SCORE-INFORMED MIDI VELOCITY ESTIMATION FOR PIANO PERFORMANCE BY FiLM CONDITIONING

Hyon Kim
Music Technology Group,
Universitat Pompeu Fabra, Barcelona
hyon.kim@upf.edu

Marius Miron
Music Technology Group,
Universitat Pompeu Fabra, Barcelona
marius.miron@upf.edu

Xavier Serra
Music Technology Group,
Universitat Pompeu Fabra, Barcelona
xavier.serra@upf.edu

ABSTRACT

Piano is one of the most popular instruments among people that learn to play music. When playing the piano, the level of loudness is crucial for expressing emotions as well as manipulating tempo. These elements convey the expressiveness of music performance. Detecting the loudness of each note could provide more valuable feedback for music students, helping to improve their performance dynamics. This can be achieved by visualizing the loudness levels not only for self-learning purposes but also for effective communication between teachers and students. Also, given the polyphonic nature of piano music, which often involves parallel melodic streams, determining the loudness of each note is more informative than analyzing the cumulative loudness of a specific time frame.

This research proposes a method using Deep Neural Network (DNN) with score information to estimate note-level MIDI velocity of piano performances from audio input. In addition, when score information is available, we condition the DNN with score information using a Feature-wise Linear Modulation (FiLM) layer. To the best of our knowledge, this is the first attempt to estimate the MIDI velocity using a neural network in an end to end fashion. The model proposed in this study achieved improved accuracy in both MIDI velocity estimation and estimation error deviation, as well as higher recall accuracy for note classification when compared to the DNN model that did not use score information.

1. INTRODUCTION

Many piano performances depart from symbolic representations: sheet music or scores. Dynamic markings are engraved in these symbolic representations denoting changes in loudness of the played notes that significantly affect expressiveness and emotional impact. Furthermore, scores encoded as MIDI contain note velocity information which marks the loudness of each note. Rendering the appropriate loudness for each piano note is a crucial piano skill and it is something that pianists hone over time.

In terms of music education, providing visual feedback of performance has been studied and found to be an effective way to improve students' skills [1, 2]. To that extend, understanding and controlling loudness is particularly important in this context [3]. Loudness estimation and visualization techniques satisfy a system requirement for giving feedback to students.

On the other hand, piano performance transcription is also an actively researched topic [4–6]. However, these studies primarily focus on detecting individual notes, rather than note loudness or dynamic symbols such as $forte$, $mezzoforte$, $piano$, $pianissimo$, $crescendo$, etc. Additionally, the transcription process is not yet fully accurate and reproducible of performance.

There are a few papers which investigated mapping from audio to MIDI velocity on note level for piano performance [7–10]. These researchers applied an NMF method to separate piano performance audio to the 88 piano keys and estimated a MIDI velocity on each note, together with score information.

In order to avoid confusion, here we refer to loudness as aggregated MIDI velocities for a certain time frame measured by a electric piano device, whilst intensity refers to maximum value of frequency sum for a note frame as defined in [7]. A series of loudness values associated with each note in the score alters the dynamics of a piece and ultimately its expressiveness [11]. Note that using MIDI velocity we predict loudness at a lower granularity than the dynamic markings, which are explicitly written in most of music scores as symbols that 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 [12, 13]. Therefore, the note level loudness itself has special meaning in the piano performance, considering its polyphonic characteristics.

It is important to note that the MIDI velocity does not directly correlate with the perceived loudness in terms of the human auditory system. Research has been conducted to investigate the relation between MIDI velocity and perceptual loudness in dB [14]. This paper showed a consistent trend of perceptual loudness in increasing dB as MIDI velocity increases nevertheless it is non-linear mapping [15]. [16, 17] researched mapping from perceptual loudness value in dB scale to these dynamic symbols for piano performance. These research indicate that estimating MIDI velocity is meaningful in terms of perceptual human hearing.

The task of estimating MIDI velocity for individual piano note involves two problems that need to be solved. The

first task involves classification, which requires identification of the specific piano key that has been pressed, and is an essential component of automatic piano transcription tasks. The second problem is a regression task that requires the estimation of numbers within the range of 0-127 for MIDI velocity on each note. To address these challenges, we propose a novel end-to-end approach, using a deep neural network (DNN) with FiLM conditioning layers [18] to incorporate score information into the DNN. We conduct experiments to estimate MIDI velocity using this approach. To evaluate the performance of these models, we measured the MIDI velocity error for each note and computed the mean value for each score. In addition to the computation of the mean MIDI velocity error for each score, this study also includes error analysis and visualizations.

2. RELATED WORK

There are only two papers which take note level MIDI velocity into account in music performance analysis and they tried to solve this by NMF [7, 10]. NMF methods have been used for source separation problems and well applied to music source separation areas as well [19]. [7] investigated an NMF method with its score information to estimate note level intensity first and then created a linear regression model to acquire note level MIDI velocity estimation. This research showed detailed analysis on the error of the NMF method and explained causes of these errors.

[8] aimed to estimate the note level intensity the from the spectrogram by filtering it according to the frequency of each note. To the best of our knowledge, there has not been any research that estimates note-level intensity using a deep neural network (DNN) method. However, with recent advancements in DNNs for audio signal processing, it may be possible to use such an approach. In our study, we compare our system with the NMF method proposed in [7] as our benchmark.

The piano performance transcription is one of the closest problems for classification from audio input. [20] 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 an 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 [7].

No research has been done on note level MIDI velocity estimation by a modern DNN when the score is given and used to inform the model. However, existing research is utilising score information to inform music instrument separation in polyphonic music [21–23]. These works utilised score or video information to get better result of source separation by creating another neural network to extract features of the additional features which are fed into an original DNN.

In this paper, we propose to use FiLM conditioning [18] 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 [18]. 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 [23].

3. METHOD

3.1 Model Architecture

We modified the piano performance transcription model in order to classify the audio into the 88 keys [20]. This model is known as the state of the art for a piano performance transcription. Figure 1 shows the structure of the entire model. This architecture first takes audio and converts it to a Log Mel-frequency Spectrogram in order to convert the waveform to be an image form. The audio window length is two seconds and hop size is one second. Sampling rate is set to 16k Hz. We take a frame of 100 segments corresponding to one second of audio. The two dimensional form of audio is processed through convolutional neural networks (CNN) as in the Figure 2. The channel size of the CNN layers are 1, 48, 64, 96, 128 as each block of CNN processes the input as the Figure 1. After the process of CNN layers, a bi-directional Gated Recurrent Unit (GRU) processes the data to check the time series of the audio data. And then the input data is classified into the 88 keys by a liner layer with the sigmoid activation function.

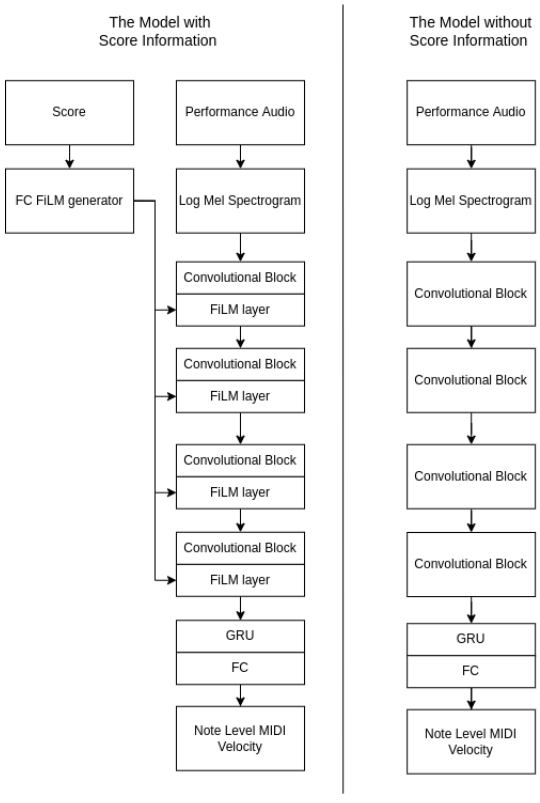

Figure 1. The model architectures of the conditional DNN and the unconditioned DNN for score informed MIDI velocity estimation

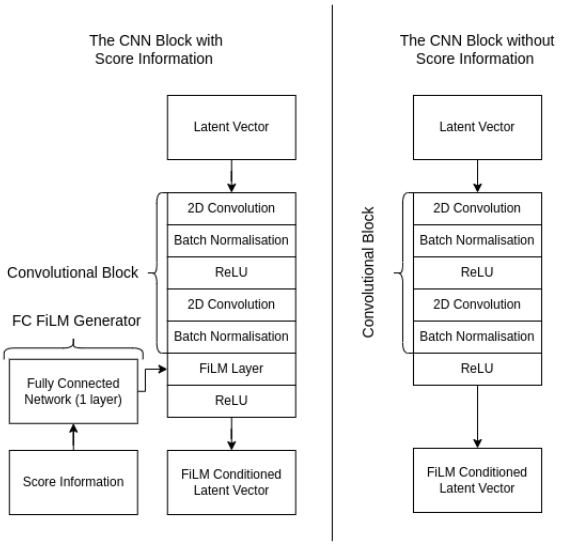

Figure 2. The detailed structures of convolutional blocks of the DNN with and without the FiLM conditioning

In order to take advantage of the classification characteristics of this network, we kept Binary Cross Entropy (BCE) classification loss for its loss function. On top of the BCE loss, we added the loss function 2 to estimate MIDI velocity which takes the $l1$ distance between the output MIDI velocity from the model and the ground truth MIDI velocity solely for the note frames. The MIDI velocity is scaled between 0-127 and represents loudness for each performed key on piano, the higher the value and the louder the sound. The employed loss function 1 is a combination of $l1$ loss and the BCE loss connected by a convex function so that we can back propagate losses for both classification and regression.

$$Loss = \theta * l1\ loss + (1 - \theta) * bce\ loss \tag{1}$$

where $\theta \in [0, 1]$ is the weight of the convex function and currently it is set to 0.5 for experimental purposes.

The $l1$ loss function in this research is defined as follows;

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

where i is an index of corresponding notes between ground truth and output within a window and N is the number of notes in the window. One data point for an input consists of two seconds and each frame contains 100 segments per second to represent the MIDI roll. We have tested other well known loss functions, mean square error, Kullback-Leibler divergence loss, the $l1$ loss. However, all of them did not work for this experiment having classification and regression aspects together.

For the purpose of inserting the score information, we also added a FiLM conditioning layer as it is introduced in Section 2. 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.

$$FiLM(x) = \gamma(z) \cdot x + \beta(z) \tag{3}$$

where vector z is a conditional vector.

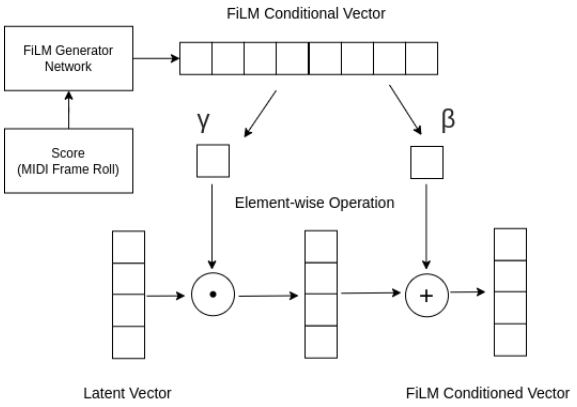

Figure 3. The diagram for the operation flow to insert a FiLM condition to the base DNN.

The Figure 3 is 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.

In this research, the FiLM generator is designed as a fully connected layer to generate conditioning parameters and it is inserted each block of CNN at the end as the Figure 2. The FiLM conditional vector is sliced for each convolutional block for a scalar multiplication and addition (affine transformation) for elements in latent vector. We also experimented with an element-wise operation by generating as many elements as the latent vector contained and performing the affine transformation for each element in the latent vector with different elements in the FiLM conditional vector. However, our current setup demonstrated the best performance.

3.2 Model Evaluation

We used the MAESTRO data set [24] for training purposes. We randomly chose 132 excerpts from the MAESTRO data set for training and the amount of data size is 2.8GB including audio and MIDI roll is selected. This limitation is made due to our limited computational resources and also in order to speed up the training process and see the result of the training process to improve the MIDI velocity estimations. The MIDI data contains the following data; note onset, note offset, note frame, MIDI velocity, pedal onset, pedal offset, pedal frame. In this data, MIDI velocity is assigned on the note frame, not on the place where onset is activated. We used the MIDI velocity data on note frame level as supervised data for training.

For the test purposes, we used the Saarland Music Data (SMD) data set [25]. The data set consists of students' piano performance both audio data in mp3 format and MIDI data which are perfectly aligned. We chose this data set

in order to compare the results against the score informed NMF method by [7]. The original sampling frequency is 44.1kHz. The amount of data is 50 classic piano excerpts and performed on an acoustic piano augmented with a MIDI interface, Yamaha Disklavier. Among the data in the SMD data set, excerpts used in previous research [7] are chosen to be tested for comparison purposes. We evaluated the model not only perfectly aligned cases with audio and score information into FiLM layer, but also a case of the score information is unaligned against audio input.

The evaluation is made by taking an $l1$ distance of MIDI velocities between ground truth and inference by the model, similarly to the previous research [7].

$$Error = \frac{\sum_i |V(i)_{\text{ground truth}} - V(i)_{\text{inference}}|}{N} \quad (4)$$

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

The inferred MIDI velocity is the maximum value within the interval of each detected and classified note frame. This is because velocity fades 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. Since this is the score informed task, the error is calculated only where note frames exist, i.e. the estimated notes are masked by its score.

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 to be the most appropriate evaluation metric for this classification problem 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 as such by the classifier.

4. RESULT AND DISCUSSION

4.1 Results on the Test Set

The results of evaluation are presented in Table 1. The table consists of the result of three models: the FiLM conditioned model (the proposed model, which uses the DNN with score), the unconditioned model (the DNN without score), and the model from [7] as a benchmark. The table shows the mean of the note-level errors, the standard deviations and the recall score for both the conditioned and unconditioned models. It should be noted that errors in the DNN models used in this study include instances of misclassified notes, where the notes are present in the ground truth but are not accurately detected by the models. This phenomenon is reflected in the recall score of the model.

The results in Table 1 indicate that the FiLM conditioning improves the MIDI velocity estimation by 0.3. The best MIDI velocity estimation of the proposed model is 5.8 for BWV875-01 by Bach, which is better than the DNN model without score conditioning, and there is a gap of 0.9 to the NMF model with score conditioning. The largest difference between the conditioned and unconditioned model is 3.3 for Bartok Sz80-03. The smallest, the biggest and the average gap towards the NMF method is 3.9 by Bach BWV875-01, 44.4 by Bach BWV888-02 and 11.0 respectively on the test set. However, it is not clear that how the gaps of the estimated MIDI velocities affects perceptual loudness. Moreover, there is a huge potential room to improve the accuracy by adding more training data into it considering the limitation of our experimental setup introduced in Section 3.

There are pros and cons for the discussion of DNN vs. NMF. For example, DNN takes longer time to train and is data hungry in general, but once data is available and optimization of training process goes well, it processes faster on test set and applicable to all unseen data in the same domain of training data. On the other hand, NMF starts optimising its parameters when it takes the test data. Therefore it takes a longer time to get the inference result (in this case, MIDI velocity) than the DNN model. Also, the optimized NMF for a data is not applicable to the other data. The NMF needs another iteration process every time new test data comes in.

Despite not improving significantly the MIDI velocity estimation itself, the FiLM conditioning has contributed to the model by increasing the classification accuracy where note frames exist. This is particularly evident in the improved recall score, which is an important metric for this classification problem. Overall, the average recall score improved from 80.9% to 85.8%. The recall scores improved on most of the excerpts, except for BWV871-01 by Bach. This suggests that the FiLM conditioning is effective in improving the performance of the proposed model, in particular, in terms of reducing the range of error and increasing the classification accuracy where note frames exist.

4.2 Visualisation of a Result

The Figure 4 illustrates the loudness of a gourd truth, a model estimation with and without score information of each note during a piano performance. It can intuitively be seen that the conditioned model classifies the audio into 88 keys more accurately. This shows that the inserted score information turned into a conditional vector on the base DNN is helping for classification and focus on MIDI velocity estimation. This visualization is particularly useful in a pedagogical setting, as dynamics, or the sequence of loudness, play a crucial role in piano performance. By providing a direct feedback on dynamics, this system can aid both students and teachers in understanding their performance. Additionally, by aligning with real-world use cases, this system is well-suited for educational purposes. This system for education aligns for the real use case scenario.

4.3 Results for Unaligned Score Cases

In practical scenarios, however, it is often the case that the score information is not perfectly aligned with the audio. The Figure 5 demonstrates the correlation between the unaligned time shift and the drop in overall accuracy of the model with score information. The unaligned time shift is a generated time gap between audio and MIDI roll in this

		The DNN with Score			The DNN without Score			The NMF with Score [7]	
Composer	**Excerpt**	**Mean**	**SD**	**Recall**	**Mean**	**SD**	**Recall**	**Mean**	**SD**
Bach	BWV849-01	8.7	5.8	89.1%	9.1	6.4	85.5%	2.6	3.0
Bach	BWV849-02	8.2	6.1	87.7%	9.0	6.6	83.8%	2.3	2.5
Bach	BWV871-01	8.3	5.7	90.7%	7.5	6.7	91.9%	1.7	2.1
Bach	BWV871-02	9.0	5.6	90.7%	9.9	6.6	89.9%	2.0	2.1
Bach	BWV875-01	5.8	5.1	91.0%	6.7	7.4	90.9%	1.9	1.9
Bach	BWV875-02	7.2	5.5	90.2%	8.3	6.5	89.7%	1.9	2.1
Bach	BWV888-01	11.4	10.9	87.8%	13.2	11.3	83.0%	2.8	3.1
Bach	BWV888-02	46.2	25.4	82.1%	47.7	25.1	77.1%	1.8	2.2
Bartok	Sz80-01	19.1	21.0	83.0%	22.2	22.5	79.4%	4.8	8.4
Bartok	Sz80-02	13.6	11.7	90.3%	13.8	12.9	81.1%	5.0	5.9
Bartok	Sz80-03	25.4	23.8	80.8%	28.7	24.3	77.6%	5.2	7.9
Beethoven	Op27No1-01	11.9	10.1	89.6%	12.3	11.2	86.7%	3.6	4.3
Beethoven	Op27No1-02	14.1	8.8	89.9%	13.6	8.4	84.0%	3.7	3.9
Beethoven	Op27No1-03	11.9	10.3	88.1%	12.5	11.5	85.9%	3.3	5.5
Beethoven	Op31No2-01	10.4	9.2	90.2%	10.5	9.0	81.7%	4.0	5.2
Beethoven	Op31No2-02	18.0	15.7	91.7%	18.6	16.4	86.2%	4.0	4.3
Beethoven	Op31No2-03	9.6	7.9	86.6%	10.2	8.4	82.9%	2.4	2.9
Brahms	Op5No1	17.6	20.1	80.6%	19.3	20.7	74.4%	6.4	8.5
Brahms	Op10No1	12.1	10.2	86.2%	11.7	9.6	78.2%	5.7	6.5
Brahms	Op10No2	13.0	11.7	84.5%	13.0	13.0	77.7%	5.2	6.7
Chopin	Op10-03	12.8	9.6	84.7%	12.1	9.1	80.2%	4.4	4.3
Chopin	Op10-04	13.0	12.5	78.9%	14.9	15.2	75.6%	3.3	4.4
Chopin	Op26No1	13.9	10.5	86.9%	13.0	9.6	81.7%	3.7	4.8
Chopin	Op26No2	13.9	11.7	87.2%	13.7	12.1	84.2%	6.8	6.9
Chopin	Op28-01	11.6	8.8	84.1%	10.5	8.8	78.3%	4.1	4.0
Chopin	Op28-03	10.0	8.0	86.0%	9.1	7.5	80.9%	3.0	3.4
Chopin	Op28-04	14.5	7.3	91.5%	13.1	7.6	86.4%	4.4	3.8
Chopin	Op28-11	12.6	7.4	86.5%	10.7	7.3	83.0%	3.6	3.6
Chopin	Op28-15	14.4	8.7	88.8%	13.9	9.6	81.2%	4.9	4.3
Chopin	Op28-17	16.4	12.6	84.8%	17.2	13.5	79.4%	5.8	6.2
Chopin	Op29	9.1	7.5	84.0%	9.4	8.1	78.9%	4.5	4.4
Chopin	Op48No1	14.1	11.2	84.2%	12.6	10.5	76.1%	5.8	6.2
Chopin	Op66	12.2	8.8	84.2%	11.9	8.3	76.7%	4.0	4.0
Haydn	Hob17No4	10.3	8.0	89.4%	11.0	8.9	87.3%	2.5	4.2
Haydn	Hob16No52-1	31.2	24.3	85.6%	31.8	24.3	81.2%	3.5	4.1
Haydn	Hob16No52-2	17.5	16.5	90.4%	18.2	16.7	84.5%	3.6	4.0
Haydn	Hob16No52-3	34.6	24.7	86.4%	37.1	24.5	81.7%	3.2	4.6
Liszt	Lecture Dante	14.7	12.9	78.3%	14.2	12.4	70.7%	6.9	8.9
Liszt	S.179	13.6	10.0	79.5%	12.6	9.8	74.9%	7.0	9.2
Liszt	S.144-2	13.8	13.6	84.9%	13.3	11.5	79.3%	4.6	7.3
Mozart	K.265	13.3	12.6	90.2%	14.5	13.9	88.6%	2.8	3.8
Mozart	K.398	20.0	20.0	89.9%	21.7	21.1	85.9%	3.1	3.4
Rachman.	Op36-1	18.3	18.2	81.2%	18.2	18.3	74.5%	5.5	7.0
Rachman.	Op36-2	14.0	12.9	83.9%	12.6	12.5	76.0%	5.7	7.2
Rachman.	Op36-3	29.3	26.1	74.7%	30.9	26.0	69.2%	5.7	8.8
Rachman.	Op39No1	14.6	13.3	75.8%	14.7	14.1	70.7%	4.4	6.2
Ravel	Jeux d'eau	19.6	17.7	76.9%	17.9	15.3	71.3%	5.6	6.9
Ravel	Valse Nobles	13.9	11.2	84.2%	13.2	10.9	76.9%	4.9	7.0
Skryabin	Op8No8	13.2	7.3	87.9%	12.6	7.5	83.8%	3.6	3.9
Average		**15.1**	**12.3**	**85.8%**	**15.4**	**12.6**	**80.9%**	**4.1**	**5.0**

Table 1. The results for each excerpt in the test set. Mean and Standard Deviation(SD) of the estimation error and recall score are listed.

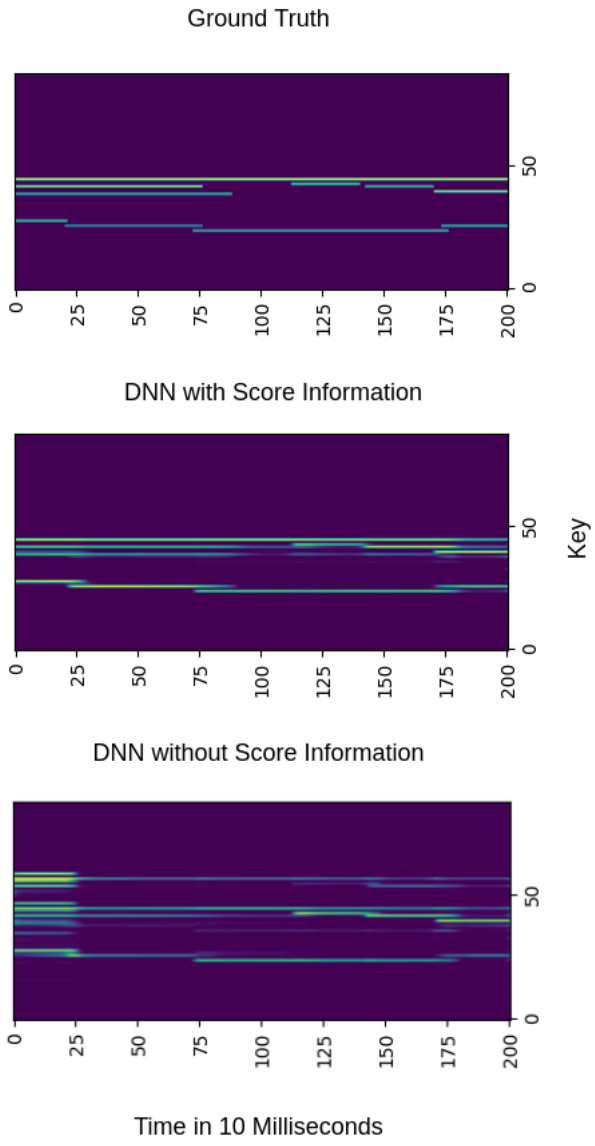

Figure 4. Visualization of MIDI velocity: The time in the horizontal axis is in 10ms i.e. two seconds in total, and the 88 keys is in the vertical axis.

experiment. The chart shows that clear correlation for the drop of overall accuracy of the model with score information.

There have been several methods and models to overcome this issue. The Dynamic Time Warping (DTW) is one of the most widely used models for aligning audio to score. [26] researched and implemented a score alignment by a DNN method. These methods are feasible to try on this model to improve accuracy.

4.4 Error Analysis

Additionally, we compared the distributions of notes in both the training set and the inferred result of the test set for factors that can induce error in the previous research [7]; pitch, ground truth MIDI velocity, and sustain pedal activation.

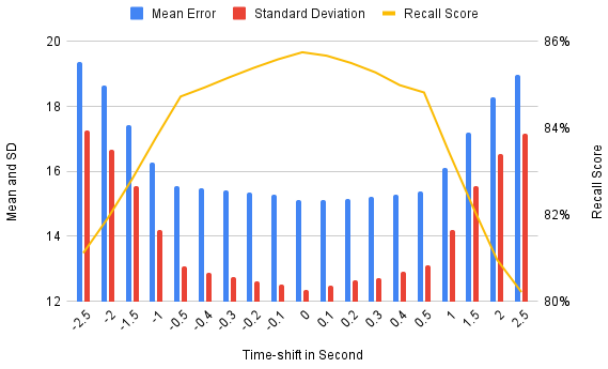

Figure 5. The relation between unaligned score and input audio in time and each metrics; the error of mean, the standard deviation and the recall score.

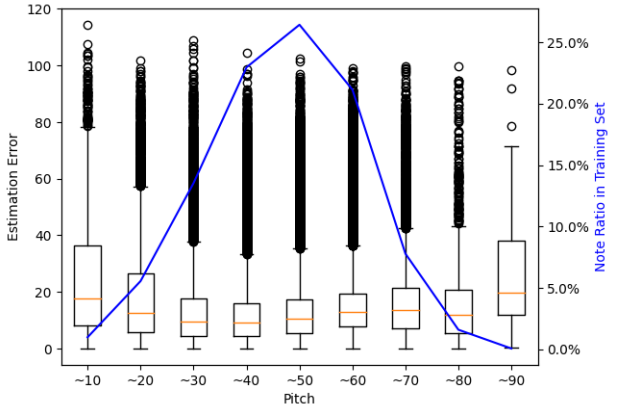

Figure 6. The Estimation Errors Based on Pitch Groups and the Ratio of Notes in the Training Set.

In Figure 6, the horizontal axis groups the pitch into 10 intervals and displays a box plot of the estimation error for correctly classified notes in the 88 keys using the model with score information. The accompanying line plot illustrates the proportion of each group present in the training set. The figure demonstrates that a more extensive training data set results in more accurate error estimation by the model. To address this, the distribution of the training set should be optimized through data augmentation techniques, with a focus on adding more samples with pitches below 30 and above 70.

The argument applies similarly to Figure 7, which groups the ground truth MIDI velocity into 10 intervals and displays the deviation of the estimation error for each group using the proposed model with score information. A close examination of the figure reveals a lack of data for ground truth velocity groups below 20 and above 110, leading to noticeable error deviations.

Figure 8 showcases the estimation errors when the sustain pedal is activated and inactive for notes correctly classified by the proposed model with score information. The training set includes approximately 65% of notes with the sustain pedal activated and the remaining without. The figure demonstrates that a larger training data set results in

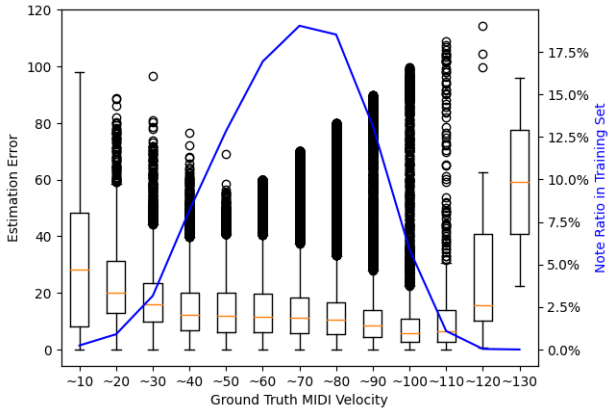

Figure 7. The Estimation Error Based on the Ground Truth MIDI Velocity Groups and the Ratio of Notes in the Training Set.

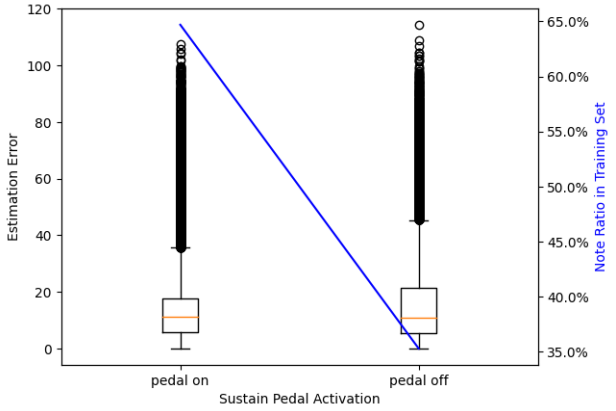

Figure 8. The Estimation Error Based on the Sustain Pedal On and Off and the Ratio of Notes in the Training Set.

improved accuracy and reduced deviation in the model's estimation of MIDI velocity.

The various aspects of estimation error require proper data augmentation for optimization. The iterative manipulation of the training set leads to enhanced accuracy for the proposed model.

The estimation error figures exhibit numerous outliers in each group, which can be addressed through post-processing statistical methods to minimize error, finding correlations between each notes.

5. CONCLUSION

In this research, we proposed an end-to-end method for estimating MIDI velocity from audio using a DNN with a FiLM conditioning mechanism that inputs score information. The model improved upon previous method [7] by being more generic and able to process all music performed on piano, not modeling on a single excerpt by the NMF. Furthermore, the proposed model demonstrated substantial improvement in note detection and classification accuracy.

The performance of the model can be further improved by increasing the amount of data and fine-tuning the training data to a specific domain. The results indicate that the proposed model is a promising step towards accurate MIDI velocity estimation in an end-to-end fashion and can be applied to real-world scenarios such as performance visualization. In the real use case scenarios, it is possible to apply the system to music education and dynamic markings transcription as in $forte$, $piano$, $mezzoforte$, $crescendo$, $decrescendo$, etc.

In the context of music education applications, the model should take care about a bigger training data set for the domain of students' performance. We also need to care about the unaligned case between audio, MIDI and score. However, visualisation of loudness such as Figure 4 gives students an objective way to see their performance and it is one of the best visualization tools if teachers model performance is recorded in MIDI format. It also gives benefits to teachers to check students' performance in a shorter time compared to listening to their performance one by one to evaluate. Therefore this loudness detection and visualisation has a clear use case scenario to support music education.

Regarding dynamic marking transcription problems, we must consider a map between MIDI velocity to perceptual loudness since dynamics markings are relative loudness and perceptual to some extent contrary to MIDI velocity which is absolute loudness. As a future work, it is important to create a map from MIDI velocity to the symbolic notations. There have been several researches to create maps from loudness to symbols of music score [14, 17]. However, this area of research needs interdisciplinary knowledge by collaborating musicologists since this is relative mapping seeing the context of loudness of performance. As can be seen, this research takes the core points to contribute various music technology areas.

However, this research is the first research utilising the DNN ability and thus there are various architectures of DNNs to try out borrowing ideas from recent advancement. There is room for further improvement by exploring different DNN architectures and increasing the amount of training data. For example, residual nets are known performing well in the U-net for source separation [22]. If we employ U-net architecture, it is possible to directly input wave forms. Also, the FiLM condition generator can be explored such as taking CNN blocks. Additionally, future work should also address the issue of unaligned audio, MIDI, and score information and the mapping of MIDI velocity to perceptual loudness and symbolic notations. The code and data used for this research can be provided upon request.

6. ACKNOWLEDGMENTS

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.

7. REFERENCES

[1] 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.

[2] L. F. Hamond, "The pedagogical use of technology-mediated feedback in a higher education piano studio: an exploratory action case study," Ph.D. dissertation, UCL (University College London), 2017.

[3] H. Kim, P. Ramoneda, M. Miron, and X. Serra, "An overview of automatic piano performance assessment within the music education context," 2022.

[4] 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, pp. 407–434, 2013.

[5] J. W. Kim and J. P. Bello, "Adversarial learning for improved onsets and frames music transcription," *arXiv preprint arXiv:1906.08512*, 2019.

[6] R. Kelz, M. Dorfer, F. Korzeniowski, S. Böck, A. Arzt, and G. Widmer, "On the potential of simple framewise approaches to piano transcription." [Online]. Available: http://arxiv.org/abs/1612.05153

[7] 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, publisher: Audio Engineering Society. [Online]. Available: https://www.aes.org/e-lib/browse.cfm?elib=20716

[8] 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.

[9] 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/

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

[15] 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.

[16] O. F. B. E. C. Kosta, K., "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.

[17] 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.

[18] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville, "FiLM: Visual reasoning with a general conditioning layer." [Online]. Available: http://arxiv.org/abs/1709.07871

[19] M. Miron, J. J. Carabias Orti, and J. Janer Mestres, "Improving score-informed source separation for classical music through note refinement," in *Müller M, Wiering F, editors. 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.* International Society for Music Information Retrieval (ISMIR), 2015.

[20] 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.

[21] E. Manilow and B. Pardo, "Bespoke neural networks for score-informed source separation," *arXiv preprint arXiv:2009.13729*, 2020.

[22] 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.

[23] 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.

[24] 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

[25] 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.

[26] T. Kwon, D. Jeong, and J. Nam, "Audio-to-score alignment of piano music using rnn-based automatic music transcription," *arXiv preprint arXiv:1711.04480*, 2017.