SCORE-INFORMED TRANSFORMER FOR REFINING MIDI VELOCITY IN AUTOMATIC PIANO TRANSCRIPTION

Zhanhong HE (zhanh.he.uwa@gmail.com) [1], **Roberto TOGNERI**[1], and **Defeng (David) HUANG**[1]

[1] *Signal Processing and Recognition Lab*, **University of Western Australia**, Perth, WA 6000 Australia

ABSTRACT

MIDI velocity is crucial for capturing expressive dynamics in human performances. In practical scenarios, a music score with inaccurate velocities may be available alongside the performance audio (e.g., music education and free online archives), enabling the task of score-informed MIDI velocity estimation. In this work, we propose a modular, lightweight score-informed Transformer correction module that refines the velocity estimates of Automatic Music Transcription (AMT) systems. Leveraging high-quality expressive datasets for rigorous evaluation, we focus on piano transcription, integrating our module into multiple existing systems. Trained exclusively on the MAESTRO dataset, our method consistently reduces velocity errors and improves cross-dataset generalization to SMD and MAPS datasets. Under this training protocol, integrating our score-informed module with HPT (named Score-HPT) establishes a new state-of-the-art, outperforming existing score-informed methods and piano transcription models while adding only 1 M parameters. Furthermore, our framework remains robust to temporal misalignments caused by imperfect audio-score synchronization, a common real-world challenge.

1. INTRODUCTION

Automatic music transcription (AMT) is a longstanding topic in Music Information Retrieval (MIR), aiming to capture symbolic music information from audio. While traditional AMT systems focus on pitch and note boundaries to construct the basic MIDI score [1], recent advancements have expanded to estimating MIDI velocity in piano performances [2]. This expansion has facilitated the creation of large-scale piano datasets [3–7], benefiting downstream MIR tasks such as expressive rendering, music education and generation [8, 9]. Despite recent progress, AMT outputs still require post-correction. Prior work has addressed temporal discrepancies through audio-to-score alignment [10–12]. By contrast, our work focuses on refining the MIDI velocity predicted by AMT systems.

Leveraging music score to aid velocity estimation from audio is known as score-informed MIDI velocity estimation. This paradigm is a practical scenario where a reference score and performance audio co-exist. In music education, for instance, students practice from a given score but require nuanced velocity feedback [13, 14]. Besides, automated dataset curation often relies on online archives that provide aligned audio-score pairs but lack detailed dynamic annotations (e.g., MuseScore.com).

Previous score-informed approaches [14, 15] applied a FiLM conditioning layer to inject score information during audio feature processing. These feature fusion methods require the design of specific model architectures from scratch. In contrast, we propose a modular approach built upon existing piano AMT systems. Fig. 1 conceptually illustrates this distinction. Specifically, we introduce a lightweight Transformer encoder as a score-informed correction module. Inspired by how musicians reference a music score to make corrections, our design refines AMT-estimated MIDI velocity with minimal architectural changes, enabling easy integration into existing systems.

To evaluate its effectiveness, we integrated the proposed module into several well-known piano transcription models, including High-Resolution Piano Transcription (HPT) [16], HPPNet [17], and a recent piano dynamic estimation (DynEst) system [18]. Furthermore, we conducted robustness evaluations under imperfect audio-score synchronization, demonstrating that our framework remains robust to temporal misalignments commonly encountered in real-world scenarios. [1]

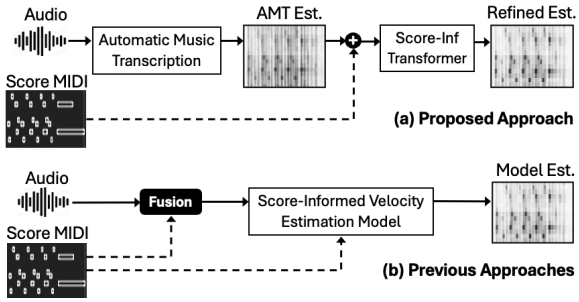


Figure 1. Comparison between the proposed and previous score-informed MIDI velocity estimation approaches.

2. RELATED WORK

2.1 Score-Informed MIDI Velocity Estimation

Early AMT systems often ignored MIDI velocity [1], fostering score-informed velocity estimation as a parallel

[1] Code and pretrained models are available at: https://github.com/zhanh-he/score-informed-amt

Copyright: © 2026. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 Unported License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

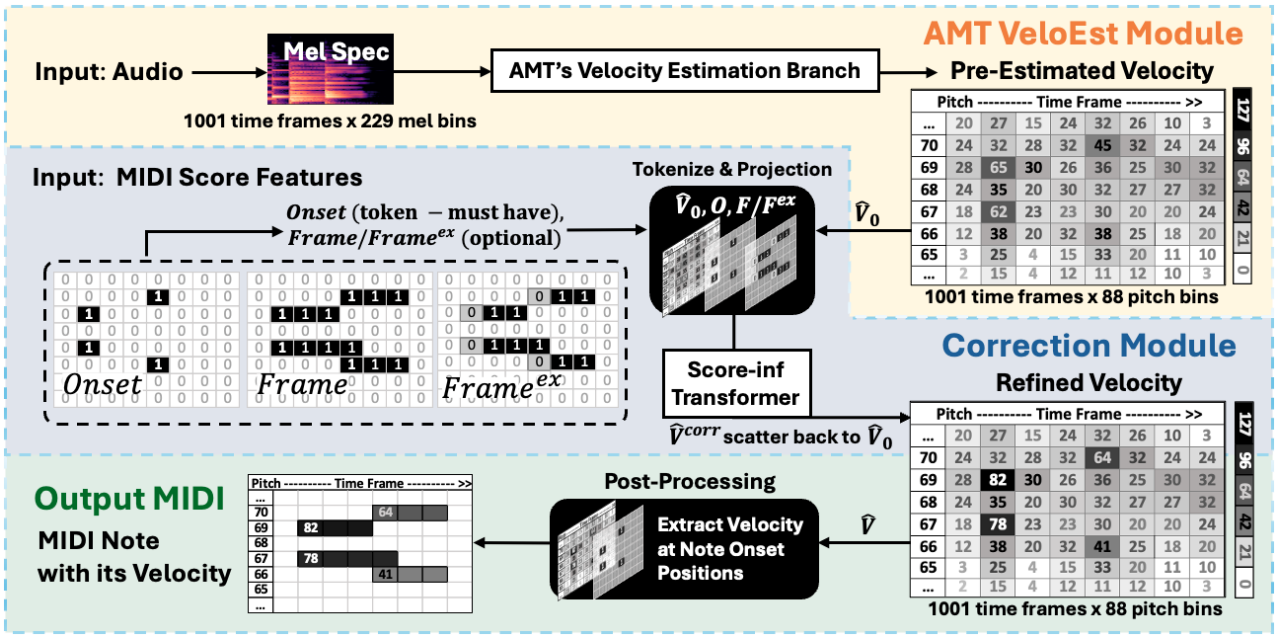


Figure 2. Proposed score-informed MIDI velocity estimation framework. The system includes a velocity estimation module, which is identical to the velocity branch of the piano AMT system. The correction module then rectifies the preliminary velocity estimates using features extracted from the MIDI score.

field. Initial methods relied on manual sound pressure measurements and statistical analysis [19–21], leading to the first parametric estimation system in 2011 [22]. Subsequent research using restricted Boltzmann machines [23] and non-negative matrix factorization [24, 25] proved the feasibility of machine learning for this task. However, their reliance on expert-defined parameters limited generalization and practical deployment.

Recently, deep learning approaches for score-informed velocity estimation have been introduced [15, 26, 27], with the FiLM U-Net model [14] achieving state-of-the-art (SOTA) results. While these end-to-end designs eliminate the need for manual parameter tuning, contemporary audio-only AMT systems have also achieved competitive velocity estimation performance. This convergence motivates our approach: rather than developing a standalone score-informed model from scratch, we propose refining the initial velocity estimates from existing piano AMT systems by leveraging available score information.

2.2 Automatic Piano Transcription

Recent piano AMT systems have increasingly been estimating MIDI velocity, thanks to high-quality expressive piano datasets [28–30]. This progression includes the Onsets and Frames [31], the T5 Transformer [32], the HPT system [16], and the Hierarchical Frequency-Time Transformer (hFT-Transformer) [33]. With a different focus, Semi-CRF [34] was proposed to refine pitch and note-boundary estimates, while its recent development using Transformer (Transkun v2) [35] achieved significant Notew/Off&Velo scores across many piano datasets, representing the current SOTA. Meanwhile, other AMT studies propose competing solutions that prioritize broader generalization [36–38], parameter efficiency [17], or extend the focus to pedal [39].

Among these systems, HPT serves as a highly prevalent baseline. It has been used in creating well-known piano datasets (e.g., EMOPIA [3], GiantMIDI [4], ATEPP [5], PiJAMA [6], Pianist8 [7]) and has been adapted for guitar transcription [40,41] as well as downstream MIR tasks. While HPT may not hold the current SOTA for piano transcription accuracy, its lightweight, modular CRNN architecture ensures computational efficiency and ease of modification. Similar advantageous qualities are found in HPPNet [17] and DynEst [18], making them ideal candidates for evaluating our proposed approach.

3. METHODOLOGY

3.1 MIDI Score Features

Consistent with the standard paradigm of score-informed MIDI velocity estimation [14], we utilize MIDI scores that are time-aligned with the audio. We rasterize the score into pianoroll-like binary matrices at the same time resolution as the AMT output. Let T be the number of frames and P the number of piano keys (T=1001 and P=88 in our setting). We define: (i) onset matrix $\mathbf{O} \in \{0,1\}^{T\times P}$, (ii) frame matrix $\mathbf{F} \in \{0,1\}^{T\times P}$, and (iii) onset-excluded sustain matrix $\mathbf{F}^{\text{ex}} \in \{0,1\}^{T\times P}$:

$$\mathbf{F}^{\text{ex}} = \mathbf{F} - \mathbf{O}. \tag{1}$$

We further extract the nonzero entries of these matrices:

$$\mathcal{I} = \{(t,p) \mid \mathbf{O}_{t,p} = 1\}, \quad \mathcal{J} = \{(t,p) \mid \mathbf{F}_{t,p} = 1\}. \tag{2}$$

They are used for tokenization in our score-informed correction module (Sec. 3.2) and masking the loss function.

3.2 Model Architecture

The proposed architecture, illustrated in Fig. 2, adopts a modular design. Since AMT systems are multi-task architectures that estimate onset, offset, frame, and velocity simultaneously, we isolate their velocity estimation branch and extend it with our score-informed correction module.

Velocity Estimation Module: This component is identical to the velocity branches of the selected piano AMT systems, including HPT [16], HPPNet [17], and DynEst [18]. All setups follow the original implementations and their public repositories. We unify the input for all baselines using log-Mel spectrograms (10 second audio segmentation, waveform normalized, 22.05 kHz sampling rate, 2048 FFT size, 100 FPS, Hann window, 229 mel bins), producing an input tensor of 1001×229. The velocity branch outputs a preliminary velocity map $\hat{\mathbf{V}}_0 \in [0,1]^{T\times P}$.

Score-informed Correction Module: The preliminary velocity $\hat{\mathbf{V}}_0$ is refined using the onset as tokens. For computational efficiency, we tokenize only the score onset coordinates $\mathcal{I}$, resulting in a variable-length sequence of $N = |\mathcal{I}|$ tokens per segment. For each onset token $i = (t_i, p_i) \in \mathcal{I}$, we form an embedding

$$\mathbf{x}_i = \mathbf{e}_{t_i}^{(t)} + \mathbf{e}_{p_i}^{(p)} + \mathbf{W}_f\, \mathbf{f}_i, \tag{3}$$

where $\mathbf{e}_{t_i}^{(t)}$ and $\mathbf{e}_{p_i}^{(p)}$ are learnable pitch/time embeddings. The $\mathbf{f}_i$ contains the onset-wise velocity $\hat{V}_{0,t_i,p_i}$ and optional score-derived features (F or F^{ex}), projected by each component's weight $\mathbf{W}_f$. Stacking all embeddings yields $\mathbf{X} \in \mathbb{R}^{N\times d}$, which is processed by the proposed lightweight Transformer encoder (a streamlined variant of the Conformer architecture [42]), as detailed in Fig. 3. We use d=128, 2 layers, and 4 attention heads, applying a dropout probability of 0.1 after each sub-module. Given the encoder outputs, a linear head predicts a scalar residual δ_i for each token, and we apply a bounded update:

$$\hat{V}^{\text{corr}}_{t_i,p_i} = \text{clip}\Big(\hat{V}_{0,t_i,p_i} + \alpha \tanh(\delta_i),\, 0,\, 1\Big), \tag{4}$$

where α controls the maximum correction magnitude (empirically set to α=0.2). The refined values replace the preliminary estimates in $\hat{\mathbf{V}}_0$ at onset positions, while non-onset entries remain unchanged.

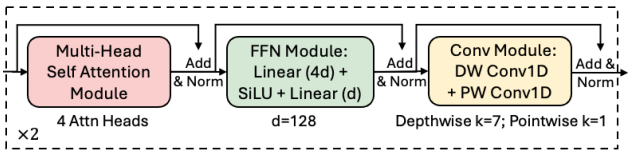


Figure 3. Proposed *Conformer-like* Transformer encoder.

3.3 Loss Function

All models are trained with the hybrid loss proposed in [14]. It combines frame-wise BCE with onset-wise ℓ_1:

$$\mathcal{L} = \frac{1}{|\mathcal{J}|}\sum_{(t,p)\in\mathcal{J}} \ell_{\text{bce}}(V_{t,p}, \hat{V}_{t,p}) + \frac{1}{|\mathcal{I}|}\sum_{(t,p)\in\mathcal{I}} \Big|V_{t,p} - \hat{V}_{t,p}\Big|. \tag{5}$$

where $\mathbf{V}$ and $\hat{\mathbf{V}}$ denote ground-truth and predicted velocity of either (i) the AMT velocity branch or (ii) the same branch with the proposed score-informed module applied. In our preliminary experiments, this multi-scale supervision consistently outperformed single-target alternatives (e.g., frame-only or onset-only losses). Consequently, we adopt this formulation throughout our study, omitting detailed ablation results for brevity.

3.4 Post-Processing

Velocity is defined at note onset. It reflects the instantaneous intensity of a keystroke rather than a time-varying loudness curve [19]. This onset-centered definition is consistent with the physical mechanism of the piano. Once a key has been fully pressed down, subsequent changes in finger pressure do not affect the tone. In other words, vibrato is technically impossible on an acoustic piano.

In our setting, a time-aligned MIDI score is available, but it does not provide reliable expressive dynamics labels such as note velocity. Instead, it only provides the ground-truth note positions, i.e., onsets and frames. We therefore bypass AMT-style note detection and focus only on velocity estimation. Consequently, we take the predicted velocity at each onset and assign this value to the entire note event, consistent with the MIDI specification.

4. EXPERIMENT

4.1 Dataset

In this study, we used the MAESTRO v3.0.0 dataset [28] with its default train/validation/test split. This dataset comprises 1,276 Yamaha Disklavier piano performances, totaling over 200 hours of precisely aligned audio-MIDI data. A Yamaha Disklavier is an acoustic grand piano with an integrated electronic system that records MIDI data from human performance. Thus, MAESTRO provides audio captured by microphones, reflecting the acoustic environment, and its tightly synchronized MIDI data.

Following previous studies [14], we trained our model exclusively on the MAESTRO train set, and evaluated it on the Saarland Music Data (SMD) dataset [29]. SMD consists of human performances recorded on a Yamaha Disklavier piano, but differs from MAESTRO in acoustic environment and recording conditions. To further assess cross-dataset generalization, we evaluated our model on 60 piano recordings from the MAPS dataset [30], exposing it to an additional set of acoustic scenarios.

4.2 Training Setup

For each selected AMT system, we trained two models in similar configurations. The baseline model is the standalone velocity branch of the original AMT system. For instance, our HPT baseline utilizes only the velocity branch, which contains about one-quarter of the full HPT system's parameters. Its score-informed counterpart (e.g., Score-HPT) integrates this same branch with our proposed correction module. This ensures a fair comparison by isolating the impact of score information on velocity estimation.

Each model was trained for 120k iterations on a single NVIDIA RTX 3090 (24 GiB). We used Adam with learning rate 1×10^{-4} (multiplied by 0.9 every 10k iterations), batch size 12, and random seed 86. All models were trained on the MAESTRO training split. We selected the best checkpoint by validation set performance and then proceeded to evaluation.

4.3 Evaluation Metrics

Consistent with [14], we use the mean absolute error (MAE) and standard deviation of the error (STD) as evaluation metrics, calculated on the standard MIDI velocity scale of $[0, 127]$. By applying onset-masking, we ensure a note-level evaluation that avoids the temporal bias of frame-wise metrics, where sustained notes would otherwise disproportionately influence the total error. These metrics are defined as:

$$\begin{aligned} \mathrm{MAE} &= \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \left| V_i - \hat{V}_i \right|, \\ \mathrm{STD} &= \sqrt{\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \Big(|V_i - \hat{V}_i| - \mathrm{MAE} \Big)^2}. \end{aligned} \tag{6}$$

While previous work [14] (which predicts a sparse map) used Recall to penalize missing velocity at note positions, this metric is not discriminative in our setting. As shown in Fig. 4, both the AMT baseline and our method output fully-populated velocity maps, so the velocity is non-zero at all potential note positions, leading to a Recall close to 100%. Since the post-processing is applied to extract velocities at known note positions, we prioritize numerical precision (MAE / STD) rather than penalizing off-note predictions.

5. RESULTS AND DISCUSSION

5.1 Validation of Score-Informed Module

To validate the proposed score-informed Transformer correction module, we first integrate it into the HPT [16]. We choose HPT as an initial testbed because it is widely used and lightweight, which allows us to isolate the contribution of the correction module before extending to additional baselines. We denote the resulting system as Score-HPT.

Table 1 summarizes the results on the MAESTRO test set. Compared with the original HPT, Score-HPT consistently achieves lower MAE and STD under all conditioning variants, with only a modest increase in model parameters (from 4.1 M to 5.0 M). Among these Score-HPT variants, *onset* + *frame*$_{ex}$ yields the best overall performance, while *onset* + *frame* performs comparably. These results confirm that the proposed score-informed module effectively refines velocity estimates upon an existing AMT system. Additionally, Fig. 4 provides a qualitative visualization of this improvement. Compared to the HPT, Score-HPT generates velocity estimates that are more closely aligned with the MIDI note onset positions, clearly demonstrating the guiding effectiveness of the proposed score-informed Transformer module.

Model & Inputs	MAESTRO test set		
	MAE ↓	STD ↓	# Params
HPT			
audio	4.6	6.1	4.1 M
Score-HPT			
audio + onset	4.0	4.8	5.0 M
audio + onset + frame	4.0	**4.7**	5.0 M
audio + onset + frame$_{ex}$	**3.9**	**4.7**	5.0 M

Table 1. Comparison of HPT and Score-HPT on the MAESTRO test set. ↓ indicate a lower value is better.

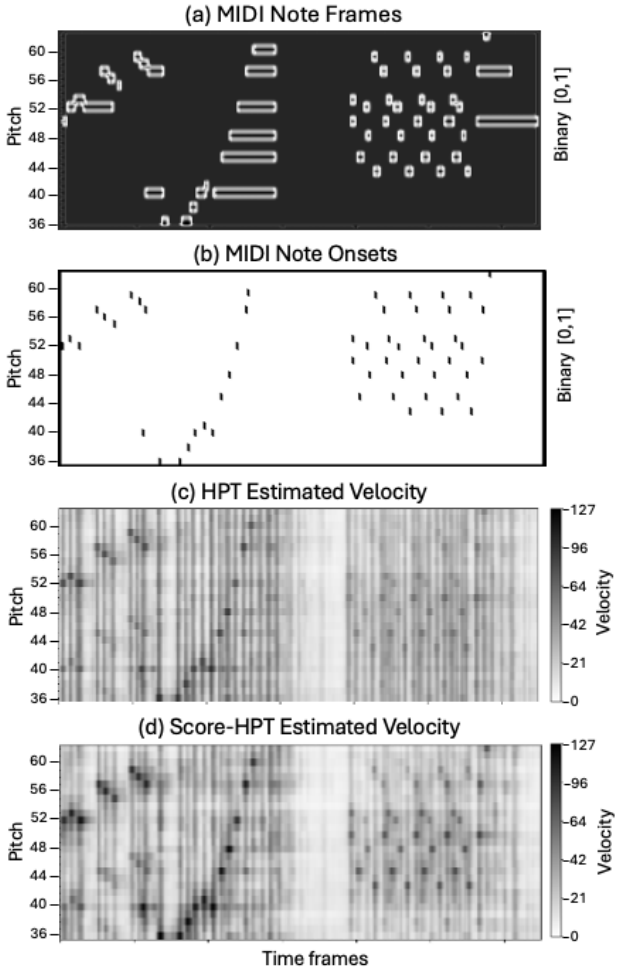


Figure 4. Velocity estimates from the HPT baseline and the proposed Score-HPT. Referencing the given MIDI note features, the refinement effect is evident. Both HPT and Score-HPT provide a fully-populated velocity estimation.

5.2 Generalization across Datasets

Fig. 5 illustrates the learning curves for the HPT and Score-HPT variants. While training on the MAESTRO train set, models were evaluated on MAESTRO validation set, alongside 10% held-out subsets of SMD and MAPS datasets (which lack official validation splits). All variants of Score-HPT consistently outperform the HPT baseline on both in-domain (MAESTRO) and out-of-distribution

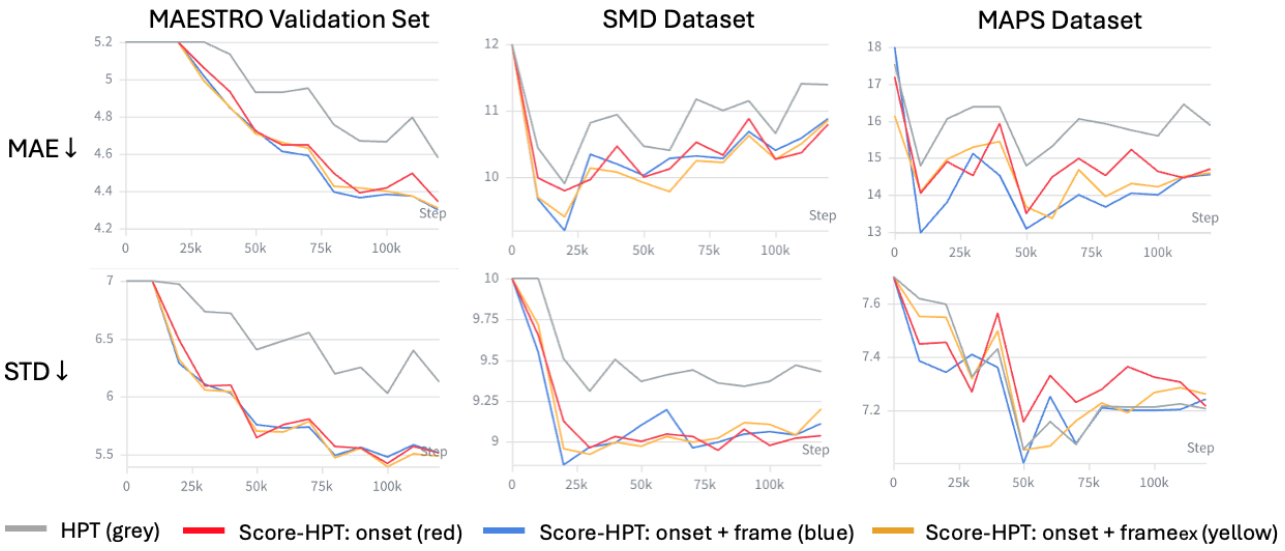


Figure 5. Comparison between HPT and Score-HPT variants using different score features. Models were trained on the MAESTRO train set for 120k iterations and evaluated every 10k iterations across three validation sets.

(SMD, MAPS) data, demonstrating superior generalization. The *onset + frame* condition yields the best overall performance, though differences between conditions are modest.

During training in MAESTRO, the in-domain performance steadily improves but out-of-distribution performance peaks early and declines. This indicates that prolonged training overfits to MAESTRO-specific acoustics and harms generalization. Similar sound biases in AMT velocity estimation were also reported in [43]. Nevertheless, Score-HPT improves both in-domain and out-of-distribution results. This highlights the benefits of score information, which provides an acoustics-invariant structural cue for reliable velocity estimation across diverse environments.

5.3 Comparison with Existing Works

Table 2 evaluates the proposed score-informed framework against existing methods and representative AMT systems on the SMD dataset. All models were trained on MAESTRO only, enabling a fair comparison. Following the findings in Sec. 5.2, we employ the *onset + frame* configuration for all proposed variants.

To further demonstrate the generality of our approach, we integrated the correction module into diverse architectures beyond HPT [16], including HPPNet [17] and DynEst [18]. Although retraining these models on log-Mel spectrograms rather than their original CQT or Bark-scale specific loudness (BSSL) features inherently degrades their performance, our score-informed module consistently yields significant improvements. This confirms that our approach serves as a general-purpose enhancement for different piano AMT backbones, while adding only a marginal overhead of ∼1M parameters. Through this integration, Score-HPT and Score-DynEst achieve new SOTA performance. These models outperform both dedicated score-informed frameworks, such as FiLM U-Net [14], and

Model	SMD dataset		
	MAE ↓	STD ↓	# Params
Score-Inf. Methods			
DiffVel [27]	19.7	13.1	–
FiLM Conv [15]	15.1	12.3	–
FiLM U-Net [14]	9.9	7.8	8.8 M
Proposed			
Score-HPT	**8.0**	**6.9**	5.0 M
Score-HPPNet	9.2	7.3	1.5 M
Score-DynEst	**8.0**	**6.9**	14.3 M
AMT Systems			
HPT [16]	10.0	8.8	4.1 M
HPPNet (no CQT) [17]	10.6	9.0	0.5 M
DynEst (no BSSL) [18]	10.4	8.8	13.3 M
hFT-Transformer [33]	9.9	7.3	5.5 M
Transkun v2 [35]	13.3	11.9	13.6 M

Table 2. Comparison of different methods on the SMD dataset. ↓ indicate a lower value is better.

top-tier AMT systems, including hFT-Transformer [33] and Transkun v2 [35]. Interestingly, although Transkun v2 [35] currently defines the SOTA in AMT, it lags behind several other models in velocity MAE and STD. A plausible explanation is that its core SemiCRF is primarily designed for pitch and note onset/offset modeling, which enables strong transcription performance (Notew/Off&Velo), whereas velocity estimation is not its main strength. Incorporating our score-informed correction module into Transkun v2 therefore remains a promising future direction.

5.4 Robustness to Audio-Score Misalignment

To evaluate the models against audio-score temporal misalignments, we conducted robustness tests using timing

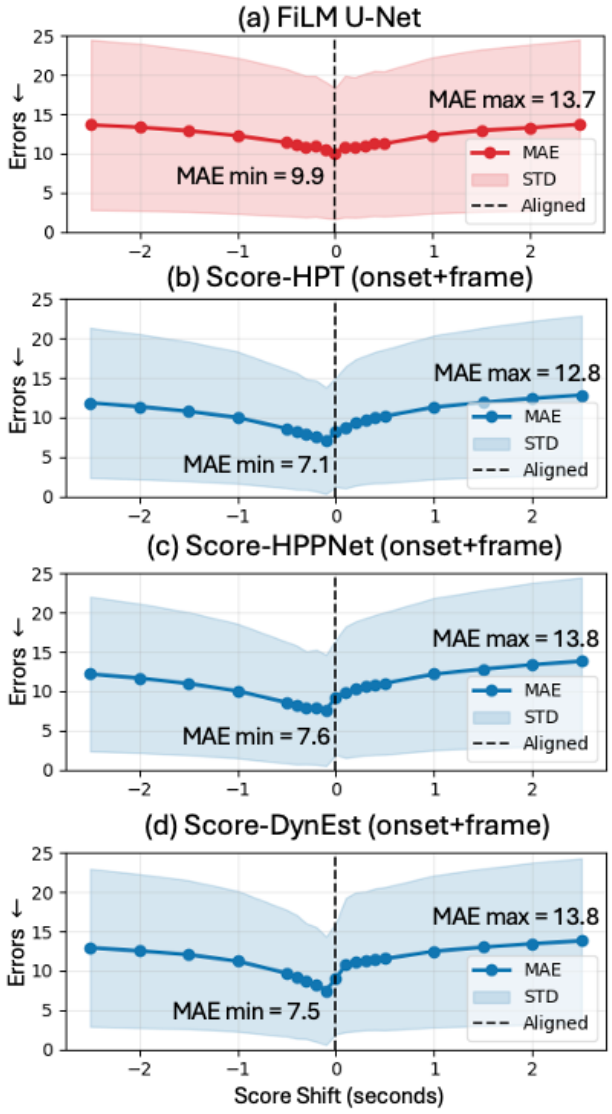


Figure 6. Robustness evaluation on data derived from the SMD dataset, using audio-score misalignments of ±0.1, 0.2, 0.3, 0.4, 0.5, 1, 1.5, 2, and 2.5 s.

shifts ranging from ±0.1 to ±2.5 seconds, following [14]. We applied these artificial shifts on SMD dataset, forming the evaluation data. For example, a +0.1 shift means that the original, perfectly aligned score is intentionally delayed by 0.1 seconds. All models evaluated here are exactly the same as those in Sec. 5.3, meaning they were trained exclusively on the MAESTRO dataset (not shifted, perfect synchronization). The final robustness results are summarized in Fig. 6.

Specifically, we introduce a necessary refinement to the evaluation pipeline of [14]. Their approach used shifted scores to feed the model but unshifted scores for post-processing to be consistent with the earlier benchmark [15]. While this intention is understandable, it introduces a timeline mismatch: if a model is guided by a shifted score, its predictions are inherently tied to that shifted temporal grid. Therefore, we argue that the model's output should be post-processed using the shifted MIDI score. Under this more internally consistent protocol, FiLM-UNet performs significantly better than originally reported in [14], revealing its true robustness as shown in Fig. 6.

Applying this refined evaluation protocol, Score-HPT achieves the best performance on the SMD dataset under misaligned settings. Notably, all methods in Fig. 6 degrade only marginally compared to the audio-score perfectly aligned scenario (score shift = 0 second). Given that MIDI velocity ranges from 0 to 127, an increase of around 6 units in MAE or STD under misalignment corresponds to a very small absolute deviation relative to the full velocity range. Furthermore, we observe that minor shifts unexpectedly improve the performance of our proposed score-informed variants, likely by offsetting the inherent fuzziness of the AMT branch's velocity prediction. Together, these findings highlight an important conclusion: score-informed MIDI velocity estimation methods are highly robust to audio-score misalignments.

6. CONCLUSION AND FUTURE WORK

In this work, we presented a novel, modular score-informed Transformer correction module for refining MIDI velocity in automatic piano transcription when a reference score is available. By operating on score onset tokens and predicting bounded residual updates, the proposed module can be attached to an existing AMT velocity branch with minimal architectural changes. Experiments with MAESTRO-only training demonstrate consistent improvements across multiple piano transcription systems (HPT, HPPNet, and DynEst), stronger cross-dataset generalization to SMD and MAPS, and state-of-the-art performance on SMD under our training protocol, surpassing both prior score-informed models and competitive piano transcription systems. We further demonstrate that the proposed framework is robust to audio-score temporal misalignments.

A key limitation of the current study is its restriction to piano datasets, constrained by the fact that high-fidelity MIDI velocity ground truth is predominantly captured via acoustic pianos equipped with velocity sensors. For many other instruments, velocity annotations are scarce or unavailable, making supervised training and evaluation difficult. An important direction for future work is to transfer the proposed framework from piano to other instruments for MIDI velocity estimation, potentially leveraging self-supervised learning, transfer learning, or continual learning to reduce dependence on explicit velocity labels.

Acknowledgments

This research was carried out while the author was in receipt of a University International Fees Scholarship and a University Postgraduate Award at The University of Western Australia. The authors acknowledge the use of Kaya, the High-Performance Computing facility at The University of Western Australia, and the technical assistance provided by the UWA IT – HPC team.

7. REFERENCES

[1] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri, "Automatic music transcription: Challenges and future directions," *Journal of Intelligent Information Systems*, vol. 41, no. 3, pp. 407–434, 2013.

[2] E. Benetos, S. Dixon, Z. Duan, and S. Ewert, "Automatic music transcription: An overview," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 20–30, 2019.

[3] H.-T. Hung, J. Ching, S. Doh, N. Kim, J. Nam, and Y.-H. Yang, "Emopia: A multi-modal pop piano dataset for emotion recognition and emotion-based music generation," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2021, pp. 318–325.

[4] Q. Kong, B. Li, J. Chen, and Y. Wang, "Giantmidi-piano: A large-scale midi dataset for classical piano music," *Transactions of the International Society for Music Information Retrieval*, May 2022.

[5] H. Zhang, J. Tang, S. R. M. Rafee, S. Dixon, and G. Fazekas, "ATEPP: A dataset of automatically transcribed expressive piano performance," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2022, pp. 446–453.

[6] D. Edwards, S. Dixon, and E. Benetos, "Pijama: Piano jazz with automatic midi annotations," *Transactions of the International Society for Music Information Retrieval*, vol. 6, no. 1, pp. 89–102, 2023.

[7] Y.-H. Chou, I.-C. Chen, J. Ching, C.-J. Chang, and Y.-H. Yang, "Midibert-piano: Large-scale pre-training for symbolic music classification tasks," *Journal of Creative Music Systems*, vol. 8, no. 1, 2024.

[8] H. Zhang, S. Chowdhury, C. E. Cancino-Chacón, J. Liang, S. Dixon, and G. Widmer, "Dexter: Learning and controlling performance expression with diffusion models," *Applied Sciences*, vol. 14, no. 15, p. 6543, 2024.

[9] Z. He, D. Cooper, D. Huang, and R. Togneri, "Filling midi velocity using u-net image colorizer," in *Proceedings of the International Symposium on Computer Music Multidisciplinary Research (CMMR)*, 2025, pp. 949–960.

[10] A. Morsi and X. Serra, "Bottlenecks and solutions for audio to score alignment research," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2022, pp. 272–279.

[11] J. Zeitler, B. Maman, and M. Müller, "Robust and accurate audio synchronization using raw features from transcription models," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2024, pp. 120–127.

[12] S. Chang, S. Dixon, and E. Benetos, "Rumaa: Repeat-aware unified music audio analysis for score-performance alignment, transcription, and mistake detection," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2025, pp. 1–5.

[13] H. Kim, P. Ramoneda, M. Miron, and X. Serra, "An overview of automatic piano performance assessment within the music education context," in *Proceedings of the International Conference on Computer Supported Education (CSEDU)*, 2022, pp. 465–474.

[14] H. Kim and X. Serra, "A method for midi velocity estimation for piano performance by a U-net with attention and FiLM," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2024, pp. 304–310.

[15] H. Kim, M. Miron, and X. Serra, "Score-informed midi velocity estimation for piano performance by film conditioning," in *Proceedings of the Sound and Music Computing Conference (SMC)*, 2023, pp. 139–147.

[16] 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.

[17] W. Wei, P. Li, Y. Yu, and W. Li, "HPPNet: Modeling the harmonic structure and pitch invariance in piano transcription," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2022, pp. 709–716.

[18] Z. He, H. Meng, D. Huang, and R. Togneri, "Joint estimation of piano dynamics and metrical structure with a multi-task multi-scale network," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026.

[19] R. B. Dannenberg, "The interpretation of midi velocity," in *Proceedings of the International Computer Music Conference (ICMC)*, 2006, pp. 193–196.

[20] W. Goebl, "The role of timing and intensity in the production and perception of melody in expressive piano performance," Ph.D. dissertation, Karl-Franzens-Universität Graz, Graz, Austria, 2003.

[21] W. M. Szeto, K. H. Wong, and C. H. Wong, "Finding intensities and temporal characteristics in piano music," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, 2005, pp. 214–219.

[22] S. Ewert and M. Müller, "Estimating note intensities in music recordings," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 385–388.

[23] S. van Herwaarden, M. Grachten, and W. B. de Haas, "Predicting expressive dynamics in piano performances using neural networks," in *Proceedings of the*

International Society for Music Information Retrieval Conference (ISMIR), 2014, pp. 47–52.

[24] D. Jeong and J. Nam, "Note intensity estimation of piano recordings by score-informed nmf," *Journal of the Audio Engineering Society*, vol. 65, no. 4/5, pp. 263–273, June 2017.

[25] D. Jeong, T. Kwon, and J. Nam, "A timbre-based approach to estimate key velocity from polyphonic piano recordings," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2018, pp. 120–126.

[26] F. Simonetta, S. Ntalampiras, and F. Avanzini, "Acoustics-specific piano velocity estimation," in *Proceedings of the IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 2022.

[27] H. Kim and X. Serra, "DiffVel: Note-level midi velocity estimation for piano performance by a double conditioned diffusion model," in *Proceedings of the International Symposium on Computer Music Multidisciplinary Research (CMMR)*, 2023, pp. 197–208.

[28] 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 *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

[29] M. Müller, V. Konz, W. Bogler, and V. Arifi-Müller, "Saarland music data (smd)," in *Late-Breaking and Demo Session of the 12th Int. Conf. on Music Information Retrieval*, 2011.

[30] V. Emiya, N. Bertin, B. David, and R. Badeau, "Maps: A piano database for multipitch estimation and automatic transcription of music," INRIA, Tech. Rep. 2010D017, 2010.

[31] C. Hawthorne, A. Stasyuk, A. Roberts, I. Simon, S. Huang, S. Dieleman, E. Elsen, J. Engel, and D. Eck, "Onsets and frames: Dual-objective piano transcription," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2018, pp. 50–57.

[32] C. Hawthorne, I. Simon, R. Swavely, E. Manilow, and J. Engel, "Sequence-to-sequence piano transcription with transformers," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2021, pp. 246–253.

[33] K. Toyama, T. Akama, Y. Ikemiya, Y. Takida, W.-H. Liao, and Y. Mitsufuji, "Automatic piano transcription with hierarchical frequency-time transformer," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2023, pp. 215–222.

[34] Y. Yan, F. Cwitkowitz, and Z. Duan, "Skipping the frame-level: Event-based piano transcription with neural semi-crfs," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021, pp. 20 583–20 595.

[35] Y. Yan and Z. Duan, "Scoring time intervals using non-hierarchical transformer for automatic piano transcription," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2024, pp. 973–980.

[36] B. Maman and A. H. Bermano, "Unaligned supervision for automatic music transcription in the wild," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2022, pp. 14 918–14 934.

[37] D. Edwards, S. Dixon, E. Benetos, A. Maezawa, and Y. Kusaka, "A data-driven analysis of robust automatic piano transcription," *IEEE Signal Processing Letters*, vol. 31, pp. 681–685, 2024.

[38] S. Strahl and M. Müller, "Semi-supervised piano transcription using pseudo-labeling techniques," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2024, pp. 173–181.

[39] Q. Wang, M. Liu, and M. Jia, "Multitask frequency-time feature-wise linear modulation for piano transcription with pedal," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 33, pp. 1049–1062, 2025.

[40] X. Riley, D. Edwards, and S. Dixon, "High resolution guitar transcription via domain adaptation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 1051–1055.

[41] X. Riley, Z. Guo, D. Edwards, and S. Dixon, "GAPS: a large and diverse classical guitar dataset and benchmark transcription model," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2024.

[42] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2020, pp. 5036–5040.

[43] L. S. Marták, P. Hu, and G. Widmer, "Sound and music biases in deep music transcription models: a systematic analysis," *EURASIP J. Audio Speech Music. Process.*, vol. 2026, p. 5, 2026.