{T} \left(n(A_i^{\text{high}}) - n(A_i^{\text{low}}) \right), \tag{4}$$

where $n(\cdot)$ denotes the MIDI note number. A_i represents the set of notes played simultaneously at onset i, with A_i^{high} and A_i^{low} denoting the highest and lowest notes in A_i, respectively. If there are no instances where multiple notes are played simultaneously, *Spread* takes its minimum value of 0.

7) QPM (Quarters Per Minute): QPM represents the tempo at which a piece is performed. In this study, the duration of one beat is defined as that of a quarter note, and *QPM* is specified as the number of beats per minute.

B. A Linear Regression-Based Music Difficulty Estimation Method

In this study, we evaluate the effectiveness of DM-whole and DM-bar through the estimation of musical difficulty. To achieve this, we propose a difficulty estimation method using a linear regression model based on DM, enabling a continuous representation of difficulty. For model training, we use widely available score data that includes whole-level difficulty annotations for each piece. An overview of the proposed method is presented in Fig. 1.

First, we train a linear regression model (LR model (1)), based on the least squares method, to estimate the whole-level difficulty (difficulty-whole) using DM-whole, which is extracted from the score data, as the explanatory variable. For the estimation of bar-level difficulty (difficulty-bar), since DM-whole is the mean of DM-bar and a linear relationship can be assumed, we use the value obtained by directly inputting DM-bar into this model as the estimate for bar-level difficulty.

It should be noted that existing score data do not include annotations for bar-level difficulty, making it difficult to directly assess the effectiveness of DM-bar. Therefore, in this paper, we indirectly evaluate the effectiveness of bar-level difficulty from the perspective of its contribution to the estimation of whole-level difficulty. Specifically, we estimate bar-level difficulty by inputting DM-bar into the aforementioned linear regression model, and then, for each score, calculate bar statistics of the estimated bar-level difficulties consisting of the maximum, mean, minimum, and the difference between the maximum and minimum values. We then train a new linear regression model (LR model (2)) to estimate the whole-level difficulty of each score using 11 explanatory variables: DM-whole (a 7 dimensional feature) and those four bar statistics. By incorporating detailed, temporally segmented information into the model, we expect to improve the performance of whole-level difficulty estimation.

IV. Experimental Evaluation

A. Data Preparation

For the experimental evaluation, we used two datasets of piano pieces labeled with difficulty levels: Mikrokosmos-difficulty (MKD) [3] and Can I play it? (CIPI) [4]. Both datasets consist of classical piano solo scores converted to MusicXML format, with each piece annotated with one of nine difficulty levels as defined by the publisher Henle Verlag. In this experiment, we excluded pieces for which the MusicXML data and difficulty labels were not in a one-to-one correspondence, as well as pieces that produced errors during DM extraction due to file corruption. As a result, we used a total of 680 pieces: 146 from MKD and 534 from CIPI.

To ensure that all nine difficulty levels were included in both the training and evaluation sets, we divided the dataset into 544

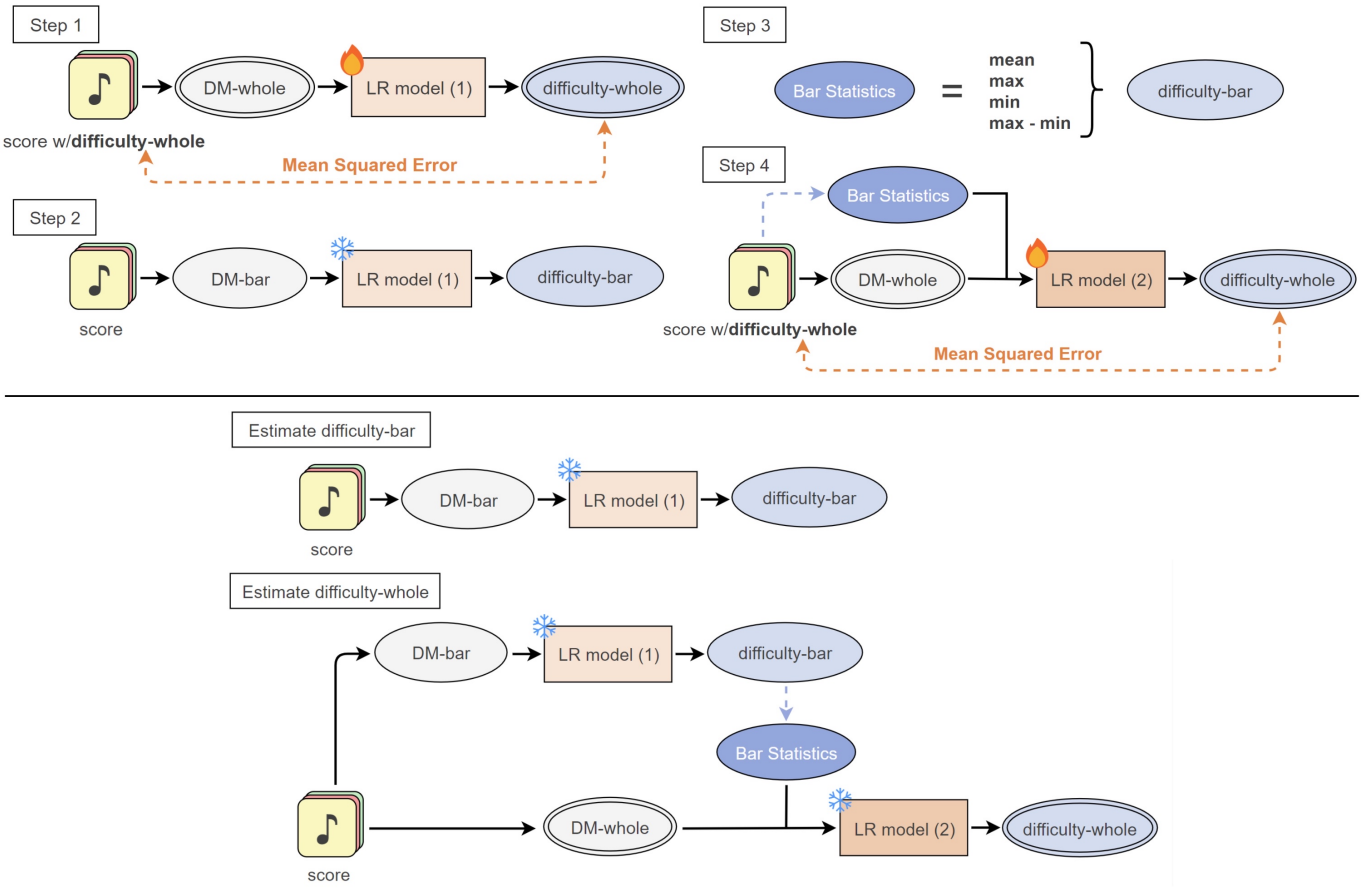

Fig. 1. Overview of the linear regression-based music difficulty estimation method. The upper figure shows the training flow and the lower figure shows the estimation flow.

pieces for training and 136 pieces for evaluation. Furthermore, DM-bar and DM-whole were extracted from each score, and each element of DM was standardized to have a mean of 0 and a variance of 1.

B. Experimental Conditions

To investigate the effectiveness of the proposed difficulty measure, we estimated musical difficulty using the method described in Section III-B and evaluated the results. Since ground-truth labels for bar-level difficulty are not available, we qualitatively evaluated the estimated bar-level difficulty by comparing it with the corresponding musical scores. For the evaluation of whole-level difficulty estimation, we used mean squared error (MSE) and the coefficient of determination (R2) as quantitative metrics. In addition, although the estimated difficulty values are output as real numbers, we also calculated accuracy (Acc) by rounding the estimates to the nearest integer within the range 1 to 9, as well as accuracy with a tolerance of ± 1 level (Acc$\pm$1), and also used these as evaluation metrics. As explanatory variables for the linear regression, we used **DM-whole**, **DM-whole with bar statistics (DM-whole w/Bar Statistics)**. We also used **DM-whole with bar count (DM-whole w/Bar Count)** as a conventional method comparable to the approaches using information extractable from the whole score [5]. The entire process from the model training to evaluation was repeated 100 times with different combinations of data splits. We then examined the mean and the 95% confidence interval for each evaluation metric and compared the results.

C. Experimental Results

Fig. 2 shows an example of bar-level difficulty estimation. In the figure, the estimated difficulties are rounded to integer values from 1 to 9, with higher values indicating greater difficulty. Bars 40 to 42, which contain a large number of notes and wide pitch intervals, making them relatively difficult to perform, are assigned a high estimated difficulty of 7. In contrast, bars 43 and 44, which have fewer notes and do not require large hand stretches, making them relatively easy to perform, are assigned lower estimated difficulties of 4 and 5, respectively. These results confirm that the bar-level difficulties by DM-bar tend to be higher for more challenging bars and lower for easier bars.

Table II shows the evaluation results for whole-level difficulty estimation. First, we focus on the values of each evaluation metric. In all cases, the average Acc is approximately 33–35%, and there is no significant improvement even when bar statistics or bar count are added. This indicates that accurately estimating difficulty on a fine nine-level scale is challenging.

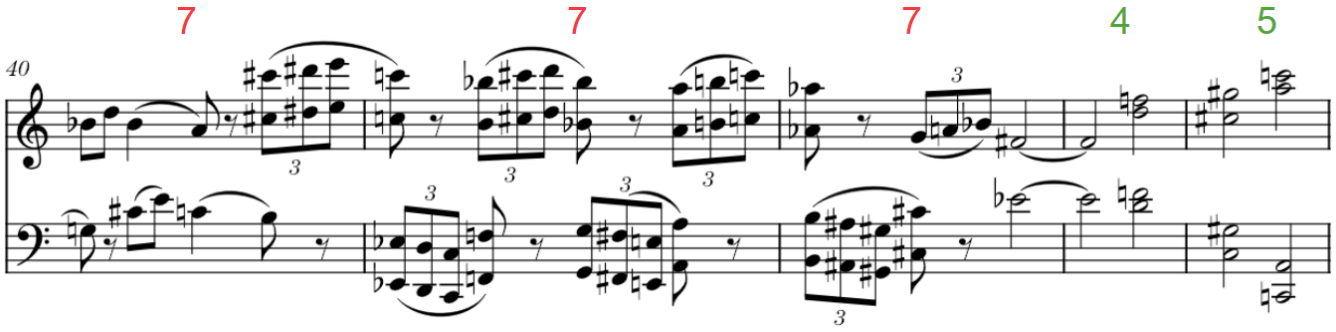

Fig. 2. Example of bar-level difficulty estimation. Measures 40–44 of *Mikrokosmos, Sz. 107* (from MKD) are shown, with estimated difficulties (1–9) displayed above the score.

TABLE II
MEANS AND 95% CONFIDENCE INTERVALS OF EVALUATION METRICS FOR WHOLE-LEVEL DIFFICULTY ESTIMATION.

Explanatory variable	MSE↓	R2↑	Acc↑	Acc±1↑
DM-whole	1.588 ± 0.041	0.597 ± 0.010	0.331 ± 0.007	0.789 ± 0.007
DM-whole w/Bar Statistics	1.439 ± 0.051	0.635 ± 0.013	0.337 ± 0.008	0.829 ± 0.006
DM-whole w/Bar Count	1.251 ± 0.033	0.683 ± 0.008	0.358 ± 0.008	0.841 ± 0.006

On the other hand, the MSE ranges from about 1.2 to 1.6, and Acc±1 is around 80%. In other words, most predicted values fall within an error margin of ±1, suggesting that the use of DM allows for a rough but reliable estimation of whole difficulty trends. Furthermore, the R2 values range from 0.6 to 0.7, confirming that the model trained with DM explains musical difficulty well. Next, we compare the evaluation results for each set of explanatory variables. The regression results obtained by adding bar statistics to DM-whole are significantly better than those using only DM-whole. On the other hand, they are still inferior to those obtained by adding bar count, implying that the ground-truth difficulty levels are strongly affected by the bar count. These findings indicate that the statistics of the estimated bar-level difficulties contribute to whole-level difficulty estimation, and that bar-level difficulty is an important feature characterizing the whole-level difficulty of a score. In addition, these results suggest that DM is effective as a measure for representing bar-level difficulty.

V. CONCLUSIONS

In this paper, we proposed a seven-dimensional difficulty measure (DM) that imposed minimal restrictions on score data format and enables local difficulty assessment, aiming for highly controllable automatic piano rearrangement. We also used DM in two forms: a bar-level difficulty measure (DM-bar) and a whole-level difficulty measure (DM-whole), and introduced a linear regression-based estimation method that captures difficulty as a continuous value. By inputting DM-bar into a linear regression model that estimates whole-level difficulty using DM-whole as explanatory variables, we enable bar-level difficulty estimation. Furthermore, by adding bar statistics—computed from the estimated bar-level difficulties—as additional explanatory variables and training a new linear regression model, we make it possible to estimate whole-level difficulty while taking bar-level difficulty into account. The experimental evaluation demonstrates that the statistics of the estimated bar-level difficulties contribute to whole-level difficulty estimation, and DM is effective as a measure for representing bar-level difficulty.

By using the proposed method for automatic difficulty labeling, it becomes easier to construct large-scale datasets, which is expected to improve the accuracy of difficulty estimation and enable applications such as automatic rearrangement systems. In future study, although this study indirectly demonstrates the usefulness of bar-level difficulty estimation for whole-level difficulty estimation, it is necessary to directly evaluate the performance of bar-level difficulty estimation. Additionally, we will apply the proposed method to an automatic piano rearrangement system and to address individual differences in difficulty perception.

ACKNOWLEDGMENT

This work was partly supported by JST, AIP Acceleration Research, JPMJCR25U5.

REFERENCES

[1] M. Suzuki, “Piano score rearrangement into multiple difficulty levels via notation-to-notation approach,” *EURASIP J. Audio Speech Music Process.*, vol. 2023, no. 1, 12 pages, Dec. 2023.

[2] Y. Ghatas, M. Fayek, and M. Hadhoud, “A hybrid deep learning approach for musical difficulty estimation of piano symbolic music,” *Alexandria Engineering Journal*, vol. 61, no. 12, pp. 10 183–10 196, 2022.

[3] P. Ramoneda, N. C. Tamer, V. Eremenko, X. Serra, and M. Miron, “Score difficulty analysis for piano performance education based on fingering,” in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 201–205.

[4] P. Ramoneda, D. Jeong, V. Eremenko, N. C. Tamer, M. Miron, and X. Serra, "Combining piano performance dimensions for score difficulty classification," *Expert Systems with Applications (ESWA)*, vol. 238, 15 pages, 2024.

[5] O. S. Véronique Sébastien Henri Ralambondrainy and N. Conruyt, "Score analyzer: Automatically determining scores difficulty level for instrumental e-learning," *Proceedings of the 13th International Society for Music Information Retrieval Conference*, pp. 571–576, 2012.

[6] S.-C. Chiu and M.-S. Chen, "A study on difficulty level recognition of piano sheet music," in *2012 IEEE International Symposium on Multimedia*, 2012, pp. 17–23.

[7] E. Nakamura and K. Yoshii, "Statistical piano reduction controlling performance difficulty," *APSIPA Transactions on Signal and Information Processing*, vol. 7, pp. 1–12, 2018.

[8] M. Terao, Y. Hiramatsu, R. Ishizuka, Y. Wu, and K. Yoshii, "Difficulty-aware neural band-to-piano score arrangement based on note- and statistic-level criteria," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 196–200.

[9] T. Fukuda, Y. Ikemiya, K. Itoyama, and K. Yoshii, "A score-informed piano tutoring system with mistake detection and score simplification," in *Proceedings of the 12th International Conference in Sound and Music Computing*, 2015, pp. 105–110.

[10] Z. Wang, K. Chen, J. Jiang, *et al.*, "Pop909: A pop-song dataset for music arrangement generation," *In Proc. Int. Society for Music Information Retrieval Conf.*, pp. 38–45, 2020.

[11] D. Edwards, S. Dixon, and E. Benetos, "Pijama: Piano jazz with automatic midi annotations," *Transactions of the International Society for Music Information Retrieval*, pp. 89–102, 2023.