DiffVel: Note-Level MIDI Velocity Estimation for Piano Performance by A Double Conditioned Diffusion Model

Hyon Kim[1] and Xavier Serra[1] *

Music Technology Group, Universitat Pompeu Fabra
hyon.kim@upf.edu, xavier.serra@upf.edu

Abstract. In any piano performance, expressiveness is paramount for effectively conveying the intent of the performer, and one of the most significant aspects of expressiveness is the loudness at the individual key or note level. However, accurately detecting note-level loudness poses a considerable technical challenge due to the polyphonic nature of piano performances, wherein multiple notes are played simultaneously, as well as the similarity of harmonic elements.
MIDI velocity is crucial for indicating loudness in piano notes. This study conducted experiments for estimating a note-level MIDI velocity expanding the DiffRoll model: the Diffusion Model for piano performance transcription. By adopting double conditioning—audio and score information—and implementing noise removal as a post-processing, our findings highlight the model's potential in estimating note level MIDI velocity.

Keywords: MIDI Velocity Estimation, Diffusion Model, Conditioned Deep Neural Network, FiLM Conditioning

1 Introduction

The assessment of piano performance can be attributed to three key factors, namely loudness, rhythm, and accurate key strokes [1]. Owing to the polyphonic nature of piano performances, multiple auditory streams coexist, such as melody line and accompaniment. This intricate aspect allows for enhanced distinguishability in the interpretations of expert pianists [2]. The expressiveness of a musical piece is significantly influenced by the series of loudness values associated with each note in the score, which contribute to the dynamic alterations throughout the composition [3].

Within the realm of music education, research has demonstrated the effectiveness of utilizing visual feedback in enhancing students' abilities [4, 5]. In this regard, the comprehension and management of loudness become especially significant [1] when

* This research was carried out under the project Musical AI - PID2019- 111403GB-I00/AEI/10.13039/501100011033, funded by the Spanish Ministerio de Ciencia e Innovación and the Agencia Estatal de Investigación.

it is visualised. The employment of techniques for estimating and visualizing loudness fulfills the crucial system requirement, enabling the provision of valuable feedback to learners.

[8,9,12] researched mapping from perceptual loudness value in dB scale to dynamic symbols for piano performance such as $forte$, $mezzoforte$, $piano$, $pianissimo$, $crescendo$, etc. Note that using MIDI velocity, we predict loudness at a lower granularity, i.e. finer scale than the dynamic markings which are explicitly written in most of music scores and indicate how loud the piece should be played. Furthermore, each note in a piano performance may have a different loudness depending on the texture of the music [11, 14]. Therefore, the note-level loudness itself has special meaning in piano performance, considering its polyphonic characteristics.

To prevent ambiguity, we use the term "loudness" to denote the combined MIDI velocities within a specific time frame as measured by an electronic piano device. On the other hand, "intensity" refers to the maximum value of the frequency sum for a note frame, as defined in [15]. It is essential to recognize that MIDI velocity does not have a direct correlation with loudness as experienced by the human auditory system. Studies have been conducted to explore the relationship between MIDI velocity and loudness measured in decibels (dB) [30]. While the research demonstrates a consistent increase in perceived loudness (in dB) with increasing MIDI velocity, it also reveals that this relationship is non-linear [13].

Since this research aims to detect MIDI velocity on each note performed, automatic piano performance transcription is a closely related area for this purpose. Piano performance transcription is also an actively researched topic [10, 23, 24]. However, these studies primarily focus on detecting individual notes, rather than note loudness or dynamic symbols in a score. Additionally, the transcription process is not yet fully accurate and reproducible of performance.

Several studies have explored the note-level loudness estimation task [6, 15–18]. These researchers employed NMF and DNN methods to isolate piano performance audio into 88 distinct keys and estimated MIDI velocity or intensity for each note. We consider this area of research as an application of Automatic Music Transcription (AMT) and to be applied to expressiveness performance modeling. The piano note-level MIDI velocity estimation task involves solving a two-fold problem. One aspect is a regression problem, requiring the estimation of numbers within the 0-127 range for MIDI velocity. The other issue is audio classification, which involves sorting audio into each piano key, typically consisting of 88 keys. To address these challenges, we propose the DiffVel as an expansion of DiffRoll [7], a diffusion model for AMT, and Feature-wise Linear Modulation (FiLM) conditioning layers [20] to incorporate score information into the DNN. We conducted experiments to estimate the MIDI velocity using this approach.

2 Related Work

2.1 Automatic Music Transcription

The piano performance transcription is one of the closest problems for classification from audio input. [29] proposed a CNN-GRU combined acoustic model which branches

into four outputs: velocity regression, onset, offset, and note frame estimation. The note frame estimation is the final goal of this model and the other three estimations are gathered as input to another acoustic model to estimate the notes at the frame level. Therefore, the estimated MIDI velocity regression is not evaluated in the paper since it is out of the scope.

Recently, diffusion models have been explored as an alternative approach. DiffWave, a state-of-the-art generative model for audio synthesis, leverages the diffusion probabilistic framework and exhibits remarkable capabilities in generating high-quality audio samples from various sources. The core idea behind DiffWave involves employing a series of de-noising score matching steps, iteratively refining the generated audio samples to achieve accurate and precise output. Building upon DiffWave, DiffRoll has been researched [7]. DiffRoll expands DiffWave into a two-dimensional representation of sound and output, taking Mel Spectrogram as a condition and forming the two-dimensional Gaussian noise input into MIDI roll. The generative model's characteristics offer considerable potential to simultaneously address classification and regression problems by tuning conditions. Exploring conditions with not only one but also multiple conditions would contribute to estimating MIDI velocity more accurately. However, the model disregards velocity estimation in the model evaluation.

For conditioning, existing research utilizes score information to inform musical instrument separation in polyphonic music [19, 21, 28]. These works employ score or video information to enhance source separation results by creating an additional neural network to extract features from the supplementary data, which are then fed into the original DNN.

2.2 Feature-wise Linear Modulation (FiLM)

In this paper, we utilized the FiLM conditioning [20] to insert score information in order to estimate note-level MIDI velocity for piano performance. FiLM conditioning is used in the image processing area and has gained improved results on object detection [20]. In previous research, natural language is used as an external condition to indicate the existence of target objects to be detected. This idea has been applied to audio source separation tasks by conditioning audio with video and score information [28].

The FiLM comprises a set of neural network layers that generate an affine transformation for a given input layer in a neural network. It consists of a base DNN which is trained in a supervised fashion and a condition generator which takes conditions such as score as input and generates β and γ to make an element-wise affine transformation in the latent space of the base DNN. In the math formula, it is described as follows;

$$FiLM(x) = \gamma(z) \cdot x + \beta(z) \tag{1}$$

where vector z is a conditional vector.

The Figure 1 shows the architecture of FiLM conditioning. This condition embedding model generates parameters, β and γ, to make an affine transformation on the latent vector x from the base DNN.

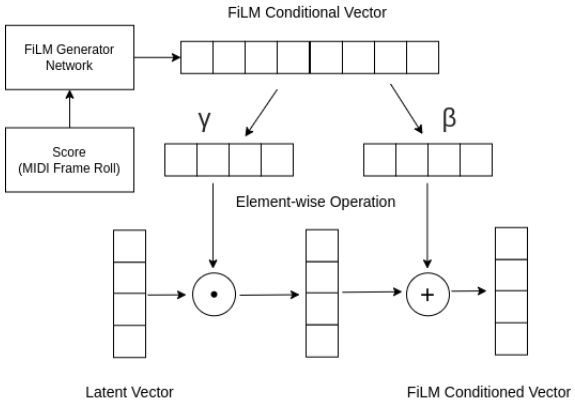

Fig. 1. The diagram illustrates the operation flow for inserting a FiLM condition into a latent vector

2.3 Note Level MIDI Velocity Estimation with Score Information

Only three papers have considered note-level MIDI velocity in music performance, employing NMF [15, 18] and DNN methods [6]. NMF methods have been used for source separation problems and effectively applied to music source separation as well [22]. [15] examined an NMF method with score information to estimate note-level intensity before creating a linear regression model to obtain note-level MIDI velocity estimation. This research provided a detailed analysis of NMF method errors and their causes. The DNN method attempted to address the estimation problem by applying the AMT method and score conditioning. The DNN architecture involves stacking convolution blocks and GRU block and inserting a FiLM conditioning generated by a fully connected linear layer. Although it did not surpass the results of the NMF method, it was the first attempt to estimate MIDI velocity using a DNN method and to generalize the model for unseen classical music inputs, as opposed to the NMF method which optimizes parameters for each test data. [16] aimed to estimate the note level intensity, rather than MIDI velocity, from the spectrogram by filtering it according to the frequency of each note.

In our study, we compare our results with the NMF method proposed in [15] and the DNN method [6] as our benchmark.

3 Method

There are two models experimented in this study: the diffusion models with and without score information by FiLM conditioning. The entire architecture is based on the DiffRoll model and the conditions, Mel Spectrogram and score, are inserted as an expansion. We used the MIDI velocity data on note frame level as supervised data for training for both models. Score information is represented in a note frame roll in the MIDI roll.

For the training data, we used the Maestro dataset [26]. The data segmentation is 20 seconds, and the number of data frames is 31 in one second. Therefore, each output from the models is a (620, 88) matrix containing onset, offset, and velocity information.

3.1 Model Architecture

The simplified overall model architecture is illustrated in Figure 2. In the diffusion model, each residual layer takes the conditions. The Mel Spectrogram transformed from input audio is added as another condition to each residual layer before the FiLM conditional vector insertion.

For the purpose of inserting the score information, we also added a FiLM conditioning layer as it is introduced in Section 2. We have tested the element-wise operations for multiplication and addition. However, the scalar multiplication and addition gave us better results. The FiLM generator is designed as a fully connected layer to generate conditioning parameters, and it is inserted after Mel Spectrum conditioning in each residual layer, i.e. the generated conditional vectors are sliced for each residual layer for the affine transformation.

For the parameter setup for DiffVel, the original setup is employed from DiffRoll: 15 residual layers, sampling rate 16000Hz, the hop size 512, the drop rate 0 to be fully supervised learning fashion, and the convolutional kernel size 9 for the residual layers. The loss function is L2 (mean square error) loss for the entire data segment, not note-level MIDI velocity error. We have tried Binary Cross Entropy (BCE) for better classification and L1 loss for MIDI velocity estimation. However, they did not work well in this diffusion model setup. Due to the limitation of computational resources, the epoch is stopped at 2000 for each training.

In the task of MIDI velocity estimation, which aims to get a number as value from the output, dealing with the input Gaussian noise is crucial. When the Gaussian noise is generated, it has a mean value of 0 and variance of 1. The diffusion step to denoise the Gaussian noise is set to 200 steps. However this noise is not perfectly removed after the diffusion steps and we need to perform denoising to each output by the post-processing.

During the post-processing, Gaussian noise removal is performed, which remained after the diffusion steps. This remainder causes a problem that it is considered as velocity during the evaluation process and causes 100% of the recall score and its note level estimation error is calculated high since the error is calculated where the note is not actually detected by the model. In this research, velocity estimation evaluation is made only on correctly detected notes. This evaluation constraint is applied to the other two models to be compared, the DNN [6] and NMF [15] models.

In order to remove the remaining noise, three methods are considered; one is to increase the diffusion steps since each step of diffusion step reduces the Gaussian noise. The second way is using a post-processing method employed by SegDiff, which averages the output from multiple inferences [25], at the expense of computational resources. However, these methods were not chosen due to the limitation of computational resources. In the process of removing the remaining Gaussian noise in the output, we calculated the distribution where the note does not exist in the ground truth score and defined a threshold to set output value 0. More precisely, a right Z-score is set based

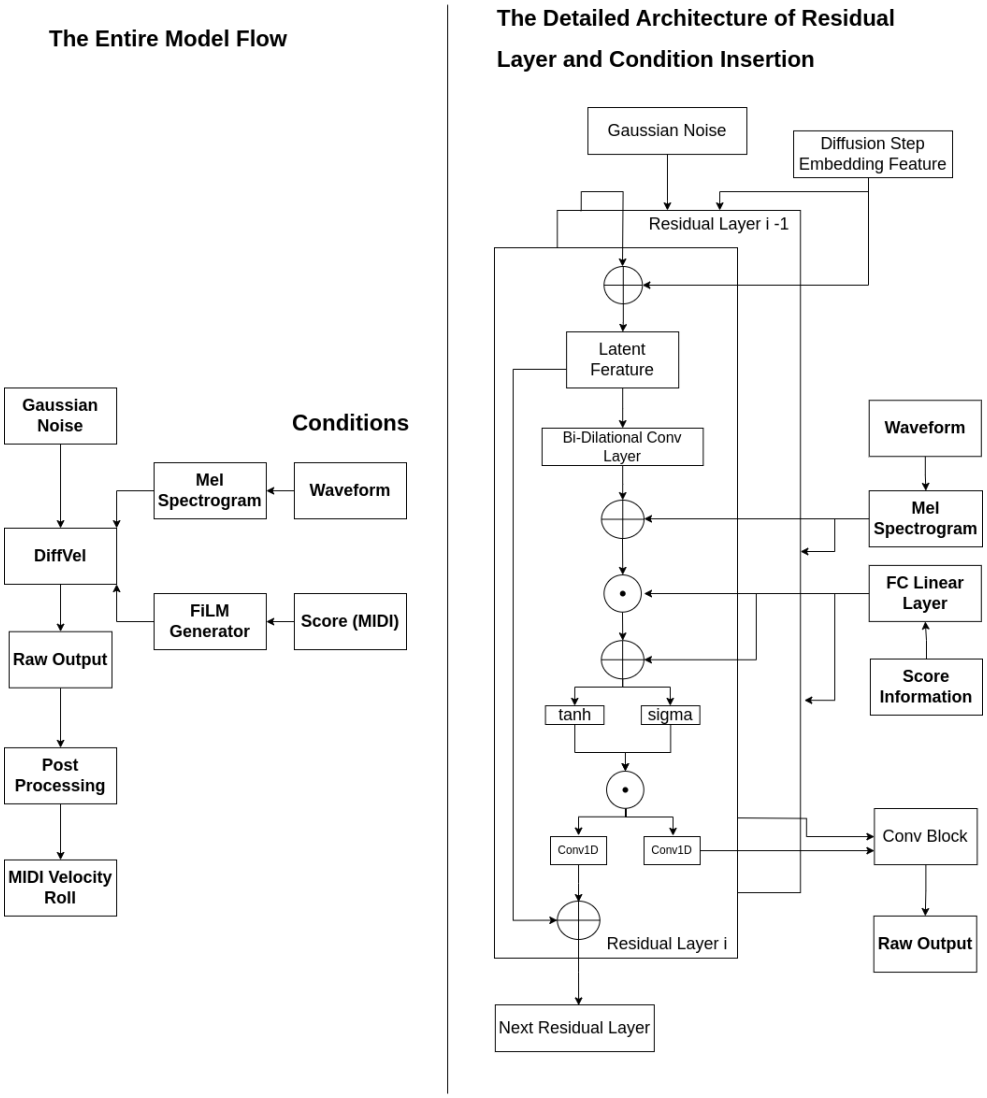

Fig. 2. The simplified overall process (right) and the detailed condition insertion into each residual layer (left)

on the distribution in order to find a threshold to set the output value to 0. Z-score is a number derived by following equation;

$$Z = \frac{x - \mu}{\sigma} \tag{2}$$

where x is observed value, μ and σ are mean and standard deviation of all output values for each score.

The reason why the right Z-scores is chosen, rather than taking the highest value in the area where the note does not exist in the ground truth, is that there are wrongly detected extra notes which have proper values to represent a MIDI velocity. These values are considered above the Z-score in the distribution of the remaining noise and do not affect the correctly detected values during the post-processing by not setting them to 0.

After removing the remaining Gaussian noise, we normalized the output to be in the range $[0, 1]$ looking at entire output value of each excerpt, not just for each output, and then scaled back to $[0, 127]$.

3.2 Evaluation

For testing purposes, we used the Saarland Music Data (SMD) dataset [27], which is also used for testing in previous researches [6, 15]. The dataset consists of students' piano performances, both audio data and MIDI data, which are perfectly aligned. The amount of data includes 50 classical piano excerpts, performed on Yamaha Disklavier. The original sampling frequency is 44.1kHz and down-sampled to 16kHz. We chose 49 excerpts from this dataset which are used in the score-informed NMF method by [15].

The model evaluation is made by taking an $L1$ distance of MIDI velocities for each note between ground truth and inference by the models, similarly to the previous research [15].

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

where i is each note and N is the number of correctly detected notes in the score.

The inferred MIDI velocity is the maximum value within the interval of each detected and classified velocity frame against the ground truth velocity frame for each note. This is because the detected velocity tends to fade after having the maximum value in the estimated MIDI velocity in a note frame as if depicting attack and fades of loudness of each note.

To evaluate the classification accuracy, recall score is chosen as the evaluation metric. This is because the estimation is masked by the given score, and recall is considered as the most appropriate evaluation metric for this classification problem when score is informed, as it takes into account both true positive and false negative. It measures the proportion of the total actual positive cases that are correctly identified by the classifier.

In this study, only correctly detected notes are evaluated, since we separate the MIDI velocity estimation accuracy and note detection accuracy as different research problem statements; the AMT and the MIDI velocity estimation as mentioned in the Section 2.

4 Results and Analysis

As we can see from the Table 1, FiLM conditioning to incorporate score information helped the estimation accuracy among the two models we have tested. The results show that FiLM conditioning improved MIDI velocity estimation but did not help with note detection for any setup. The result represents all note-wise errors inferred on the SMD dataset.

In terms of Gaussian noise removal, the right Z-score = 3 improved the overall accuracy significantly by sorting output values to correctly detected notes and the remaining noise after the diffusion steps. When post-processing is not performed and the noise remains, the evaluation method considers MIDI velocity detected and recall score is always 100%.

	Single Conditioning			Double Conditioning with Score		
Z-score	Mean	SD	Recall	Mean	SD	Recall
Raw Output	32.8	20.5	100%	28.6	19.2	100%
1	24.7	16.7	60%	21.0	14.5	56%
2	24.0	16.1	58%	20.2	13.6	54%
3	23.7	15.8	56%	**19.7**	**13.1**	53%

Table 1. The mean and standard deviation (SD) of the MIDI velocity estimation error for the models are based on Z-score for noise removal

The Figure 4 shows an example of the remaining Gaussian noise removal. The pale red color shown in the raw output is the remaining Gaussian noise from the input to the model, and setting Z-score determines the threshold to set the value to zero, attempting not to touch the detected notes. It can be intuitively seen that the remaining noise is removed without changing the value of the detected note velocities based on Z-score values.

Proposed Model		Conv-FiLM with Score [6]		NMF with Score [15]	
Mean	SD	Mean	SD	Mean	SD
19.7	13.1	15.1	12.3	4.1	5

Table 2. The comparison of results for the proposed model and previous research

We also compared the results to the previous models that have the same setup: a score-informed MIDI velocity estimation task. The Table 2 displays the mean and standard deviation (SD) values for proposed and previously researched models. The proposed model exhibited a mean value of 19.7 and an SD of 13.1, indicating the poorest performance among the three methods. In contrast, the Conv-FiLM DNN with Score

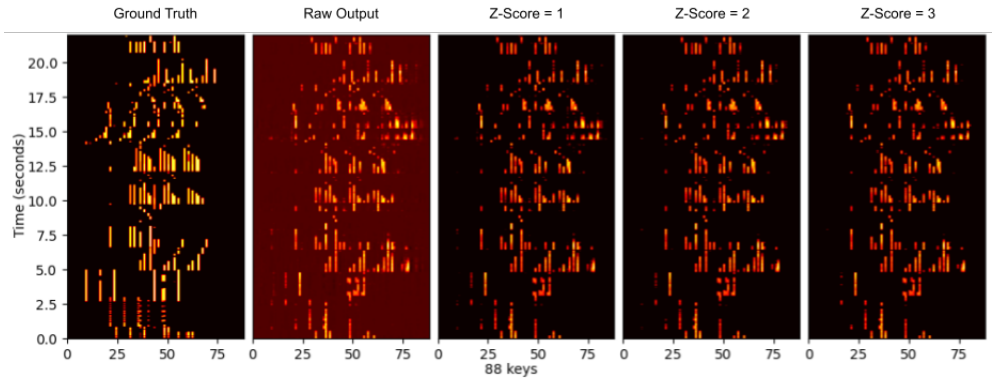

Fig. 3. The visualization and comparison of Gaussian noise removal for raw output and after removal are based on Z-score = 1, 2 and 3.

approach achieved a mean of 15.1 and an SD of 12.3, while the NMF with score method demonstrated the best performance with the lowest mean and SD values, at 4.1 and 5, respectively. Although the proposed model currently underperforms compared to the other models, it is important to note that the difference between their mean values is not substantial and we do not know significance in the sense of perceptual loudness yet.

Figure 4 displays the deviation of the error in each range of MIDI velocity, pitch, and sustain pedal activation respectively for both models. The box-charts for pitch and sustain pedal are similar figure for both models. These charts demonstrate that the more training data notes you have, the more accurate your MIDI velocity estimation will be, looking at the note ratio in the training dataset. This implies that data augmentation, such as pitch shifting, is necessary for low and high pitch notes in the training data. When looking at the error based on the MIDI velocity group, it is interesting to observe that FiLM conditioning improved the model's estimation for lower velocity notes, but resulted in worse estimation for higher velocity notes compared to the model without score information. It was also observed that both models tend to estimate MIDI velocity lower than the ground truth. Further analysis is required to interpret this phenomenon.

5 Discussion and Future Work

In this study, experiments on a diffusion model with double conditions for note level MIDI velocity estimation for piano performance have been conducted. We discovered that FiLM conditioning for score information insertion improved the estimation error and standard deviation on the overall test data.

We need to investigate the how the MIDI velocity error gives us the human perceptual sense to give the true evaluation of the model. As is mentioned in the introduction, there is still no research has been conducted for creating mapping from MIDI velocity to perceptual loudness. This will be one of our future works to keep this research move forward.

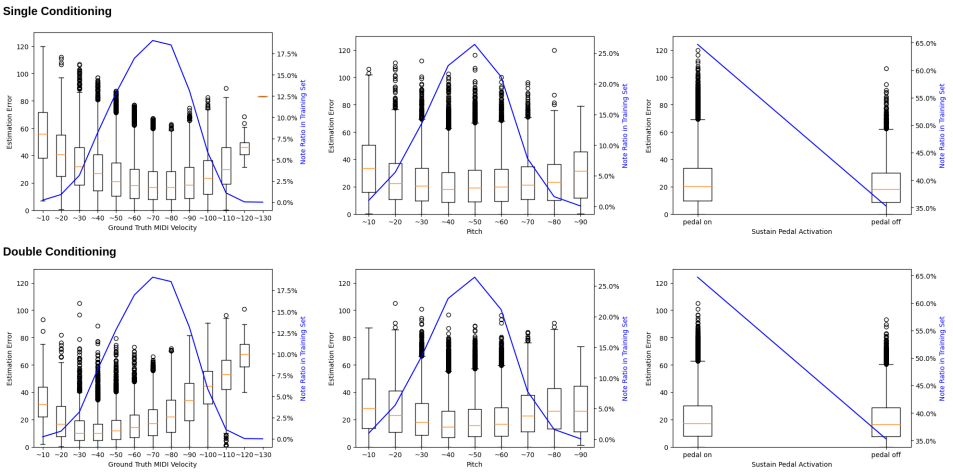

Fig. 4. The error analysis is based on ground truth MIDI velocity, pitch and sustain pedal activation. The box charts in the upper row display results for the single-conditioned model. Similarly, the box charts in the lower row show the results of the model with score information.

One of the downside of the models is they take significant amount of time and computational cost for an inference and model convergence in training phase. In this study, for example, it took about 2.5 minutes for 20 seconds of MIDI velocity roll output. This problem would be a blocker for a use case which requires real time processing.

The model achieved a similar result to the DNN used in a previous study [6], which indicates that the direction of this research is promising, and further exploration is warranted. Due to computational limitations, the training was stopped at 2000 epochs. However, the losses on validation set are still showing the trend of decrease on each model. This indicates that further training could improve accuracy within a short period of time with high confidence. Moreover, the recent rapid development and evolution on generative models including the diffusion model will improve its transcription accuracy, and at the same time, it would lead more attention to the FiLM conditioning to realize multi-modality for certain use-case scenarios such as education purposes which needs score and audio information.

Since it has been observed that FiLM conditioning improves the estimation results, further investigation into the condition generator is necessary for better estimation and note detection, rather than a simple fully connected linear layer. Moreover, the proposed diffusion model is adaptable to multiple conditioning techniques, making feature engineering a particularly suitable strategy for optimization within the DiffVel setup. By refining the features used in the model, it may be possible to extract more meaningful patterns and relationships from the data, ultimately leading to improved results. Additionally, incorporating more data and extending the training process could potentially enhance the proposed model's performance. Therefore, future research can focus on these aspects to optimize the proposed model and potentially achieve better performance than the existing approaches.

In real-world use cases, such as music education, score alignment must be taken into account for conditioning. A Dynamic Time Warping will be used to address this issue in a future work.

The code and the dataset used for this research would be provided upon request.

References

1. Kim, Hyon and Ramoneda, Pedro and Miron, Marius and Serra : An overview of automatic piano performance assessment within the music education context, Xavier : 2022 : SCITEPRESS–Science and Technology Publications
2. Federico Simonetta, Federico Avanzini, Stavros Ntalampiras : A Perceptual Measure for Evaluating the Resynthesis of Automatic Music Transcriptions : arXiv:2202.12257 [cs.SD]
3. Grachten, Maarten and Widmer, Gerhard : Linear Basis Models for Prediction and Analysis of Musical Expression : Journal of New Music Research, volume 41, number 4, pages 311–322, 2012
4. Hamond, Luciana Fernandes and Welch, Graham and Himonides, Evangelos : The pedagogical use of visual feedback for enhancing dynamics in higher education piano learning and performance : Opus, 25, 3, pages 581–601 year 2019
5. Hamond, Luciana Fernandes: The pedagogical use of technology-mediated feedback in a higher education piano studio: an exploratory action case study : 2017 UCL (University College London)
6. H. Kim, M. Miron, X. Serra : Score-Informed MIDI Velocity Estimation for Piano Performance by FiLM Conditioning : Proc. Int. Conf. Sound and Music Computing, 2023
7. Kin Wai Cheuk, Ryosuke Sawata, Toshimitsu Uesaka, Naoki Murata, Naoya Takahashi, Shusuke Takahashi, Dorien Herremans, Yuki Mitsufuji : DiffRoll:Diffusion-based Generative Music Transcription with Unsupervised Pretraining Capability : arXiv:2210.05148 [cs.SD]
8. Kosta, Katerina and Ramírez, Rafael and Bandtlow, Oscar F and Chew, Elaine : Mapping between dynamic markings and performed loudness: a machine learning approach : Journal of Mathematics and Music, volume 10, number 2 pages 149–172, 2016, Taylor & Francis
9. Kosta, K., O. F. Bandtlow, E. Chew : Outliers in Performed Loudness Transitions: An Analysis of Chopin Mazurka Recordings. : International Conference for Music Perception and Cognition (ICMPC), pages 601-604, 2016, California, USA
10. Benetos, Emmanouil and Dixon, Simon and Giannoulis, Dimitrios and Kirchhoff, Holger and Klapuri, Anssi : Automatic music transcription: challenges and future directions : Journal of Intelligent Information Systems, volume 41, page 407–434, 2013, Springer
11. Sarah Kim and Jeong Mi Park and Seungyeon Rhyu and Juhan Nam and Kyogu Lee : Quantitative analysis of piano performance proficiency focusing on difference between hands, PLoS ONE volume 16, 2021
12. Katerina Kosta and Oscar F. Bandtlow and Elaine Chew : Dynamics and relativity: Practical implications of dynamic markings in the score : Journal of New Music Research, volume 47, number 5, pages 438-461, 2018, Routledge : `https://doi.org/10.1080/09298215.2018.1486430`
13. Qu, Yang and Qin, Yutian and Chao, Lecheng and Qian, Hangkai and Wang, Ziyu and Xia, Gus : Modeling Perceptual Loudness of Piano Tone: Theory and Applications : arXiv preprint arXiv:2209.10674
14. Goebl, W. : Melody lead in piano performance: expressive device or artifact? : The Journal of the Acoustical Society of America, volume 110, number 1, pages 563-72, 2001, Acoustical Society of America

15. Jeong, Dasaem and Kwon, Taegyun and Nam, Juhan : Note-Intensity Estimation of Piano Recordings Using Coarsely Aligned MIDI Score, volume 68, pages 34–47, number 1, Journal of the Audio Engineering Society, JAES, Audio Engineering Society
16. Ewert, Sebastian and Müller, Meinard : Estimating note intensities in music recordings : 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 385–388, 2011, IEEE
17. Devaney, Johanna and Mandel, Michael : An evaluation of score-informed methods for estimating fundamental frequency and power from polyphonic audio : 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 181–185
18. Jeong, Dasaem and Nam, Juhan : Note intensity estimation of piano recordings by score-informed NMF : Audio Engineering Society Conference: 2017 AES International Conference on Semantic Audio, 2017, Audio Engineering Society
19. Manilow, Ethan and Pardo, Bryan : Bespoke neural networks for score-informed source separation : arXiv preprint arXiv:2009.13729, 2020
20. Perez, Ethan and Strub, Florian and de Vries, Harm and Dumoulin, Vincent and Courville, Aaron : FiLM: Visual Reasoning with a General Conditioning Layer : `http://arxiv.org/abs/1709.07871`,
21. Meseguer-Brocal, Gabriel and Peeters, Geoffroy : Conditioned-U-Net: Introducing a control mechanism in the U-Net for multiple source separations : arXiv preprint arXiv:1907.01277, 2019
22. Miron, Marius and Carabias Orti, Julio J and Janer Mestres, Jordi : Improving score-informed source separation for classical music through note refinement : Proceedings of the 16th International Society for Music Information Retrieval (ISMIR) Conference; 2015 Oct 26-30; Málaga, Spain. Canada: International Society for Music Information Retrieval; 2015.
23. Kim, Jong Wook and Bello, Juan Pablo : Adversarial learning for improved onsets and frames music transcription : arXiv preprint arXiv:1906.08512, 2019
24. Kelz, Rainer and Dorfer, Matthias and Korzeniowski, Filip and Böck, Sebastian and Arzt, Andreas and Widmer, Gerhard : On the Potential of Simple Framewise Approaches to Piano Transcription : arXiv:1612.05153 [cs]
25. Tomer Amit, Tal Shaharbany, Eliya Nachmani, Lior Wolf : SegDiff: Image Segmentation with Diffusion Probabilistic Models : arXiv:2112.00390 [cs.CV]
26. Curtis Hawthorne and Andriy Stasyuk and Adam Roberts and Ian Simon and Cheng-Zhi Anna Huang and Sander Dieleman and Erich Elsen and Jesse Engel and Douglas Eck : Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset : International Conference on Learning Representations, 2019
27. Müller, Meinard and Konz, Verena and Bogler, Wolfgang and Arifi-Müller, Vlora : Saarland music data (SMD), 2011
28. Slizovskaia, Olga and Haro, Gloria and Gómez, Emilia : Conditioned source separation for musical instrument performances : IEEE/ACM Transactions on Audio, Speech, and Language Processing : volume 29, pages 2083–2095, 2021,
29. Kong, Qiuqiang and Li, Bochen and Song, Xuchen and Wan, Yuan and Wang, Yuxuan : High-resolution piano transcription with pedals by regressing onset and offset times : IEEE/ACM Transactions on Audio, Speech, and Language Processing : volume 29, pages 3707–3717, 2021
30. Roger B. Dannenberg : The Interpretation of MIDI Velocity : International Conference on Mathematics and Computing