

Score-Informed BiLSTM Correction for Refining MIDI Velocity in Automatic Piano Transcription

Zanhong He¹, Roberto Tognari¹, Defeng (David) Huang¹,

¹University of Western Australia, Perth, Australia

Abstract—MIDI is a modern standard for storing music, recording how musical notes are played. Many piano performances have corresponding MIDI scores available online. Some of these are created by the original performer, recording on an electric piano alongside the audio, while others are through manual transcription. In recent years, automatic music transcription (AMT) has rapidly advanced, enabling machines to transcribe MIDI from audio. However, these transcriptions often require further correction. Assuming a perfect timing correction, we focus on the loudness correction in terms of MIDI velocity (a parameter in MIDI for loudness control). This task can be approached through score-informed MIDI velocity estimation, which has undergone several developments. While previous approaches introduced specifically built models to re-estimate MIDI velocity, thereby replacing AMT estimates, we propose a BiLSTM correction module to refine AMT-estimated velocity. Although we did not reach state-of-the-art performance, we validated our method on the well-known AMT system, the high-resolution piano transcription (HPT), and achieved significant improvements.

1. INTRODUCTION

Automatic music transcription (AMT) is a longstanding topic in Music Information Retrieval (MIR), dedicated to extracting musical notes from audio recordings and converting them into MIDI scores. Traditional AMT systems concentrated on estimating the pitch and note boundary to construct the basic MIDI score [1]. Recent developments have extended to tasks such as instrument identification in music ensembles and MIDI velocity estimation for piano performances [2]. These advances have enabled large-scale transcription, resulting in many datasets [3]–[8] that are valuable for MIR downstream research. In most cases, AMT outputs require corrections. This challenge has driven the development of audio-to-MIDI alignment to correct timing discrepancies [9]–[11] and motivated our research to correct the MIDI velocity.

MIDI velocity controls the loudness of each musical note. Together with note timing, it shapes the expressiveness of a performance. Accurate velocities make MIDI scores valuable guides in music education [12] and crucial data for music generative research [13]. However, manually correcting velocities is laborious: people perceive loudness differently [14], and the fine granularity of the MIDI velocity (from 0 to 127) makes human judgments inconsistent. Refining AMT-estimated velocities with artificial intelligence offers an optimal solution. This task, known as score-informed MIDI velocity estimation, assumes audio recordings paired with a perfectly time-aligned MIDI score whose velocities are imprecise or missing. Such a MIDI score, an ideal prerequisite for this task, would be obtained by applying timing corrections to the AMT output, either through alignment techniques or manually.

In this paper, we propose a bidirectional long short-term memory (BiLSTM) module designed specifically to refine, rather than replace entirely, the velocity estimates provided by an existing AMT system, leveraging information from the corresponding MIDI score. Figure 1 conceptually illustrates this distinction. While previous score-informed approaches [12], [15], [16] involved developing comprehensive models tailored for complete velocity re-estimation from scratch, our strategy of implementing a correction module that builds upon the output of

an established AMT system, offering adaptability to AMT baseline systems and reduced development overhead. The use of a BiLSTM allows the module to effectively capture the sequential context of notes within the performance, crucial for accurate velocity adjustments.

To evaluate our approach, we integrated the correction module into the high-resolution piano transcription (HPT) system [17]. Although the results did not surpass the state-of-the-art (SOTA) in score-informed velocity estimation, our method yielded significant improvements when applied on the HPT system. This demonstrates the effectiveness of adding correction module as a practical strategy.

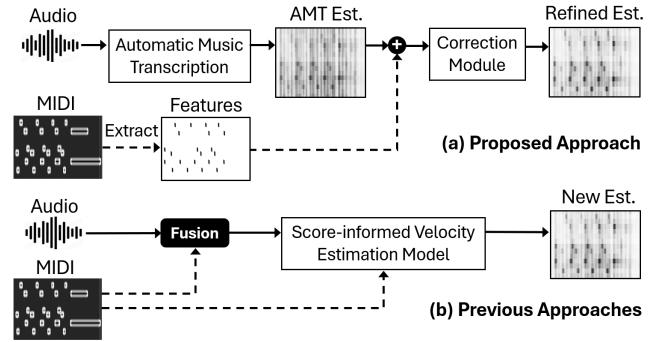


Fig. 1: Comparison between the proposed and previous approaches.

2. RELATED WORKS

2.1. Score-Informed MIDI Velocity Estimation

Historically, AMT systems could only predict MIDI notes without considering velocity information [1]. Consequently, score-informed MIDI velocity estimation emerged as a research focus. Early work relied on manual measurements of sound pressure level and statistical methods [18]–[20], culminating in 2011 with the first automatic velocity estimation system based on parametric modeling [21]. Subsequent efforts applied restricted Boltzmann machines [22] and non-negative matrix factorization [23], [24], demonstrating the feasibility of machine learning methods. Nevertheless, these approaches require expert-defined parameters for each inference, limiting their generalization across data and practical deployment.

Deep learning approaches for score-informed velocity estimation were first introduced in 2023 [12], [15] and developed further in [16]. These methods eliminate the need for expert parameter tuning. However, at the same time, AMT systems capable of estimating velocity have achieved comparable performance. This pivots our focus toward refining the AMT-estimated velocity instead of developing a new score-informed MIDI velocity estimation model from scratch.

2.2. Automatic Music Transcription

Recent AMT systems began estimating MIDI velocity, starting with OaF [25], and continued through the T5 transformer [26], the HPT system with its convolutional recurrent neural networks (CRNN)

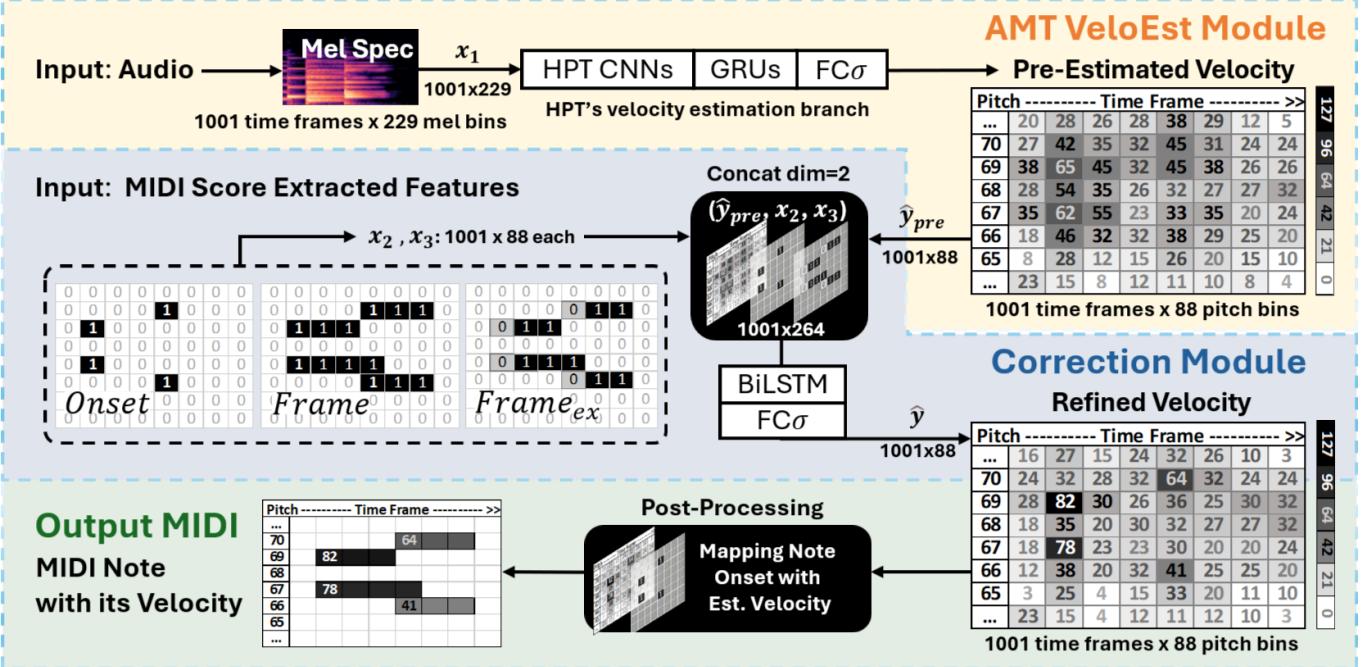


Fig. 2: Proposed Score-HPT architecture and workflow. The system includes a velocity estimation module, which is the same as in the HPT system. The correction module then rectifies the preliminary velocity estimates using features extracted from the MIDI score. In this figure, velocity values are de-normalized from (0,1) to their original [0, 127] scale for better visualization.

[17], and a transformer replacement for HPT’s GRU modules [27]. Semi-CRF [28], [29] was proposed to refine pitch and note-boundary estimates, while harmonic attention [30] was proposed for the same purpose. More recently, the hFT-transformer [31] achieved significant Notew/Off&Velo scores on the MAESTRO dataset [32], representing the current SOTA. Meanwhile, other AMT studies propose competing solutions that prioritize broader generalization [33], [34] or fewer model parameters [35].

Among AMT systems, HPT is a widely used baseline, supporting extended tasks such as creating large piano datasets (e.g., EMOPIA [3], GiantMIDI [4], ATEPP [5], PiJAMA [6], Pianist8 [8]), adapting to guitar transcription [7], [36], music style and emotion classification [8], music synthesis and source separation [37], [38]. While HPT is not the current SOTA in piano transcription, its lightweight and modular CRNN architecture offers ease of modification and computational efficiency. These qualities make it an ideal foundation for our approach and enable broader research applications.

3. METHODOLOGY

3.1. Model Architecture

Figure 2 illustrates our proposed architecture, Score-HPT, which extends the HPT system [17] by adding a score-informed velocity correction module.

HPT Velocity Estimation: Since HPT is a multitask structural system, we isolate its MIDI velocity estimation branch. This branch provides preliminary velocity estimates from the audio signal. All settings in this component are the same as in HPT [17], including the CRNN architecture and mel-spectrogram extractor. Audio is converted into mel-spectrogram tensors of size 1001 × 229 (frames × bins) as input. The branch’s output is a 1001 × 88 matrix of preliminary velocity estimates, with values normalized from 0-127 to [0, 1] for stable training.

Score-informed BiLSTM Correction: The preliminary velocity estimates are then refined by our correction module. These pre-estimates are concatenated in parallel with MIDI score extracted features, where their overlap provides crucial cues for the correction. The combined inputs are processed by a BiLSTM layer with 256 hidden units per direction, yielding a 512-dimensional hidden state for each frame. This is then passed through a fully connected layer with Sigmoid activation, mapping to 88 output units, providing a refined velocity estimate for each piano key.

Mapping Note Onset: This is a standard post-processing step in generating the MIDI output, because velocity represents the intensity of a keystroke at that instant, rather than a time-varying loudness [18]. Standard AMT systems perform this mapping using predicted onsets, where timing and velocity estimation errors are often interdependent. For our task, we assume corrected MIDI note timing and utilize ground truth note information, thereby focusing velocity estimation.

3.2. MIDI Score Features Extraction

While HPT provides preliminary velocity estimates from audio, we combine them by concatenating three types of pianoroll-like features derived from the MIDI score:

- 1) the note-on event matrix $Onset \in \{0, 1\}^{T \times P}$, marking the key attacks;
- 2) the note duration matrix $Frame \in \{0, 1\}^{T \times P}$, marking the note active frames;
- 3) the onset-excluded frame matrix $Frame_{ex} \in \{0, 1\}^{T \times P}$, marking the note sustain frames:

$$Frame_{ex} = Frame - Onset, \quad (1)$$

here, $T = 1001$ is the number of time frames and $P = 88$ is the number of piano keys. These matrix features help the model distinguish velocity patterns at the note start from those during sustain or silence.

4. EXPERIMENT

4.1. Dataset

In this study, we used the MAESTRO v3.0.0 dataset [32] with its default train/validation/test split. This dataset comprises 1,276 Yamaha Disklavier piano performances captured during the International Piano-e-Competition, totaling over 200 hours of precisely aligned audio-MIDI data. Yamaha Disklavier is the acoustic grand piano with an integrated electronic system that records MIDI data directly from human actions. Thus, MAESTRO provides audio captured by microphones, reflecting the acoustic environment, and its tightly synchronized MIDI data.

Following previous studies [16], we trained our model on the MAESTRO train set only, and evaluated it on the same 49 performances from the Saarland Music Data (SMD) dataset [39]. SMD were real recordings on the Yamaha Disklavier piano but differ in acoustic environment and recording conditions. To further assess generalization, we applied the MAPS dataset [40], which comprises 60 recordings of Yamaha Disklavier pianos. This exposes the model to yet another set of acoustic scenarios.

4.2. Training Setup

To track how model evolves during training, we retrained the HPT velocity estimation branch and compared it with Score-HPT. We used the velocity binary cross-entropy (BCE) loss function, separate from HPT’s multitask loss design [17]:

$$l_{\text{velo}} = \sum_{i=1}^{|I|} l_{\text{bce}}(y_i, \hat{y}_i), \quad I = \{(t, p) \mid \text{Onset}_{t,p} = 1\} \quad (2)$$

Here, the summation is performed for all elements i within the set I . This set I includes all time-pitch coordinates (t, p) where an actual note onset occurs (i.e., $\text{Onset}_{t,p} = 1$, acting as a mask). The y_i and \hat{y}_i are the ground truth and model-estimated velocity, respectively, and l_{bce} is the standard BCE loss function as implemented in PyTorch. The training ran for 200k iterations on a single Nvidia P100 16GiB GPU, taking roughly two days. We used the Adam optimizer with an initial learning rate of 1×10^{-4} decayed by 0.9 every 10k iterations, a batch size of 12, and a fixed random seed of 13. Training was carried out exclusively on the MAESTRO training set. We selected the best checkpoint of each model based on validation performance, and then proceeded to evaluation.

4.3. Evaluation Metrics

To ensure consistency with previous studies [16], we adopt the mean absolute error (MAE), standard deviation of error (STD), and Recall to evaluate the model. While training utilizes normalized velocity values (y_i, \hat{y}_i) within a range $[0, 1]$ for the BCE loss, all evaluations are performed using the denormalized ground truth (Y_i) and model-estimated (\hat{Y}_i) velocities, which are on the 0-127 scale. The MAE is defined as:

$$\text{MAE} = \frac{1}{|I|} \sum_{i \in I} |Y_i - \hat{Y}_i|, \quad (3)$$

where $I = \{(t, p) \mid O_{t,p} = 1\}$ is the set of active onset positions. The STD is computed analogously, as the standard deviation of the per-note absolute errors. Specifically, the Recall is derived from the Notew/Off&Velo metrics in mir_eval toolkit [41]. Notew/Off&Velo employs two thresholds: a 50ms timing tolerance for onset/offset detection (i.e., note-on/off event) and a $\pm 10\%$ velocity error tolerance. An estimate is counted as correct only if it satisfies both timing and velocity criteria.

5. RESULTS AND DISCUSSION

5.1. Validation of Score-HPT

Both Table 1 and Fig. 3 can demonstrate the effectiveness of our proposed method, integrating the score-informed BiLSTM correction module into HPT velocity estimation branch. The resulting system, Score-HPT, with all configurations outperformed its HPT baseline on the MAESTRO test set across all metrics. Specifically, the "audio + onset" configuration performed the best, while the "audio + onset + frame_{ex}" configuration delivered similar performance. Focusing on Recall, the results are summarized as follows:

- Original HPT - (79.8%): Reported by Kong et al. [17], this Recall obtained by mapping both HPT estimated onsets and velocities. It represents the vanilla HPT transcription, with relatively low Recall highlights the need for timing and velocity corrections.
- HPT - 90.6%: Here, HPT estimated velocities are mapped with ground-truth onsets instead of its predictions. This isolates the timing errors, assuming the timing correction is done, showing the maximum gain achievable with audio-to-MIDI alignment.
- Score-HPT - 95.6%: HPT-estimated velocities are refined by our BiLSTM correction module guided by precise MIDI score features (i.e. onsets and/or frames). The highest Recall confirms the effectiveness of score-informed correction for closing the gap to perfect transcription.

Table 1: Comparison of HPT and Score-HPT on the MAESTRO test set. ↑ and ↓ indicate whether higher or lower values are better.

Model & Inputs			MAESTRO test set		
x1	x2	x3	MAE ↓	STD ↓	Recall ↑
HPT					(79.8%)
audio			5.05	4.85	90.6%
Score-HPT					
audio	+ onset		3.48	3.43	95.6%
audio	+ frame		3.56	3.52	95.3%
audio	+ frame _{ex}		3.49	3.49	95.3%
audio	+ onset	+ frame	3.51	3.46	95.5%
audio	+ onset	+ frame _{ex}	3.50	3.43	95.6%
audio	+ frame	+ frame _{ex}	3.52	3.51	95.3%

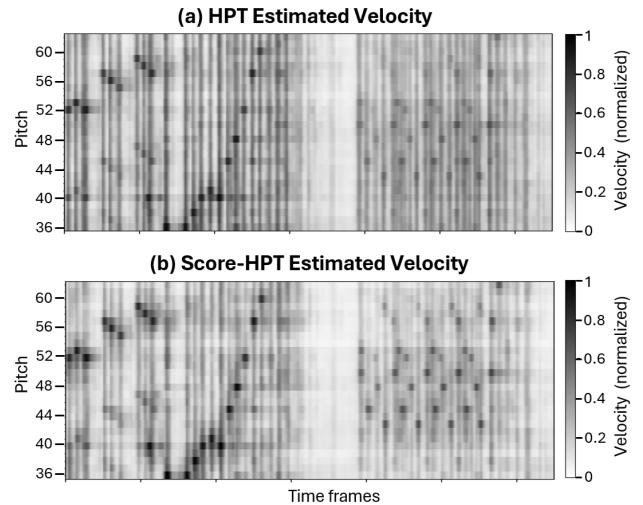


Fig. 3: Visualization of velocity estimates from the baseline HPT and the proposed Score-HPT. The refinement effect is evident, with Score-HPT producing clearer contrast and fewer spurious activations.

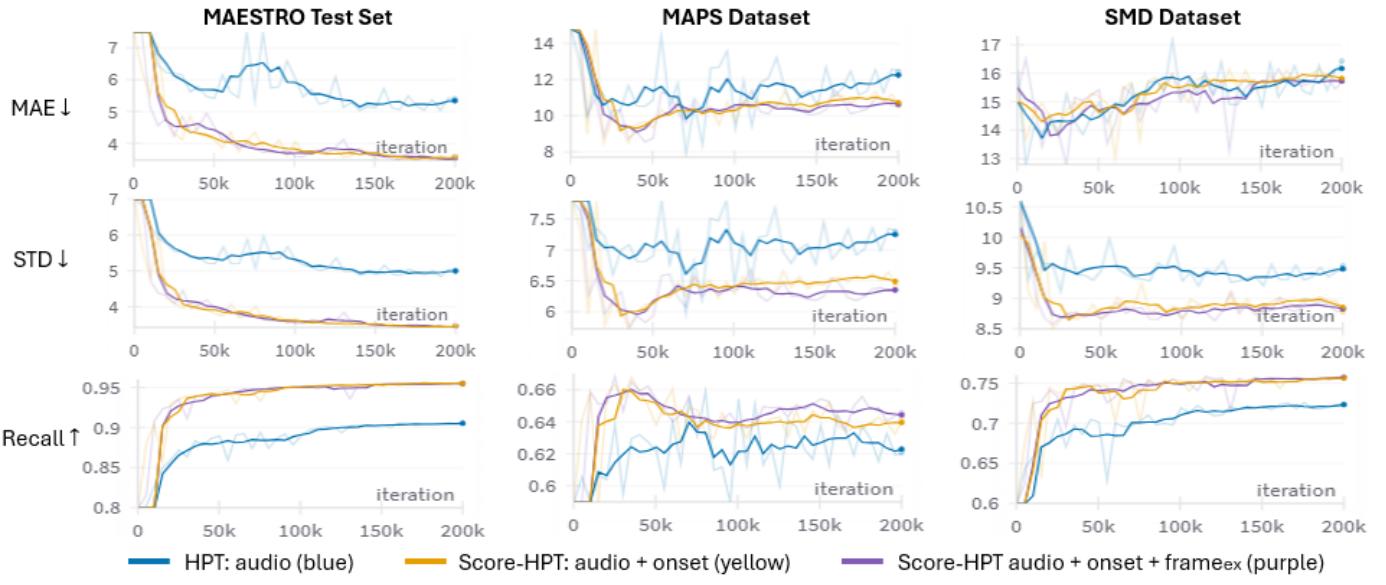


Fig. 4: Comparison of model performance over 200k training iterations for the HPT versus Score-HPT “audio + onset” and “audio + onset + frame_{ex}” configurations. Curves display metrics evaluated every 5k iterations, smoothed over a 20k iteration window for visualization. Training conducted exclusively on the MAESTRO train set, with evaluated on MAESTRO test set, MAPS and SMD datasets.

5.2. Generalization across Datasets

Figure 4 illustrates the enhanced generalization capabilities of our score-HPT variants, which consistently outperformed baseline HPT across the MAESTRO test set and out-of-distribution MAPS and SMD datasets after training on the MAESTRO. This underscores the benefit of score-informed correction, particularly for architectures like HPT known for generalization limitations [34], as the score offers robust structural and rhythmic context less affected by acoustic variations, leading to more reliable velocity estimations.

Performance trends also diverged over training (Fig. 4): while MAESTRO test performance steadily improved, out-of-distribution performance, especially on MAPS, peaked early and then declined. This suggests prolonged MAESTRO training caused overfitting to its acoustics properties, hindering generalization. The score-HPT’s superior performance on out-of-distribution data highlights the regularizing effect of score information in mitigating this overfitting.

5.3. Comparison with Existing Works

Table 2 compares Score-HPT performance on SMD dataset against existing methods, including both score-informed approaches and AMT systems that also provide velocity estimates. All models were trained exclusively on the MAESTRO train set. Consistent with the trends observed in Fig. 4, the Table 2 confirms that Score-HPT significantly outperforms the HPT baseline on the SMD dataset across all reported metrics (MAE, STD, and Recall), despite a minor trade-off between the variants.

Notably, our proposed Score-HPT does not surpass the current SOTA score-informed method, FiLM U-Net [16]. Our method’s ultimate performance ceiling is inherently dependent on the performance of the baseline AMT system whose velocity estimates it refines. Due to limited resources, we were unable to build our approach on the SOTA AMT system, hFT-Transformer [31], whose performance rivals FiLM U-Net but requires an Nvidia A100 80GiB GPU for training. Instead, we validated that our correction module significantly boosts the performance of HPT, which itself is a representative and widely used AMT system.

Table 2: Comparison of different methods on the SMD dataset. ↑ and ↓ indicate whether higher or lower values are better.

Model & Inputs	SMD dataset			
	MAE ↓	STD ↓	Recall ↑	
Score-informed Methods				
DiffVel [15]	multi.	19.7	13.1	53.0%
FiLM Conv [12]	multi.	15.1	12.3	85.8%
FiLM U-Net [16]	multi.	9.9	7.8	89.7%
Score-HPT (ours)				
audio + onset	multi.	13.0	9.0	70.3%
audio + onset & frame _{ex}	multi.	13.5	8.8	72.6%
AMT Systems				
HPT [17]	audio	13.9	9.4	68.7%
hFT-Transformer [31]	audio	9.9	7.3	78.0%

6. CONCLUSION AND FUTURE WORKS

In this work, we proposed a novel approach utilizing a score-informed BiLSTM correction module. Unlike previous score-informed methods that aim to replace the AMT system’s velocity estimation, our approach focuses on refining AMT-estimated velocity. By integrating this module with the HPT system, our Score-HPT demonstrated significant performance improvements on the MAESTRO dataset compared to its baseline. Furthermore, it exhibited strong generalization capabilities, maintaining superior performance over the baseline when evaluated on the out-of-distribution MAPS and SMD datasets, effectively mitigating some of HPT’s known generalization weaknesses by leveraging the structural information from the MIDI score.

This work has several limitations. Firstly, the method was not applied to a SOTA AMT system. Secondly, the score-informed method assumes a MIDI score perfectly time-aligned with the audio, which may not be achievable in practice. These limitations highlight future work: exploring the application of the correction module to more advanced AMT systems, and investigating the robustness of our approach with imperfectly aligned scores as in [16].

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 *Proc. of the 22nd Int. Society for Music Information Retrieval Conf. (ISMIR)*, 2021.
- [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 *Proc. of the 23rd Int. Society for Music Information Retrieval Conf. (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] X. Riley, Z. Guo, D. Edwards, and S. Dixon, “GAPS: a large and diverse classical guitar dataset and benchmark transcription model,” in *Proc. of the 25th Int. Society for Music Information Retrieval Conf. (ISMIR)*, San Francisco, United States, 2024.
- [8] 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.
- [9] C. Raffel, “Learning-based methods for comparing sequences, with applications to audio-to-midi alignment and matching,” Ph.D. dissertation, Columbia University, 2016.
- [10] A. Morsi and X. Serra, “Bottlenecks and solutions for audio to score alignment research,” in *Proc. of the 23rd Int. Society for Music Information Retrieval Conf. (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 *Proc. of the 25th Int. Society for Music Information Retrieval Conf. (ISMIR)*, San Francisco, USA, 2024, pp. 120–127.
- [12] H. Kim, M. Miron, and X. Serra, “Score-informed midi velocity estimation for piano performance by film conditioning,” in *Proc. of the Sound and Music Computing Conf. (SMC)*, 2023, pp. 139–147.
- [13] S. Oore, I. Simon, S. Dieleman, D. Eck, and K. Simonyan, “This time with feeling: Learning expressive musical performance,” *Neural Computing and Applications*, vol. 32, pp. 955–967, 2020.
- [14] C. Bigras, V. Duda, and S. Hébert, “Sensory and affective dimensions in loudness perception: Insights from young adults,” *Hearing Research*, vol. 454, p. 109147, 2024.
- [15] H. Kim and X. Serra, “DiffVel: Note-level midi velocity estimation for piano performance by a double conditioned diffusion model,” in *Proc. of the Int. Symposium on Computer Music Multidisciplinary Research (CMMR)*, Tokyo, Japan, 2023, pp. 197–208.
- [16] ——, “A method for midi velocity estimation for piano performance by a U-net with attention and FiLM,” in *Proc. of the 25th Int. Society for Music Information Retrieval Conf. (ISMIR)*, San Francisco, USA, 2024, pp. 304–310.
- [17] 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.
- [18] R. B. Dannenberg, “The interpretation of midi velocity,” in *Proc. of the Int. Computer Music Conf. (ICMC)*, 2006, pp. 193–196.
- [19] W. Goebel, “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.
- [20] W. M. Szeto, K. H. Wong, and C. H. Wong, “Finding intensities and temporal characteristics in piano music,” in *Proc. of the Int. Conf. on Digital Audio Effects (DAFx)*, 2005, pp. 214–219.
- [21] S. Ewert and M. Müller, “Estimating note intensities in music recordings,” in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011, pp. 385–388.
- [22] S. van Herwaarden, M. Grachten, and W. B. de Haas, “Predicting expressive dynamics in piano performances using neural networks,” in *Proc. of the 15th Int. Society for Music Information Retrieval Conf. (ISMIR)*, 2014, pp. 47–52.
- [23] 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.
- [24] D. Jeong, T. Kwon, and J. Nam, “A timbre-based approach to estimate key velocity from polyphonic piano recordings,” in *Proc. of the 19th Int. Society for Music Information Retrieval Conf. (ISMIR)*, Paris, France, 2018, pp. 120–126.
- [25] 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 *Proc. of the 19th Int. Society for Music Information Retrieval Conf. (ISMIR)*, 2018, pp. 50–57.
- [26] C. Hawthorne, I. Simon, R. Swavely, E. Manilow, and J. Engel, “Sequence-to-sequence piano transcription with transformers,” in *Proc. of the 22nd Int. Society for Music Information Retrieval Conf. (ISMIR)*, 2021, pp. 246–253.
- [27] L. Ou, Z. Guo, E. Benetos, J. Han, and Y. Wang, “Exploring transformer’s potential on automatic piano transcription,” in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Process. (ICASSP)*, 2022, pp. 776–780.
- [28] 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*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34, 2021, pp. 20 583–20 595.
- [29] Y. Yan and Z. Duan, “Scoring time intervals using non-hierarchical transformer for automatic piano transcription,” in *Proc. of the 25th Int. Society for Music Inf. Retrieval Conf. (ISMIR)*, 2024, pp. 973–980.
- [30] Q. Wang, M. Liu, C. Bao, and M. Jia, “Harmonic-aware frequency and time attention for automatic piano transcription,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 3492–3506, 2024.
- [31] K. Toyama, T. Akama, Y. Ikemiya, Y. Takida, W.-H. Liao, and Y. Mitsufuji, “Automatic piano transcription with hierarchical frequency-time transformer,” in *Proc. of the 24th Int. Society for Music Information Retrieval Conf. (ISMIR)*, Hybrid Conference, 2023, pp. 215–222.
- [32] 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.
- [33] B. Maman and A. H. Bermano, “Unaligned supervision for automatic music transcription in the wild,” in *International Conference on Machine Learning*. PMLR, 2022, pp. 14 918–14 934.
- [34] 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.
- [35] W. Wei, P. Li, Y. Yu, and W. Li, “HPPNet: Modeling the Harmonic Structure and Pitch Invariance in Piano Transcription,” in *Proc. of the 23rd Int. Society for Music Information Retrieval Conf. (ISMIR)*, 2022, pp. 709–716.
- [36] X. Riley, D. Edwards, and S. Dixon, “High resolution guitar transcription via domain adaptation,” in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2024, pp. 1051–1055.
- [37] L. Lin, Q. Kong, J. Jiang, and G. Xia, “A unified model for zero-shot music source separation, transcription and synthesis,” in *Proc. of the 22nd Int. Society for Music Information Retrieval Conf. (ISMIR)*, 2021, pp. 381–388.
- [38] K. W. Cheuk, K. Choi, Q. Kong, B. Li, M. Won, J.-C. Wang, Y.-N. Hung, and D. Herremans, “Jointist: Simultaneous improvement of multi-instrument transcription and music source separation via joint training,” *arXiv preprint arXiv:2302.00286*, 2023. [Online]. Available: <https://arxiv.org/abs/2302.00286>
- [39] 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 (ISMIR)*, 2011.
- [40] 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.
- [41] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, and D. P. W. Ellis, “mir_eval: A transparent implementation of common MIR metrics,” in *Proc. of the 15th Int. Society for Music Information Retrieval Conf. (ISMIR)*, Taipei, Taiwan, 2014, pp. 367–372.