EVALUATING HIGH-RESOLUTION PIANO SUSTAIN PEDAL DEPTH ESTIMATION WITH MUSICALLY INFORMED METRICS

Hanwen Zhang[1,2,*], *Kun Fang*[1,2,*], *Ziyu Wang*[3,4], *Ichiro Fujinaga*[1,2]

[1] Schulich School of Music, McGill University, Montréal, QC, Canada
[2] CIRMMT (Centre for Interdisciplinary Research in Music Media and Technology), Montréal, QC, Canada
[3] Courant Institute of Mathematical Sciences, New York University, New York, USA
[4] Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, UAE

ABSTRACT

Evaluation for continuous piano pedal depth estimation tasks remains incomplete when relying only on conventional frame-level metrics, which overlook musically important features such as direction-change boundaries and pedal curve contours. To provide more interpretable and musically meaningful insights, we propose an evaluation framework that augments standard frame-level metrics with an *action*-level assessment measuring direction and timing using segments of press/hold/release states and a *gesture*-level analysis that evaluates contour similarity of each press-release cycle. We apply this framework to compare an audio-only baseline with two variants: one incorporating symbolic information from MIDI, and another trained in a binary-valued setting, all within a unified architecture. Results show that the MIDI-informed model significantly outperforms the others at action and gesture levels, despite modest frame-level gains. These findings demonstrate that our framework captures musically relevant improvements indiscernible by traditional metrics, offering a more practical and effective approach to evaluating pedal depth estimation models.

***Index Terms*—** Music information retrieval, Piano performance analysis, Sustain pedal depth estimation, Evaluation metrics, Musically informed evaluation

1. INTRODUCTION

Although the sustain pedal is a continuous control crucial for producing distinct sound effects essential to classical piano playing, previous studies treated it as an auxiliary signal within piano transcription systems [1, 2, 3, 4] and mostly focused on binary on/off detection [5, 6, 7, 8, 9, 10]. Previous work [11] demonstrates the importance and benefits of continuous pedal depth estimation. However, compared with mature MIR tasks such as melody extraction, beat tracking, and chord recognition, pedal depth estimation remains underexplored, and its evaluation still largely relies on standard generic frame-level metrics (MSE, MAE, F1, etc.), which are not specifically designed for musically meaningful information, such as whether action boundaries are correct, and whether phrase-level contours are reasonable. Binary F1-score can ignore important cases such as half-pedal and pedal fluttering, while multi-class F1-score overlooks expressive musical aspects in the curve. Frame-level regression metrics (MSE/MAE) better reflect continuous values, but still over-penalize small timing shifts or minor jitters, even when two pedal curves are nearly equivalent in perceived effect.

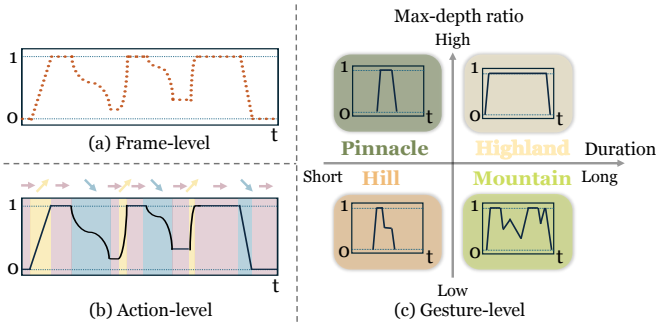

Fig. 1. An illustration of the three-level analysis proposed and used in this work. (a) Frame: each time step is treated as the basic unit. (b) Action: consecutive frames sharing the same direction/state are merged into segments labeled *press* (yellow), *hold* (pink), or *release* (blue). (c) Gesture: complete pedal press-release cycles, characterized and classified according to canonical contour shapes.

In this paper, we propose evaluating pedal depth estimation models not only with existing frame-level metrics but also using two additional levels of metrics that are musically more interpretable and practically more informative: (1) an *action-level* evaluation, which assesses predictions based on segments of press/hold/release; and (2) a *gesture-level* evaluation, which treats each press–release cycle as a unit and compares contour similarity within that unit. Evaluation at action and gesture levels extends the generic frame-based analysis by providing richer insights into model performance that align with actual practice and pedagogical concepts in piano playing.

For our experiments, we design three models within a unified Transformer-based architecture to demonstrate how the three levels yield complementary, interpretable judgments. The goal is to explore whether our new framework can reveal differences in model performance missed in frame-level metrics, and whether evaluation at the action or gesture level will help us better understand model behavior in musical contexts. Results show that continuous-valued models outperform binary-valued at predicting all *press*/*release* actions and plotting gesture shapes; adding MIDI further strengthens

*Equal contribution.

action boundary clarity and gesture contour matching, though fine micro-dynamics in long *mountain*-like gestures remain challenging for our models. Overall, this framework provides more musically meaningful analysis and thus more informative model performance diagnosis for piano sustain pedal depth estimation tasks.

2. RELATED WORK

Early sustain pedal depth estimation research projects evolved from sensor-based, small-scale classification [5, 6, 8, 7] to binary pedal detection within transcription systems [9, 10, 2, 3, 4], but evaluation still relies on frame-level metrics. [11] demonstrated feasibility but further underscored the absence of a unified, reproducible evaluation paradigm, prompting us to reconsider both what to evaluate and how, via a musically informed approach.

More broadly, temporal music information retrieval (MIR) research has developed a family of task-aligned evaluation protocols across continuous and discrete, low- and high-level targets[12]. Continuous signals (e.g., melody) are typically scored by frame-wise accuracy[13, 14]. Discrete events (e.g., onsets, beats) use tolerance-window precision/recall/F1 and continuity measures[15, 16]. Chord recognition, structural segmentation, and pattern discovery are segment/structural tasks; accordingly, evaluation typically emphasizes duration-weighted overlap, vocabulary mappings, and pairwise consistency, and specifically for segmentation adopts a hierarchical scheme that separates boundary localization from structural labeling[17, 18, 19, 20, 21]. Following this trajectory, we position pedal depth estimation as a composite temporal task that simultaneously involves a continuous curve (depth), discrete events (press/release boundaries), and structural semantics (contours). It therefore requires measurements at the corresponding levels and a unified aggregation, rather than a single-scale, generic metric.

3. PEDAL PREDICTION BASELINES

Before introducing our proposed pedal evaluation methods, we first describe the models used for comparison and present conventional metrics. Section 3.1 defines the task and introduces controlled model variants based on Transformer; Section 3.2 reviews standard frame-level metrics; Section 3.3 discusses their limitations, thus indicating the need for our music-informed evaluations in Section 4.

3.1. Pedal Prediction Model Variants

We formulate sustain pedal prediction as a temporal sequence estimation task. We consider two input representations: *audio* denotes log-mel spectrograms (229 bins) and MFCCs (20 dims) computed on $\sim$5 s windows (500 frames), while *MIDI* denotes frame-aligned 88-dim note-velocity vectors obtained from audio via transcription using [4]. The model outputs a frame-wise continuous pedal depth sequence $x_{1:T} \in [0, 1]$ (MIDI CC64 normalized by 127), frame-wise pedal onset $o_{1:T}$ and offset $f_{1:T}$ binary event sequences, and a segment-level global depth $g \in [0, 1]$ that equals the segment pedal level average using the following loss:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{pedal}} + \lambda_2 \mathcal{L}_{\text{global}} + \lambda_3 \mathcal{L}_{\text{onset}} + \lambda_4 \mathcal{L}_{\text{offset}}, \quad (1)$$

where $\mathcal{L}_{\text{pedal}}$ and $\mathcal{L}_{\text{global}}$ are MSE losses on frame-wise and segment-level depth, and $\mathcal{L}_{\text{onset}}$, $\mathcal{L}_{\text{offset}}$ are BCE losses on event sequences; the weights $\lambda_{1..4}$ are fixed across variants.

Building on [11], we keep multi-task heads and streamline the input stage: mel features are encoded by a small CNN, MFCCs by an MLP, the fused representation is passed to a Transformer encoder (8 heads, standard FFN), and the output heads produce the targets above. To isolate the roles of output format and input modality, we train three controlled variants under identical architecture, hyperparameters, and optimization:

1. BINARY: binarized labels trained with BCE; output in $[0, 1]$.
2. AUDIO: continuous regression on mel bins + MFCCs.
3. AUDIO+MIDI: same as AUDIO, with an additional fused, frame-aligned MIDI (pitch and velocity) stream.

3.2. Frame-Level Evaluation

Standardized evaluation protocols for continuous pedal estimation are still under discussion; most evaluations still borrow criteria from binary prediction [11]. Conventional frame-wise metrics measures: (i) classification scores after discretization (binary and 4-class) using Precision/Recall/F1, and (ii) regression errors on the continuous depth $x_{1:T}$ using MAE and MSE. These summarize per-frame signal fidelity and enable comparison to prior work.

3.3. Challenges in Existing Evaluation

We highlight the following issues and limitations of frame-based metrics for pedaling: (i) Not all frames matter equally: boundary frames at harmonic changes or intended releases are musically critical, whereas interior frames within a sustained harmony can tolerate substantial flutter. Uniform frame weighting over-penalizes acceptable fluctuations and under-emphasizes boundary accuracy. (ii) Contour insensitivity: frame-wise metrics measure absolute differences but ignore the general contour: phase-shifted yet musically equivalent curves and equal long holds with different micro-oscillations can receive large errors despite similar intent. (iii) Low diagnostic value: aggregate frame scores do not indicate what to improve (timing, duration, contour) or where failures occur (which gesture types), offering limited guidance for model design and error analysis.

These observations motivate the music-informed analysis introduced next (Sections 4.1 and 4.2), which emphasizes boundary correctness, segment coverage, and contour similarity. Figure 2 showcases two excerpts where the MSE and MAE point at relatively poor predictions, even though the general pattern is captured and aligned properly in time.

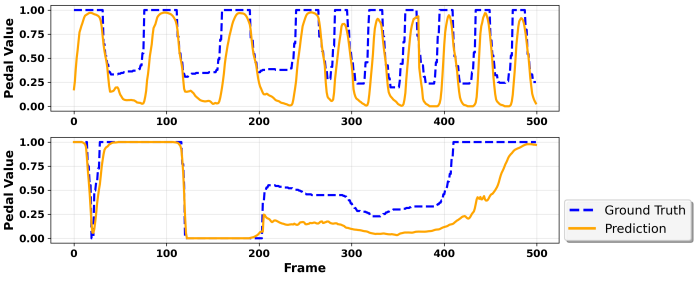

Fig. 2. Two examples with large MSE and MAE but correctly following the pattern. The top one has MSE 0.0983 and MAE 0.2425. The bottom one has MSE 0.0866 and MAE 0.2053.

4. EVALUATION FRAMEWORK

Beyond frame-level metrics, we propose two novel perspectives for evaluation inspired by piano pedagogy: (i) an action-based eval-

uation (Section 4.1), in which we segment the pedal curve into *press*/*hold*/*release* and score temporal coverage and alignment; and (ii) a gesture-based evaluation (Section 4.2), where we consider complete pedal gestures from pedal on to off and compare characteristic contour shapes that implies pedaling in expressive piano performance.

4.1. Action-Based Evaluation

Instead of working with quantized or raw depth values, we now treat pedaling as a sequence of discrete directional actions (*press*, *hold*, and *release*) derived from the sustain pedal signal $x_{1:T} \in [0, 1]$. This representation reflects performers' intentional control of the pedal at different musical moments. As noted in lessons with Claudio Arrau [22], pianists and teachers pay close attention to pedaling and predominantly describe it in terms of actions. Compared to exact depth values, the exact timing and type of action better correspond to how pianists conceptualize the use of sustain pedal.

However, ground truth signals from optical sensors and model outputs inevitably contain local jitter. To preserve original temporal resolution while avoiding purely frame-level analysis, we run linear regression on a sliding window centered on each frame. We use a window size of 19, a slope threshold of 0.005, and a minimum R^2 of 0.5 to identify action states for both ground truth and model predictions. Therefore, for each frame at time t, its action state is determined by the regression slope of the segment centered at t.

The action-based evaluation can now be framed as a three-class classification task for *press*, *hold*, *release* classes using action states extracted from by the steps mentioned above, and the evaluation follows the standard precision, recall, and F1 framework. We report scores for each action class as well as class-balanced macro and weighted averages, all shown in Table 2.

Model	Binary (P↑/R↑/F1↑)			4-Class (P↑/R↑/F1↑)			MSE ↓	MAE ↓
[11]	0.8975	0.8971	0.8973	0.6849	0.6971	0.6863	0.0425	0.1339
BINARY	0.8929	0.8909	0.8914	0.5622	0.6615	0.5655	0.0880	0.1714
AUDIO	0.9043	0.9037	0.9039	0.7013	0.7153	0.7045	0.0416	0.1237
AUDIO+MIDI	0.9379	0.9370	0.9372	0.7529	0.7661	0.7546	0.0280	0.0986

Table 1. Frame-level results.

4.2. Gesture-Based Evaluation

Prior work [11] shows sustain pedaling is not a binary down–up switch but nuanced press–to–release cycles whose shapes depend on musical intent. Therefore, we propose *gestures* as an important perspective to evaluate pedal depth estimation. We therefore evaluate from the perspective of *gestures*, defined as contiguous segments that begin when depth first exceeds a small threshold ϵ (press onset) and end at the subsequent return below ϵ (release). Formally, let $x_{1:T} \in [0, 1]$ be the frame-wise pedal depth at a fixed frame rate. A gesture $g = x_{t:t+n}$ is a maximal interval satisfying:

$$x_{t-1} \leq \epsilon, x_{t+n+1} \leq \epsilon, \text{and } x_j > \epsilon \, \forall j \in [t, t + n].$$

That is, g starts at the first frame exceeding ϵ, remains above ϵ, and ends at the first return below ϵ. To classify various types of gestures, we further compute a *max-depth ratio*, which measures how long the pedal stays or fluctuates near its maximum depth within the segment:

$$r(g) = \frac{\left|\{x_j \geq \theta \max(g) | x_j \in g\}\right|}{n}.$$

For a given piece, the pedal value sequence can thus be segmented into gestures, which are then classified along two axes, *duration* and *max-depth ratio* $r(g)$, into four canonical shapes (illustrated in Figure 1).

- ***Pinnacle***: (short, high r): a quick press–release with a dominant peak; used for brief accents or as catch pedal.
- ***Hill***: (short, low r): a short press–release with gentle shaping (e.g., partial releases/half-pedal control); More decorative and nuanced than Pinnacle.
- ***Highland***: (long, high r): press–hold–release with an extended plateau, building resonance and/or sound volume (common in large Romantic piano repertoires).
- ***Mountain***: (long, low r): sustained depth with modulation/oscillation before release, enabling extended blending and color (typical in Impressionist compositions).

The remaining intervals (i.e., those with depth $< \epsilon$ in between the four gestures listed above) are labeled as *plain*, a category that does not correspond to any gesture type.

Classifying pedal gestures and evaluating them separately provides a clearer and more musically meaningful interpretation. Pedaling, as Banowetz describes, is a "downward journey" [23], and contiguous pedal cycles, which we define as *gestures*, appear to be the most appropriate unit for evaluating pedal depth estimation in a musical context.

For all non-*plain* gestures in the corpus: duration (frames) has mean 147.8, median 70.0, std 402.9; max-depth ratio has mean 0.645, median 0.688, std 0.220. We set our max-depth ratio threshold to $\theta = 0.65$ and duration threshold to 100 frames to define our quadrants.

We already see that frame-wise scores can miss contour similarity in Section 3.3. To compare gesture shapes while tolerating small local deviations, we adopt (i) Fourier-descriptor analysis: compute discrete Fourier coefficients and reconstruct with the first $K{=}11$ coefficients to suppress high-frequency noise/model artifacts, and compare the reconstructed low-pass signals via MSE; (ii) 5-point analysis: compute duration-weighted MSE using only five landmarks per gesture (start, end, median, mean, max). According to these new metrics, the top example in Figure 2 yields an MSE of 0.0586 using the Fourier method, whereas the bottom example obtains an MSE of 0.0274 in the 5-point analysis, now both being less harshly graded than by raw MSE scores.

5. EXPERIMENTS

This section presents our experiments with three models detailed in 3.1 and their performance evaluation using our proposed new framework: Section 5.1 details the dataset and unified training setup; Section 5.2 shows results and findings across models; Section 5.3 summarizes consistent strengths and remaining limitations.

5.1. Dataset and Training

We use MAESTRO v3.0.0 [24], which, among other piano corpora [25, 26], is well suited for sustain pedal depth estimation for its professional recordings, expert performances, a curated repertoire, and synchronized optical pedal depth.

All models are trained on a single NVIDIA H100 (80 GB) with batch size 32, using AdamW ($\beta_1{=}0.9$, $\beta_2{=}0.999$, weight decay 0.01) and a OneCycle scheduler (peak LR $5{\times}10^{-4}$; 10% warm-up; init factor $1/25$; final factor $1/100$; cosine). We train for 15 epochs

	Press			*Hold*			*Release*			*Overall*	
Model	P ↑	R ↑	F1 ↑	P ↑	R ↑	F1 ↑	P ↑	R ↑	F1 ↑	F1 (Macro) ↑	F1 (Weighted) ↑
BINARY	0.6083	0.3815	0.4689	0.8147	0.9179	0.8632	0.6427	0.4310	0.5160	0.6160	0.7648
AUDIO	0.5383	0.6959	0.6070	0.8877	0.8028	0.8431	0.5544	0.6850	0.6128	0.6876	0.7815
AUDIO+MIDI	0.6325	0.7745	0.6964	0.9171	0.8568	0.8859	0.6807	0.7721	0.7235	0.7686	0.8392

Table 2. Action-level results: per-action Precision/Recall/F1, plus macro- and weighted-F1 overall.

	Mountain		*Highland*		*Hill*		*Pinnacle*		*Plain*		Weighted	
Model	5-pts ↓	Fourier ↓	5-pts ↓	Fourier ↓	5-pts ↓	Fourier ↓	5-pts ↓	Fourier ↓	5-pts ↓	Fourier ↓	5-pts ↓	Fourier ↓
BINARY	0.1009	0.0760	0.0939	0.0286	0.1357	0.1373	0.1236	0.1110	0.1475	0.0526	0.1143	0.0634
AUDIO	0.0689	0.0284	0.0880	0.0146	0.0899	0.0521	0.0941	0.0503	0.1332	0.0471	0.0946	0.0329
AUDIO+MIDI	0.0457	0.0273	0.0441	0.0116	0.0511	0.0358	0.0518	0.0291	0.0737	0.0247	0.0530	0.0225

Table 3. Gesture-level results: MSE for each gesture category and overall, using two shape measures: (i) 5-point, and (ii) Fourier (11 coefficients).

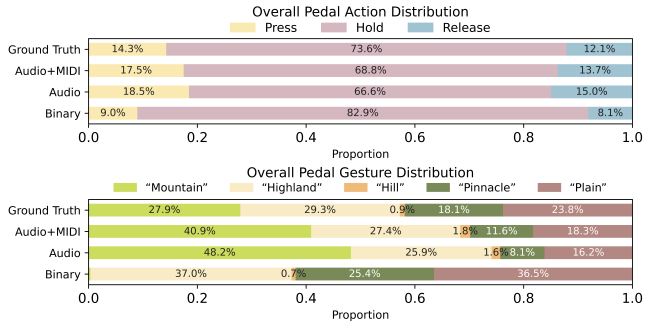

Fig. 3. Overall distribution of pedal *actions* (top) and *gestures* (bottom) across models. Each row corresponds to one model and the ground truth; bars are stacked by category, with percentages computed from summed frame counts in each category. Labels show per-category proportions.

and select the best checkpoint (∼150k steps, epoch 13). To ensure fair comparison, all hyperparameters and architecture are shared across variants, except the additional MIDI input.

5.2. Evaluation Across Models

We use conventional frame-based metrics (Table 1) as well as our action- (Table 2) and gesture-based metrics (Table 3). Overall, the AUDIO+MIDI model performs the best, followed by AUDIO, with BINARY the last. This pattern highlights that estimating continuous pedal values is beneficial, and adding MIDI input further enhances performance by providing helpful structural priors.

Continuous estimation matters (vs. binary classification). Although BINARY achieves a comparable frame-wise F1 (0.8914), it exhibits systematic bias at both action and gesture levels (Figure 3, Table 2). At action level, recall scores for *press* and *release* are very low, indicating that many actions are missed. This likely comes from data imbalance in binary setting where most frames are "off," encouraging under-prediction of pedal activation. At gesture level, BINARY largely collapses to predicting *highland* and *pinnacle*, and rarely recognizes more complex patterns such as *mountain* (Fig. 3). Even for *highland* and *pinnacle*, which BINARY predicts frequently, gesture-level MSE remains unsatisfactory (Table 3).

Models trained for continuous estimation have remarkably improved action-level recalls and modestly but consistently higher precisions than BINARY (Table 2). In addition, at the gesture level, AUDIO and AUDIO+MIDI can identify gesture categories other than *highland* and *pinnacle*, yielding distributions closer to ground truth (Figure 3), and they also result in substantially lower per-gesture MSE, a positive sign for more accurate shape modeling. These findings agree with [11] and reinforce the importance of continuous-valued prediction for piano pedal estimation.

Adding MIDI input provides structural priors. AUDIO+MIDI, on the other hand, predicts all gesture categories with a distribution the closest to ground truth (Figure 3). It also preserves shape more faithfully within each gesture. In particular, gesture-level MSE drops notably for short, rapid gestures such as *pinnacle* and *hill*, indicating greater temporal sensitivity than the AUDIO baseline (Table 3). The substantially improved *5-point* scores and the modestly better *Fourier* (11-coefficient) scores compared to AUDIO suggest that the extra MIDI input mainly helps predict the general trend more than high-frequency micro-oscillations. In addition, scores across gestures are more balanced for AUDIO+MIDI, which suggests that MIDI input helps the model generalize more consistently for both short and long gestures.

5.3. Strengths and Limitations of Current Pedal Models

We summarize consistent patterns across models: where they succeed reliably and where errors persist.

Strengths. *Highland* is consistently the best-predicted gesture: its distribution rate is the closest to ground truth (Figure 3), and among all gestures, segments identified as *highland* almost always have the lowest MSE according to Fourier and 5-point analyses (Table 3). This indicates that models are effective at capturing gestures with relatively stable, long pedal movements, where the resulting sound is more pronounced.

Limitations and biases. (i) Under-detection of high max-depth gestures. Models tend to under-predict *Pinnacle*/*Highland* (missed or confused with *Hill*/*Mountain*), suggesting difficulty with gestures having high max-depth ratios. (ii) Short-and-fast gestures remain hard. *Pinnacle*/*Hill* are better captured but still with relatively high MSE, likely because the brief duration and rapid changes exceed the temporal resolution that per-frame prediction can reliably resolve.

6. CONCLUSION

We studied sustain pedal depth estimation with controlled experiments and evaluated using our proposed new framework with three levels at *frame*, *action*, and *gesture*. Across three model variants, we demonstrated that continuous-valued estimation is crucial for accurately capturing actions and preserving gesture characteristics and

pre-extracted MIDI pitch velocity information provides useful structural priors for better contour shaping in gestures. While conventional frame-wise metrics can miss musically acceptable behavior, our proposed three-level approach provides a more comprehensive view of strengths and limitations of our models. These results shows the value of our musically informed evaluation and point to future work on boundary-sensitive objectives, contour-aware losses, and perceptual validation for piano sustain pedal depth estimation tasks.

7. REFERENCES

[1] Curtis Hawthorne, Erich Elsen, Jesse Song, Adam Roberts, Ian Simon, Colin Raffel, Jonathon Engel, Sageev Oore, and Douglas Eck, "Onsets and frames: Dual-objective piano transcription," in *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, Paris, France, 2018.

[2] Qiuqiang Kong, Bochen Li, Xuchen Song, Yuan Wan, and Yuxuan 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, October 2021.

[3] Yujia Yan, Frank Cwitkowitz, and Zhiyao Duan, "Skipping the frame-level: Event-based piano transcription with neural semi-CRFs," in *Advances in Neural Information Processing Systems*, Virtual, 2021, vol. 34, pp. 20583–20595.

[4] Yujia Yan and Zhiyao Duan, "Scoring time intervals using non-hierarchical transformer for automatic piano transcription," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, San Francisco, United States, 2024.

[5] Beici Liang, György Fazekas, and Mark B. Sandler, "Detection of piano pedaling techniques on the sustain pedal," in *Proceedings of Audio Engineering Society Convention 143*, New York, United States, 2017.

[6] Beici Liang, György Fazekas, Mark B. Sandler, and Andrew McPherson, "Piano pedaller: A measurement system for classification and visualisation of piano pedalling techniques," in *Proceedings of The International Conference on New Interfaces for Musical Expression (NIME)*, Copenhagen, Denmark, 2017, pp. 325–329.

[7] Beici Liang, György Fazekas, and Mark B. Sandler, "Measurement, recognition, and visualization of piano pedaling gestures and techniques," *Journal of the Audio Engineering Society*, vol. 66, no. 6, pp. 448–456, June 2018.

[8] B. Liang, G. Fazekas, and M. B. Sandler, "Piano legato-pedal onset detection based on a sympathetic resonance measure," in *Proceedings of the 26th European Signal Processing Conference (EUSIPCO)*, Rome, Italy, 2018.

[9] Beici Liang, György Fazekas, and Mark B. Sandler, "Piano sustain-pedal detection using convolutional neural networks," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom, 2019, pp. 241–245.

[10] Beici Liang, Gyorgy Fazekas, and Mark B. Sandler, "Transfer learning for piano sustain-pedal detection," in *2019 International Joint Conference on Neural Networks (IJCNN)*, Budapest, Hungary, 2019, pp. 1–6, IEEE.

[11] Kun Fang, Hanwen Zhang, Ziyu Wang, and Ichiro Fujinaga, "High-resolution sustain pedal depth estimation from piano audio across room acoustics," in *Proceedings of the 26th International Society for Music Information Retrieval Conference (ISMIR)*, Daejeon, Korea, September 2025.

[12] Colin Raffel, Brian McFee, Eric J. Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, and Daniel P. W. Ellis, "mir_eval: A Transparent Implementation of Common MIR Metrics," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Taipei, Taiwan, 2014.

[13] G. E. Poliner, D. P. W. Ellis, A. F. Ehmann, Emilia Gómez, S. Streich, and B. Ong, "Melody transcription from music audio: Approaches and evaluation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1247–1256, 2007.

[14] Justin Salamon, Emilia Gómez, Daniel P. W. Ellis, and Gaël Richard, "Melody extraction from polyphonic music signals: Approaches, applications, and challenges," *IEEE Signal Processing Magazine*, vol. 31, no. 2, pp. 118–134, March 2014.

[15] M. E. P. Davies, N. Degara, and M. D. Plumbley, "Evaluation methods for musical audio beat tracking algorithms," Technical Report C4DM-TR-09-06, Centre for Digital Music, Queen Mary University of London, London, England, October 2009.

[16] Sebastian Böck, Florian Krebs, and Markus Schedl, "Evaluating the online capabilities of onset detection methods," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2012, pp. 49–54.

[17] J. Pauwels and G. Peeters, "Evaluating automatically estimated chord sequences," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2013, pp. 749–753, IEEE.

[18] Y. Ni, M. McVicar, R. Santos-Rodriguez, and T. De Bie, "Understanding effects of subjectivity in measuring chord estimation accuracy," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 12, pp. 2607–2615, 2013.

[19] D. Turnbull, G. Lanckriet, E. Pampalk, and M. Goto, "A supervised approach for detecting boundaries in music using difference features and boosting," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2007, pp. 51–54.

[20] M. Levy and M. Sandler, "Structural segmentation of musical audio by constrained clustering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 318–326, 2008.

[21] H. M. Lukashevich, "Towards quantitative measures of evaluating song segmentation," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2008, pp. 375–380.

[22] Victoria A. von Arx, *Piano Lessons with Claudio Arrau: A Guide to His Philosophy and Techniques*, Oxford University Press, May 2014.

[23] Joseph Banowetz, *The Pianist's Guide to Pedaling*, Indiana University Press, Bloomington, IN, 1985.

[24] Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck, "Enabling factorized piano music modeling and generation with the MAESTRO dataset," in *Proceedings of the 7th International Conference on Learning Representations*, New Orleans, Louisiana, United States, 2019.

[25] Francesco Foscarin, Andrew McLeod, Philippe Rigaux, Florent Jacquemard, and Masahiko Sakai, "ASAP: a dataset of aligned scores and performances for piano transcription," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2020, pp. 534–541.

[26] Valentin Emiya, Nancy Bertin, Bertrand David, and Roland Badeau, "MAPS - A piano database for multipitch estimation and automatic transcription of music," Research report, July 2010.