

A TRANSCRIPTION-BASED SYSTEM FOR MIREX 2026 MUSIC PERFORMANCE DIFFICULTY PREDICTION

Qinyuan Liu

Zheyu Jiang
KaraOrchee

Xiaoyu Guan

qinyuan.liu@gmail.com, crystal.j.jiang@gmail.com, jjacksonsguan@gmail.com

ABSTRACT

We present our system for the Music Performance Difficulty Prediction task of MIREX 2026. The system first transcribes each solo piano recording with a piano transcription model that we trained. Three scorers then estimate the difficulty from its output: gradient-boosted trees over piece-level statistics, a Transformer over the notes, and a Transformer over the model’s audio embeddings. A recording’s score is the mean of the three. The scorers are trained on PSyllabus recordings and on our own renderings of the dataset’s MIDI files. On the five published PSyllabus splits, the system reaches a Kendall’s τ_c of 0.820 (0.804 with its scores mapped to grades), against 0.775 for the multimodal baseline released by the dataset’s authors. Our system also outperforms this baseline on the five other metrics we report, in each split. On composers withheld from development, its τ_c on recordings and renderings combined, 0.789, is within the range seen in development. On one RTX 5090 GPU, the whole system processes about 100 seconds of audio per second.

1. INTRODUCTION

Most work on piano difficulty estimation [1–3] uses symbolic scores. Ramoneda et al. [4] extended the problem to audio with PSyllabus, 7,901 recorded pieces graded on 11 levels by the Piano Syllabus community, plus grades from 13 other grading systems. Other work probes pretrained audio encoders for difficulty [5, 6]. The PSyllabus models take a constant-Q transform (CQT), a piano roll from the transcription model of Kong et al. [7], or both.

However, a piano roll keeps only the notes that a transcription model detects and leaves out the model’s internal representation of the audio, a sequence of embeddings. Our system for the MIREX 2026 task [8] uses both: it (1) transcribes each recording with a transcription model we trained, (2) passes the notes and embeddings to three scorers with different inputs, and (3) averages their scores.

2. SYSTEM

Transcription. We trained the transcription model on the training split of MAESTRO v3.0.0 [9] (962 performances). For each recording, it outputs notes (onset, offset, pitch and velocity) and, in the same pass, the sequence of audio embeddings. Audio is first converted to mono, resampled and normalized in loudness.

Scorers. Gradient-boosted trees take piece-level statistics of the notes and a piece-level summary of the embeddings. A Transformer takes the sequence of notes, together with piece-level statistics and the same summary. A second Transformer of the same design takes the sequence of embeddings instead of the notes, with the piece-level statistics only. Each scorer’s score is the mean of several independently trained instances, and the system averages the three.

Training data. Only PSyllabus [4] is used to train the scorers.¹ From the videos it lists, we obtained audio with detectable notes for 7,254 of its 7,901 entries (92%); the rest could not be retrieved or had no detectable notes. Each recording starts at the entry’s listed time. An entry inside a longer video is cut to about the length of its MIDI transcription in the dataset. Because the task also evaluates a synthesized rendering of each hidden piece’s score and aggregates the predictions per piece [8], we also train the scorers on our renderings of the transcribed MIDI files that PSyllabus distributes for the same pieces. The submitted models are trained on all 7,254 entries, the withheld composers included. The Hidden Voices benchmark [4] (Section 3) served only as a check, never for training or model selection. MAESTRO and PSyllabus are public, and the task draws its hidden repertoire from outside public difficulty datasets [8].

Label-space handling and post-processing. Targets are the PSyllabus grades (0–10). For the trees only, a grade-10 piece that its source grading system places above level 10 gets a target above 10. Because the task ranks systems by Kendall’s τ_c alone [8], which depends only on the order of the scores, the system outputs a continuous score and needs no mapping to the task’s undisclosed scale. For the categorical metrics in Section 3, we map scores to grades with thresholds fitted on cross-validated predictions for the training data. The system outputs 0 when the input is silent, when it detects fewer than five notes, or when the pipeline fails.



© Q. Liu, Z. Jiang, and X. Guan. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0).
Attribution: Q. Liu, Z. Jiang, and X. Guan, “A Transcription-Based System for MIREX 2026 Music Performance Difficulty Prediction”, *MIREX 2026 technical report*.

¹ <https://doi.org/10.5281/zenodo.14794592>

Metric	MM [4]	PR+era [4]	Ours
τ_c	0.775 ± 0.004	0.774 ± 0.008	0.820 ± 0.004
τ_c (rec.+rend.)	—	—	0.819 ± 0.004
τ_b	0.782 ± 0.004	0.779 ± 0.008	0.811 ± 0.005
Macro-MSE ↓	1.766 ± 0.090	1.780 ± 0.094	1.427 ± 0.085
Balanced acc.	0.377 ± 0.016	0.381 ± 0.010	0.401 ± 0.009
Acc±1	0.787 ± 0.006	0.789 ± 0.013	0.826 ± 0.005
Spearman’s ρ	0.895 ± 0.004	0.894 ± 0.006	0.919 ± 0.004

Table 1. Published PSyllabus splits, mean $\pm$ standard deviation over the five splits. MM (early fusion of CQT and piano roll) and PR+era (piano roll, era as an auxiliary task) are the released models with the best mean on at least one metric, scored from their released predictions on the same pieces. Rec.+rend.: recording and rendering scores averaged. —: no released predictions for renderings.

3. EVALUATION

Before developing this system, after only a first feature baseline, we withheld 15% of the composers (184). We made all design choices by five-fold, composer-disjoint cross-validation on the other composers (the development folds), and used the withheld composers twice: to evaluate a first version, and as a pre-set check that one late change to its training data had to pass. We evaluated the final system once on the five published splits of [4]; earlier versions were also scored there, but we chose between them on the development folds. To approximate the task’s per-piece aggregation, we also report τ_c (rec.+rend.), the τ_c of the mean of each piece’s scores for its recording and our rendering. Except for Hidden Voices, the evaluations use one instance of each Transformer.

Published splits. These are five random 60/20/20 splits of PSyllabus, stratified by grade, with 1,589 test pieces each; we score the 1,441–1,466 per split for which we have audio. We train our scorers on each split’s training and validation pieces; the released models were trained on the training pieces, with the validation pieces used for early stopping. Our τ_c and ρ are computed on the continuous scores. The other metrics use grades from the thresholds: macro-MSE (squared error averaged over the true grades), balanced accuracy (recall averaged over grades) and Acc±1 (the share of pieces within one grade). We compute τ_b , macro-MSE and balanced accuracy as in the code released with [4], which reports them as τ_c , MSE and Acc; re-scored this way on the full test sets, the released predictions reproduce the means it reports. Acc±1 and ρ follow the task page [8].

Table 1 compares our system with MM and PR+era. Our system reaches a τ_c of 0.820, against 0.775 for MM, and outperforms the best released model on every metric in every split. The released predictions are integer grades, and τ_c gives no credit for pairs a system scores as equal. Mapped to grades, our system still reaches 0.804, higher in every split. Unlike the hidden test set [8], these splits are not composer-disjoint, and about 80% of their test pieces were in our development folds, so these results are not independent of our design choices.

Withheld composers. On the pieces of the 165 withheld composers with audio (1,393), our τ_c (rec.+rend.) is 0.789, within the range of the five development folds (0.761–0.818, mean 0.786). This suggests that our design choices did not overfit the development folds.

Grading systems. Because the hidden scale is not disclosed, we also compare our scores with the grades that 13 other grading systems assign to PSyllabus pieces. On the development folds, τ_c (rec.+rend.) ranges from 0.60 (ANZCA, 278 pieces) to 0.87 (ABRSM, 956 pieces).

Robustness. On development pieces rendered with either of two soundfonts not used in training (1,472 and 740 pieces, respectively), τ_c (rec.+rend.) is 0.001 and 0.003 lower, so the choice of soundfont matters little. Hidden Voices [4, 10] provides a second check. On the 34 pieces with both our audio and released predictions, our τ_c is 0.759, within the range the released models reach when each model’s five split versions are averaged (0.702–0.769; 0.673–0.727 for a single version on average).

Limitations. Our renderings come from MIDI files transcribed from performances [4] rather than from scores, so they only approximate the organizers’ renderings. Our system also differs from the released models in transcription, scorers and training data at once, so the comparison cannot show which of them makes the difference.

Compute declaration. Peak GPU memory footprint is 2.3 GB for pieces under 5 minutes and 2.7 GB for an 87-minute recital. After a one-time model load of about 7 s, inference on one RTX 5090 GPU, end to end including transcription, processes about 100 s of audio per second: the average inference time was 0.5 s per recording for 13 recordings of 52 s on average (about 1 s for a test piece of two such recordings), and 45 s for the recital.

4. AI USAGE STATEMENT

We used LLM-based assistants to help draft the text, including the related work. The authors directed the research, its implementation and the revision of the text, verified all numbers and references, and take full responsibility for the paper.

5. REFERENCES

- [1] V. Sébastien, H. Ralambondrainy, O. Sébastien, and N. Conrui, “Score Analyzer: Automatically determining scores difficulty level for instrumental E-Learning,” in *Proc. of the Int. Society for Music Information Retrieval Conf. (ISMIR)*, 2012, pp. 571–576.
- [2] P. Ramoneda, N. C. Tamer, V. Eremenko, X. Serra, and M. Miron, “Score difficulty analysis for piano performance education based on fingering,” in *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 201–205, arXiv:2203.13010.
- [3] 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*, vol. 238, p. 121776, 2024, arXiv:2306.08480.

- [4] P. Ramoneda, M. Lee, D. Jeong, J. J. Valero-Mas, and X. Serra, “Can audio reveal music performance difficulty? Insights from the Piano Syllabus dataset,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 1129–1141, 2025, arXiv:2403.03947.
- [5] H. Zhang, J. Liang, and S. Dixon, “From audio encoders to piano judges: Benchmarking performance understanding for solo piano,” in *Proc. of the Int. Society for Music Information Retrieval Conf. (ISMIR)*, 2024, arXiv:2407.04518.
- [6] P. Alonso-Jiménez, P. Ramoneda, R. O. Araz, A. Poltronieri, and D. Bogdanov, “OMAR-RQ: Open music audio representation model trained with multi-feature masked token prediction,” 2025, arXiv:2507.03482.
- [7] Q. Kong, B. Li, X. Song, Y. Wan, and Y. Wang, “High-resolution piano transcription with pedals by regressing onset and offset times,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3707–3717, 2021, arXiv:2010.01815.
- [8] P. Ramoneda and H. Zhang, “MIREX 2026: Music Performance Difficulty Prediction,” Task description, 2026, https://www.music-ir.org/mirex/wiki/2026:Music_Performance_Difficulty_Prediction, accessed 24 Sep. 2026.
- [9] C. Hawthorne, A. Stasyuk, A. Roberts, I. Simon, C.-Z. A. Huang, S. Dieleman, E. Elsen, J. Engel, and D. Eck, “Enabling factorized piano music modeling and generation with the MAESTRO dataset,” in *Proc. of the Int. Conf. on Learning Representations (ICLR)*, 2019, arXiv:1810.12247.
- [10] University of Colorado Boulder, “Hidden Voices: Piano music by Black women composers,” <https://www.colorado.edu/project/hidden-voices/>, accessed 24 Sep. 2026.