A METHOD FOR MIDI VELOCITY ESTIMATION FOR PIANO PERFORMANCE BY A U-NET WITH ATTENTION AND FILM

Hyon Kim
Universitat Pompeu Fabra
hyon.kim@upf.edu

Xavier Serra
Universitat Pompeu Fabra
xavier.serra@upf.edu

ABSTRACT

It is a well known fact that the dynamics in piano performance gives significant effect in expressiveness. Taking the polyphonic nature of the instrument into account, analysing information to form dynamics for each performed note has significant meaning to understand piano performance in a quantitative way. It is also a key element in an education context for piano learners.

In this study, we developed a model for estimating MIDI velocity for each note, as one of indicators to represent loudness, with a condition of score assuming educational use case, by a Deep Neural Network (DNN) utilizing a U-Net with Scaled Dot-Product Attention (Attention) and Feature-wise Linear Modulation (FiLM) conditioning. As a result, we prove that effectiveness of Attention and FiLM conditioning, improved estimation accuracy and achieved the best result among previous researches using DNNs and showed its robustness across the various domain of test data.

1. INTRODUCTION

In the realm of piano performance, the loudness of each note plays a pivotal role, alongside other factors such as tempo and precise keystrokes [1]. When analyzing piano performances, the loudness of each note is quantitatively represented by MIDI velocity. Given the polyphonic nature of the piano, measuring the overall loudness within a specific timeframe fails to provide meaningful insights into the performance's quality. Loudness can be observed at various granularities, ranging from note-level loudness and frame-level aggregated loudness to the transcription of symbolic loudness representations. Each note in a piano performance can exhibit varying loudness levels, contingent on the music's texture [2, 3]. The unique loudness of each note, especially in the context of the piano's polyphonic attributes, holds significant meaning. Mastery over the loudness of individual notes is paramount, particularly in educational settings. To hone this control, score information serves as an essential benchmark. Visualization further enhances this educational endeavor [4]. Consequently, this study operates under the assumption that score information is accessible.

To ensure clarity in our terminology, we define "loudness" as the aggregated MIDI velocities within a designated timeframe, as gauged by an electronic piano device. In contrast, "intensity" refers to the peak value of the frequency sum for a note frame, as delineated in [5]. It is imperative to recognize that MIDI velocity does not directly correspond to the loudness as perceived by the human auditory system. Previous research has probed the relationship between MIDI velocity and perceived loudness in decibels (dB) [6,7]. These investigations consistently reveal a non-linear relationship, where an increase in MIDI velocity corresponds to a rise in perceived loudness.

Furthermore, studies such as those by [8, 9] have explored the mapping from perceptual loudness values in dB scale to dynamic symbols in piano performance, including symbols like $forte$, $mezzoforte$, $piano$, $pianissimo$, $crescendo$, and so forth. The dynamics and expressiveness of a musical composition are shaped by the loudness values attributed to each note in the score [1]. Notably, MIDI velocity offers a more nuanced prediction of loudness compared to traditional dynamic markings found in most music scores. These markings provide relative directives on the loudness with which a piece should be played. The loudness of individual notes in a piano performance can fluctuate based on the texture of the music [2,3]. Given the polyphonic characteristics of piano performances, note-level loudness is of paramount importance.

Recognizing the significance of delving into note-level loudness granularity, this study primarily centers on MIDI velocity estimation, particularly within an educational context where score information is presumed available.

2. RELATED WORK

In this section, we delve into pertinent works within the domain of Machine Learning methods and their applications for the task.

Note Level Intensity Estimation: The task of note-level loudness estimation has been the focus of multiple studies [5, 10–13]. These investigations have utilized both Non-Negative Matrix Factorization (NMF) and DNN methodologies to segregate piano performance audio into 88 distinct keys, subsequently estimating MIDI velocity or intensity for each note. This research domain can be

viewed as an extension of Automatic Music Transcription (AMT) and Music Performance Assessment, with potential applications in modeling performance expressiveness. The task of piano note-level MIDI velocity estimation is multifaceted, encompassing both a regression problem, where MIDI velocity values within the 0-127 range are estimated, and an audio classification challenge, which categorizes audio into one of the typical 88 piano keys. A limited number of studies have tackled the note-level MIDI velocity estimation task for an actual piano performance data, employing techniques like NMF [5] and DNN methods [13, 14]. The study by [5] integrated with score information to estimate note-level intensity, subsequently developing a linear regression model for note-level MIDI velocity estimation. The DNN methods [13] have sought to address the estimation challenge by incorporating AMT techniques and score conditioning. These DNN architectures amalgamate convolution blocks and GRU blocks, introducing FiLM conditioning generated by a fully connected linear layer. A diffusion model together with FiLM conditioning [14] inserts a score and performance audio information for its generative task to express note frames with MIDI velocity information. While the DNN approach did not outperform the NMF method, it marked a pioneering effort to estimate MIDI velocity using DNNs, aiming to create a model that could generalize to unseen classical music inputs, in contrast to the NMF method that optimizes parameters for individual test data. In our research, we juxtapose our findings with these preceding studies.

U-Net: The U-Net architecture incorporates layered residual connections. The concept of a residual network emerged as a solution to counteract the vanishing or exploding gradient issues encountered during the DNN training phase. U-Net has been employed for piano performance transcription, specifically for reconstructing spectrograms [15]. Its efficacy in music source separation tasks within the field of music information retrieval is well-documented. Notably, research has been conducted on a FiLM-conditioned U-Net for music source separation [16]. In our study, we leverage a U-Net structure with convolutional layers to process the mel spectrogram, a two-dimensional representation of audio. We anticipate that the U-Net will enhance classification accuracy, converting audio to the 88 piano keys.

Feature-wise Linear Modulation (FiLM): Our study employs FiLM conditioning to integrate score information, aiming to estimate note-level MIDI velocity for piano performances [17]. Historically, FiLM conditioning has found applications in image processing, yielding enhanced results when conditioned with natural language for tasks like object detection [17]. This concept has been extended to audio source separation tasks, where audio is conditioned with supplementary information such as video and scores [18]. Structurally, FiLM encompasses neural network layers that produce an affine transformation for a specified input layer. It integrates a base DNN, trained in a supervised manner, with a condition generator. This generator processes conditions, such as scores, to produce the parameters β and γ for an element-wise affine transformation in the latent space of the base DNN. Mathematically, this is represented as: $FiLM(\boldsymbol{x}) = \gamma(\boldsymbol{z}) \cdot \boldsymbol{x} + \beta(\boldsymbol{z})$. Here, the vector $\boldsymbol{z}$ serves as the conditional vector. Figure 1 visually represents the FiLM conditioning architecture, illustrating how the condition embedding model generates the parameters β and γ for the affine transformation on the latent vector $\boldsymbol{x}$ derived from the base DNN.

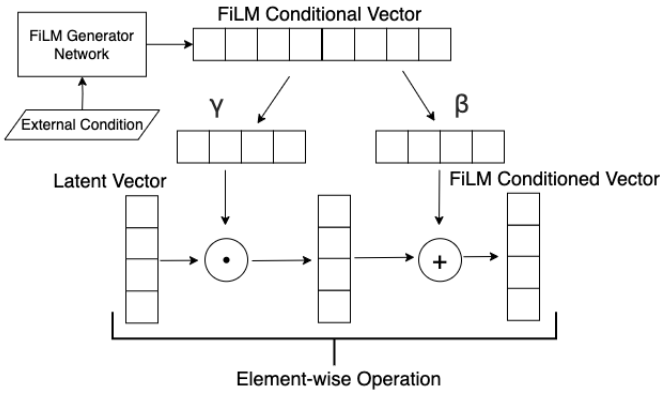

Figure 1. Visualization of FiLM operation

The Scaled Dot-Product Attention (Attention): The Attention, introduced by [19], has been instrumental in advancing the field of deep learning. This mechanism computes attention weights by scaling the dot products of queries and keys, which facilitates a dynamic focusing of the model on relevant parts of the input data. Its efficiency and simplicity allow for significant improvements in model performance by enabling the capture of long-range dependencies within the data, without the constraints imposed by previous sequence processing models. The architecture is utilised in an image processing area [20] and a speech processing area [21] together with U-Nets. This mechanism has also been applied to music information retrieval such as source separation [22] and showed its performance together with computational efficiency for the task. These researches show that the Attention mechanism works for capturing its target information from complex input data.

Our model incorporates this Attention within the U-Net architecture to leverage its proven benefits, thereby enhancing our model's ability to understand and generate nuanced responses based on the context provided by the input sequence in a musical sense.

3. METHOD

Figure 2 illustrates the comprehensive architecture of our proposed model. Initially, the model processes audio input, transforming it into a Log Mel-frequency Spectrogram. This transformation facilitates the conversion of the waveform into an image-like format. The audio processing parameters include a window length of two seconds, a hop size of one second, and a sampling rate of 16k Hz, resulting in a model output resolution of 100 frames per second. The overarching model architecture can be cate-

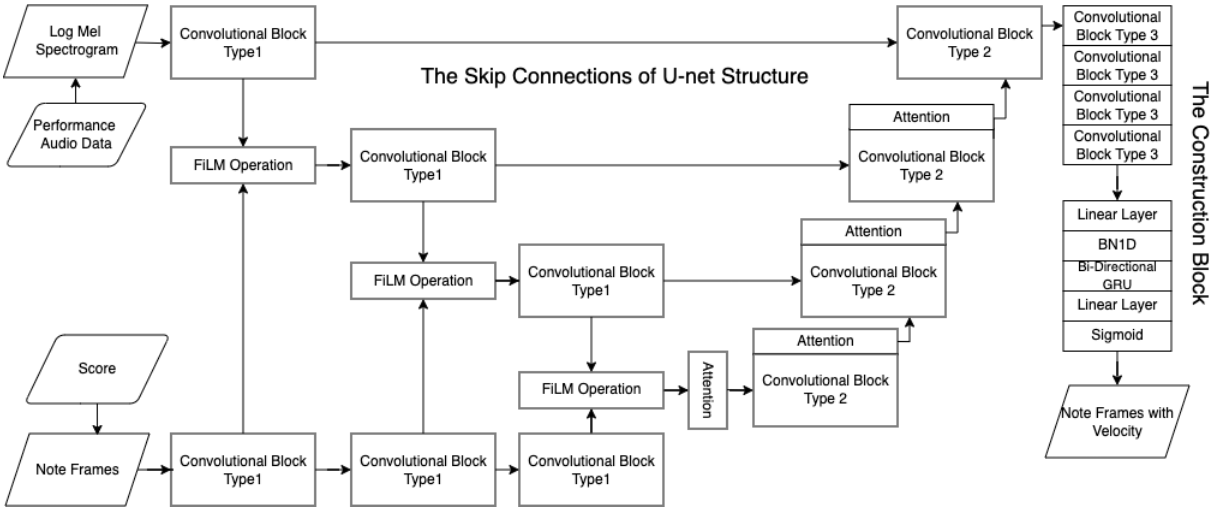

Figure 2. The entire architecture of the proposed model.

gorized into three distinct convolutional blocks, as showcased in Figure 3.

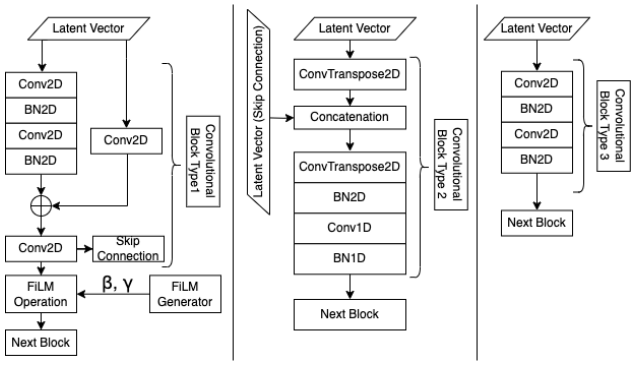

Figure 3. Schematic of the three convolutional blocks utilized in the model.

Convolutional blocks of type 1 and 2 collectively form the U-Net structure. Type 1 blocks also play a pivotal role in encoding note frame information. In this study, note frames are derived from a MIDI roll. Corresponding blocks in the encoding phase generate FiLM conditioning parameters, denoted as β and γ, for each affine transformation. Several methods for inserting FiLM parameters are described in [16]. In our model, through empirical study, we generate parameters to ensure element-wise correspondence for each latent space vector, as depicted in Figure 1. Consequently, each output from the score encoders generates twice as many parameters for the output of each block in the encoder of the U-Net.

To ensure uniformity in the processed latent features, we employ convolutional layers of the same hierarchical level to produce FiLM parameters for each layer within the U-Net. For the skip connections, non-conditioned latent vectors from each block are relayed to the corresponding type 2 block, while FiLM-conditioned latent vectors are channeled to the subsequent layer of the type 1 block.

In the decoder section of the U-Net architecture, Attention modules are incorporated before each convolutional block type 2. This configuration enhances the network's ability to focus on relevant features by dynamically adjusting the importance of different areas of the input image. The Attention mechanism, which calculates attention scores by scaling the dot-product of queries and keys, enables the model to prioritize specific features over others, improving the precision of MIDI velocity estimation.

The construction block consists of convolutional block type 3 followed by the block containing bi-directional GRU. It processes inputs through a sequence of layers including linear transformations for dimensionality reduction based on the input feature type, batch normalization, and a bidirectional GRU for capturing temporal dynamics. The network concludes with a fully connected layer applying a sigmoid function to output note frames with velocity information. Dropout and ReLU activations are utilized throughout to enhance performance and prevent overfitting.

For training our model, we employed the MAESTRO dataset [23]. MAESTRO is a dataset composed of about 200 hours of virtuosic piano performances captured with fine alignment (up to 3 ms) between note labels and audio waveforms. Notably, other DNN models targeting MIDI velocity estimation, such as [13] and [14], have also employed this dataset. This usage facilitates a more legitimate comparison of model performance across different studies.

Our chosen loss function, represented by Eq. 1, amalgamates the $l1$ loss and the Binary Cross-Entropy (BCE) loss. This design facilitates back-propagation of losses for both classification and regression tasks.

$$Loss = \theta \cdot l1 \text{ loss} + (1 - \theta) \cdot \text{BCE loss} \tag{1}$$

Here, $\theta \in [0, 1]$ signifies the weight for the $l1$ and the BCE loss. For our empirical setup, we set θ to 0.5. The $l1$ loss function, as defined in Eq. 2, is articulated as:

$$l1 \text{ loss } = \frac{\sum_i |V(i)_{\text{ground truth}} - V(i)_{\text{model output}}|}{N} \tag{2}$$

In this equation, $V(i)$ represents MIDI velocity with the index i of corresponding notes between the ground truth and the model output within a specified window, while N denotes the total number of notes present in that window. Each input data point spans two seconds, with each frame encompassing 100 segments per second to depict the MIDI roll. The velocities used to compute the loss are normalized to the range $[0, 1]$ to match the scale of the BCE.

For the evaluation phase, we employed the Saarland Music Data (SMD) dataset [24]. SMD provides audio recordings along with perfectly synchronized MIDI files for various piano pieces. The pieces were performed by students of the Hochschule für Musik Saar on a hybrid acoustic/digital piano (Yamaha Disklavier). We selected 49 excerpts from this dataset, consistent with the test sets used in prior studies [5, 13, 14], ensuring a fair and comparable assessment. The model's error is quantified using the formula presented in Eq. 3:

$$Error = \frac{\sum_i |V(i)_{\text{ground truth}} - V(i)_{\text{inference}}|}{N} \quad (3)$$

In this equation, i represents individual notes, and N denotes the total number of notes accurately identified in the score. The inferred MIDI velocity is determined by the peak value within the interval of each detected and categorized velocity frame, juxtaposed with the ground truth velocity frame for the respective note. This approach is adopted because the detected velocity typically exhibits a peak followed by a decline in the estimated MIDI velocity within a note frame, mirroring the attack and decay patterns of each note's loudness. Differently from loss function, the output values are not normalised but are scaled to the range $[0, 127]$. The recall score serves as our primary evaluation metric for classification accuracy, given that the model's output is constrained by the provided score information.

4. RESULTS AND DISCUSSION

Result and Comparison: Table 1 presents the comparative outcomes of our model against previous works in the field. The proposed model consistently outperforms other DNN-based methods across all metrics, demonstrating notable improvements. The enhancements are particularly evident when comparing the best and worst outcomes of our model with those of other models. The results highlight that the U-Net designed with Attention and FiLM conditioning with score information significantly boosts performance.

Among the test set, the most favorable outcome is observed for "Bach BWV875-01 002," which recorded mean error, standard deviation, and recall values of 4.6, 3.3, and 95.6%, respectively. Conversely, "Chopin Op028-17" exhibited the least favorable results for mean error and standard deviation, with values of 16.0 and 11.9 respectively, and a recall of 87.5%. Additionally, "Ravel Jeux d'eau" demonstrated the lowest recall score in the dataset

Model	Mean	SD	Recall
DNN Based Model			
DiffVel [14]	19.7	13.1	53.0%
Convolutional Net [13]	15.1	12.3	85.8%
Proposed Model	**9.9**	**7.8**	**89.7%**
NMF Based Model			
Score-Informed NMF [5]	4.1	5.0	N.A.

Table 1. Comparative results of models for note-level MIDI velocity estimation with score information. SD: Standard Deviation

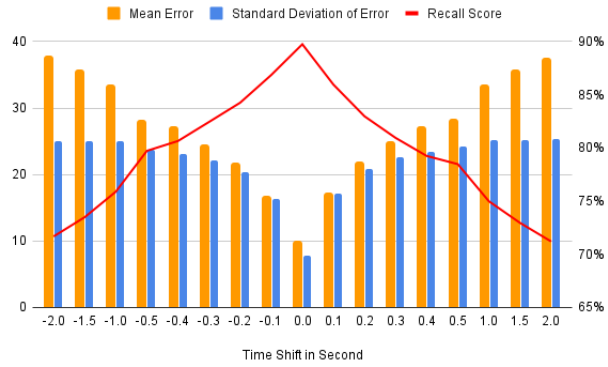

Figure 4. Mean and SD of errors for misaligned score information

at 80.7%, with corresponding mean error and standard deviation values of 12.1 and 10.2. These results illustrate the varied performance of our model across different musical pieces, underscoring its effectiveness as well as areas for potential improvement.

The analysis also highlights the strengths and weaknesses of both DNN and NMF-based methods. DNNs are capable of capturing complex relationships within the training data due to their nonlinear nature, but they are computationally demanding and require extensive data to optimize parameters effectively. In contrast, NMF-based methods, such as the one described by [5], optimize parameters for individual excerpts using score information in the test set, offering a more tailored approach. This specificity, however, can limit their generalizability compared to DNNs, which aim to develop a more generic model suitable for diverse musical excerpts. Notably, the proposed model is trained on a distinct domain, specifically a piano performance dataset different from the test set, to ensure a fair comparison and robust assessment of its performance. This strategy helps in evaluating the model's ability to generalize across different musical contexts effectively.

Misaligned Condition Insertion: In real-world applications, alignment discrepancies frequently occur between scores and their corresponding audio, affecting the accurate feeding of note frames. Figure 4 elucidates the model's sensitivity to temporal misalignments, exhibiting a correlation between the degree of time shift and the model's performance metrics.

These shifts are synthetically generated by inserting the

conditions a specified number of seconds ahead or behind each input frame, with a two-second duration per input. It is clear that misaligned data affects to the model accuracy proportionally. Addressing this misalignment, data augmentation can be employed during the training phase to acclimate the model to varying degrees of data condition misalignments, thereby enhancing its flexibility.

Nonetheless, this alignment challenge may become negligible with the integration of a dedicated note frame detection model, as delineated by models like the one in [25]. Utilizing such models for precise note frame detection allows for a subsequent, more accurate analysis of MIDI velocity estimation by the proposed model, streamlining the workflow and potentially increasing performance accuracy.

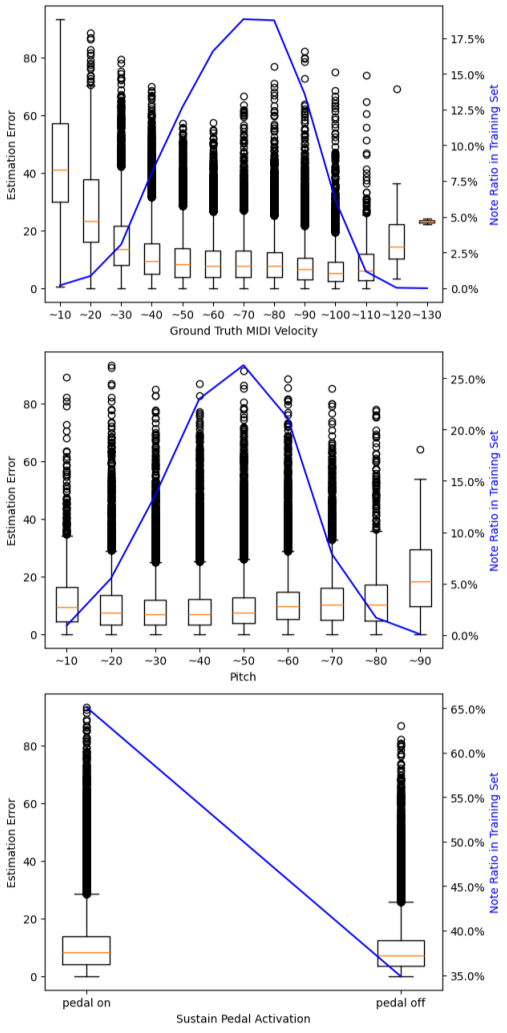

Figure 5. Error distributions based on various ground truth aspects: pitch, sustain pedal activation, and MIDI velocity intervals together with the ratio of notes appeared in the training set.

Error Analysis: Further analysis was conducted to evaluate the error distribution across different pitch groups, ground truth MIDI velocities, and sustain pedal activation states, as depicted in Figure 5. The analysis indicates that error is inversely correlated with the volume of data in the training set: the greater the quantity of data processed by the model, the more accurate the MIDI velocity estimates, highlighting the benefits of extensive data representation.

The results further reveal that enhanced training data volumes lead to improved estimation outcomes across various data dimensions. This suggests that applying data augmentation to achieve a balanced distribution in pitch and velocity bins can result in higher estimation accuracy. However, such augmentation must maintain the musicological context, including harmony and expressiveness, making this a complex yet critical task for effective model training.

Ablation Study: In our ablation study, we evaluate the individual and combined contributions of FiLM conditioning and the Attention modules to our model's performance, based on the U-Net architecture. These components were chosen for their theoretical abilities to enhance feature representation and focusing mechanisms, respectively. The study aims to clarify their roles within our proposed deep neural network architecture. We examine four configurations of our model: (i) with both FiLM and Attention (proposed model), (ii) with FiLM but without Attention, (iii) with Attention but without FiLM, and (iv) without either FiLM or Attention, as shown in Table 2.

Model Configuration	Mean	SD	Recall
With FiLM:			
With Attention	**9.9**	**7.8**	**89.7%**
Without Attention	10.0	7.8	89.4%
Without FiLM:			
With Attention	12.1	10.5	73.0%
Without Attention	13.0	10.5	68.5%

Table 2. Ablation Study: Detailed Performance Comparison Highlighting the Impact of FiLM Conditioning and Attention.

The ablation study highlights the significant impact of FiLM Conditioning and a relatively lesser contribution from the Attention in enhancing the performance of the proposed model. The observed synergy when integrating these modules indicates a promising avenue for future research and development in deep neural network architectures. While the Attention module improves model performance, its effectiveness is not as pronounced as that of FiLM Conditioning. This suggests that the model's ability to concentrate on relevant features, and thereby its predictive performance, is significantly enhanced by FiLM Conditioning. Notably, we could see universal improvement on the recall score on all the excepts on the test set in any comparison among combination of (i) to (iv).

According to Table 1, the study also demonstrates that incorporating a U-Net mechanism, particularly its skip connections, can enhance accuracy for the task at hand, outperforming previous models.

Robustness of the Model: We conducted a comparative analysis against the state-of-the-art transcription model that additionally estimates the MIDI velocity [26]. The results, detailed in Table 3, indicate that our proposed

model achieves comparable performance in MIDI velocity estimation. Notably, our model demonstrates enhanced robustness across various test datasets, as evidenced by the recall scores, in comparison to the model proposed by [26] which is also trained on the MAESTRO dataset.

Model	Mean	SD	Recall
Proposed Model	9.9	7.8	**89.7%**
The hFT Model [26]	9.9	**7.3**	78.0%

Table 3. Comparison to the SOTA Transcription Model

This comparison highlights the efficacy of our model, particularly in its ability to generalize across different datasets, which is crucial for practical applications. The fact that both models yield identical mean scores for MIDI velocity estimation but our model exhibits a higher recall rate suggests our model's capability in accurately capturing the nuances of musical expression. Furthermore, despite the slightly higher standard deviation in our model's performance, the significantly higher recall rate underscores its robustness and reliability in diverse testing scenarios. This finding is particularly relevant for applications requiring high fidelity in musical transcription and velocity estimation, indicating a promising direction for future research and development in music transcription technologies. Also the ablation study indicates that adding FiLM conditioning can improve the model accuracy for the task, after experimental process of designing the parameter generators and methods to insert the parameters, which yields another research topic.

5. CONCLUSION AND FUTURE WORKS

In this study, we explored the complexities of MIDI velocity estimation, leveraging an U-Net architecture enriched with the Attention and FiLM conditioning to integrate score information. Our results underscore the superiority of this approach among DNN methodologies. Our empirical evaluations further attest to the pivotal role of FiLM conditioning in bolstering result accuracy. This enhancement transcends specific model architectures, with FiLM conditioning amplifying precision across various models, ranging from feed-forward designs with convolution and GRU blocks to diffusion models, combining the previous researches [13, 14]. The Attention also contributes to improve on both MIDI velocity estimation and recall score. The model also showed that comparable results towards SoTA transcription model and the robustness across the sources of test set compared to other state of the art transcription models which also estimates MIDI velocity.

Generally, FiLM conditioning has proven effective for MIDI velocity estimation tasks. Enhanced transcription of note onset, offset, and frames could further refine performance, positioning this model as a robust solution for MIDI velocity estimation across diverse datasets. This suggests that utilizing DNN models, such as the onsets and frames model proposed by [25], which demonstrates superior accuracy in note frame detection without FiLM conditioning, could be advantageous. In situations where score data are not available, a cascaded approach can be employed: first, use a DNN for accurate note frame detection, and then leverage the detected MIDI for FiLM conditioning, circumventing the need for score-to-audio alignment.

As we look to the future, our objective is to expand the range of score information, transitioning from MIDI note frame to more comprehensive formats, MusicXML to be encoded. Such a shift is anticipated to offer increased resilience, especially in situations where achieving precise alignments poses challenges. Data augmentation on training data is also considered as crucial task for obtaining more robust estimation, as mentioned. Additionally, addressing issues such as omitted notes and extraneous notes is essential to tailor the model more effectively for educational applications, catering to both novice learners and seasoned professionals, considering currently the set up only considers the student is god enough to follow the score for performance visualization purposes.

The potential applications of this research are manifold, extending from the development of visualization tools that bolster musical communication to advanced transcription techniques. Such annotations, especially those denoting expressiveness, carry significant implications, particularly in pedagogical contexts where teacher-student interactions are crucial.

The code and model developed for this study are available upon request.

6. ACKNOWLEDGMENTS

"IA y Música: Cátedra en Inteligencia Artificial y Música" (TSI-100929-2023-1), funded by the Secretaría de Estado de Digitalización e Inteligencia Artificial, and the European Union-Next Generation EU, under the program Cátedras ENIA 2022 para la creación de cátedras universidad-empresa en IA.

7. REFERENCES

[1] M. Grachten and G. Widmer, "Linear basis models for prediction and analysis of musical expression," *Journal of New Music Research*, vol. 41, no. 4, pp. 311–322, 2012.

[2] W. Goebl, "Melody lead in piano performance: expressive device or artifact?" *The Journal of the Acoustical Society of America*, vol. 110, no. 1, pp. 563–72, 2001.

[3] S. Kim, J. M. Park, S. Rhyu, J. Nam, and K. Lee, "Quantitative analysis of piano performance proficiency focusing on difference between hands," *PLoS ONE*, vol. 16, 2021.

[4] L. F. Hamond, G. Welch, and E. Himonides, "The pedagogical use of visual feedback for enhancing dynamics in higher education piano learning and performance," *Opus*, vol. 25, no. 3, pp. 581–601, 2019.

[5] D. Jeong, T. Kwon, and J. Nam, "Note-intensity estimation of piano recordings using coarsely aligned MIDI score," vol. 68, no. 1, pp. 34–47, 2020, publisher: Audio Engineering Society. [Online]. Available: https://www.aes.org/e-lib/browse.cfm?elib=20716

[6] R. B. Dannenberg, "The interpretation of midi velocity," in *International Conference on Mathematics and Computing*, 2006.

[7] Y. Qu, Y. Qin, L. Chao, H. Qian, Z. Wang, and G. Xia, "Modeling perceptual loudness of piano tone: Theory and applications," *arXiv preprint arXiv:2209.10674*, 2022.

[8] K. Kosta, O. F. Bandtlow, and E. Chew, "Outliers in performed loudness transitions: An analysis of chopin mazurka recordings." in *International Conference for Music Perception and Cognition (ICMPC)*, California, USA, 2016, pp. 601–604.

[9] K. Kosta, R. Ramírez, O. F. Bandtlow, and E. Chew, "Mapping between dynamic markings and performed loudness: a machine learning approach," *Journal of Mathematics and Music*, vol. 10, no. 2, pp. 149–172, 2016.

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

[11] J. Devaney and M. Mandel, "An evaluation of score-informed methods for estimating fundamental frequency and power from polyphonic audio," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 181–185. [Online]. Available: http://ieeexplore.ieee.org/document/7952142/

[12] D. Jeong and J. Nam, "Note intensity estimation of piano recordings by score-informed nmf," in *Audio Engineering Society Conference: 2017 AES International Conference on Semantic Audio*. Audio Engineering Society, 2017.

[13] H. Kim., M. Miron, and X. Serra, "Score-informed midi velocity estimation for piano performance by film conditioning," in *Proc. Int. Conf. Sound and Music Computing*, 2023.

[14] H. Kim and X. Serra, "Diffvel: Note-level midi velocity estimation for piano performance by a double conditioned diffusion model," in *Proceedings of the 16th International Symposium on Computer Music Multidisciplinary Research*, Tokyo, Japan, 2023. [Online]. Available: http://hdl.handle.net/10230/57790

[15] K. W. Cheuk, Y.-J. Luo, E. Benetos, and D. Herremans, "The effect of spectrogram reconstruction on automatic music transcription: An alternative approach to improve transcription accuracy," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 9091–9098.

[16] G. Meseguer-Brocal and G. Peeters, "Conditioned-u-net: Introducing a control mechanism in the u-net for multiple source separations," *arXiv preprint arXiv:1907.01277*, 2019.

[17] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[18] O. Slizovskaia, G. Haro, and E. Gómez, "Conditioned source separation for musical instrument performances," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2083–2095, 2021.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023.

[20] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention u-net: Learning where to look for the pancreas," 2018.

[21] R. Giri, U. Isik, and A. Krishnaswamy, "Attention wave-u-net for speech enhancement," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 249–253.

[22] T. Sgouros, A. Bousis, and N. Mitianoudis, "An efficient short-time discrete cosine transform and attentive multiresunet framework for music source separation," *IEEE Access*, vol. 10, pp. 119 448–119 459, 2022.

[23] 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 *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=r1lYRjC9F7

[24] M. Müller, V. Konz, W. Bogler, and V. Arifi-Müller, "Saarland music data (smd)," in *Proceedings of the international society for music information retrieval conference (ISMIR): late breaking session*, 2011.

[25] C. Hawthorne, E. Elsen, J. Song, A. Roberts, I. Simon, C. Raffel, J. Engel, S. Oore, and D. Eck, "Onsets and frames: Dual-objective piano transcription," 2018.

[26] 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 24th International Society for Music Information Retrieval Conference*, 2023.