NOTE LEVEL MIDI VELOCITY ESTIMATION FOR PIANO PERFORMANCE

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

Marius Miron
Universitat Pompeu Fabra
marius.miron@upf.edu

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

ABSTRACT

Piano is one of the most popular music instruments. During the piano performance, loudness is an important factor for expressiveness, alongside tempo, changes in dynamics play with expectation, convey various emotions, and render expressiveness. Due to the polyphonic characteristics and with the goal of better analysing the expressiveness of performance of piano with multiple notes playing simultaneously, it is more useful to find loudness for each note than looking at accumulated loudness for a single time frame. Most of the research in this topic uses Non-negative Matrix Factorization (NMF) techniques to find note level loudness. In contrast, we propose to use Deep Neural Networks (DNNs) conditioned with score information to estimate the loudness based on MIDI velocity for each note performed by piano players. To our best knowledge, this is a novel research for note level MIDI velocity estimation by a DNN model in end to end fashion having FiLM conditioning.

1. INTRODUCTION

Loudness is one of the most important aspects of music performance. It is considered a component of expressiveness and it renders changes in the perceived dynamics [1]. There is also research to find a bridge between MIDI velocity and loudness [2] and it is considered as an indicator for the loudness of performance [3].

There is substantial work done on mapping from audio to MIDI velocity on note level for piano performance [4–7]. These researchers applied an NMF method to separate a piano performance audio to the 88 piano keys and estimated a MIDI velocity on each note, together with score information.

We consider this area of research as an extension of expressiveness modelling and piano performance transcription since input audio must be classified to 88 keys of piano and regressed to the MIDI scale.

In this research, we conducted experiments to estimate MIDI velocity using a novel end-to-end approach, a DNN utilising FiLM layers [8] for conveying score information into the DNN.

2. METHOD

We employed the piano performance transcription model previously used to classify the audio into 88 notes [9]. However, this model does not evaluate its estimation of MIDI velocity accuracy but cares about classification of audio to each key. We modified the model and the model architecture for this research is illustrated as figure 1.

In order to take advantage of the classification characteristics of this network, we used Binary Cross Entropy (BCE) as the loss function. On top of the BCE loss, we added an $l1$ loss function to estimate MIDI velocity which takes $l1$ distance between the output MIDI velocity from the model and ground truth MIDI velocity where note onsets occur only. The employed loss function 2 is $l1$ loss and BCE loss connected by a parameter so that we can back propagate losses for both classification and regression. The loss function is defined as following;

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

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

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

$$l1\ loss = \frac{\sum_i |V(i)_{\text{gt}} - V(i)_{\text{e}}|}{N} \tag{2}$$

where i is an index of each data point of notes corresponding MIDI roll as ground truth (gt) and estimation output (e) within an input frame. N is the number of all data points in a MIDI roll. The frame for one data point is two seconds and each frame contains 100 segments per second to represent MIDI roll.

We added a FiLM conditioning in order to insert a score information. FiLM is a fully connected linear layer to create parameters for affine transformation on an arbitrary layer of the DNN which gives an output for inference [8]. This idea has been applied to audio source separation tasks by adding video and score information from the FiLM network [10]. We employed a FiLM conditioning architecture from the research. We inserted the FiLM layer for applying parameters to each convolution block as in the figure 1. In this research the FiLM generator is designed by a fully connected layer to generate conditioning parameters.

Late-Breaking / Demo Session Extended Abstract, ISMIR 2022 Conference

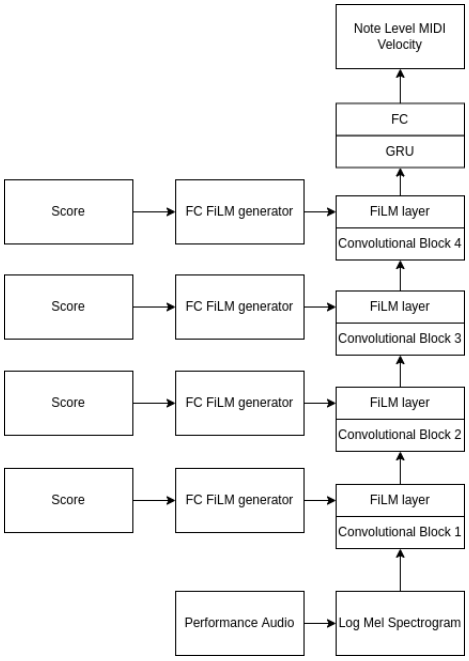

Figure 1. The model architecture for score informed MIDI velocity estimation

We used the Maestro dataset [11] for training and the SMD dataset [12] for the testing purposes. The amount of 2.8 GB of data for both audio and MIDI data for 132 excerpts is selected from the Maestro dataset for corresponding to maestro mini for training. This setup is for speeding up the experiment process and we also did not have a computation resource to train with the full dataset. The SMD dataset contains performed notes and its MIDI velocity. We compare solely the note onset frame, i.e. MIDI velocity on estimated note onset and MIDI velocity at the same point are compared. Furthermore, we use perfectly aligned score as input to the film layers. The evaluation is made by taking the $l1$ distance of MIDI velocities between ground truth and inference by the model.

$$Error = \frac{\sum_j |V(j)_{\text{gt}} - V(j)_{\text{inf}}|}{N} \tag{3}$$

where j is each note onset frame and inf stands for inference of the model and N is number of notes in a score. This error equation 3 is taken on each note at where onset happens in the ground truth MIDI roll, i.e. the inference output is masked by an onset roll of the ground truth in order to utilise the given score information.

3. RESULTS

The Table 1 is the preliminary result of this experiment.

As we can see from the table 1, the error of mean value ranged from 10.83 to 20.07 and the average of the error is 14.23 on the scale of 0-128 of MIDI velocity. This result is not as good as the NMF model [4]. We expect the accuracy to improve after training on the whole dataset and post-processing the results.

Composer	Excerpt	Mean	SD
Bach	BWV849-01	13.68	12.58
Bach	BWV849-02	16.96	15.63
Bach	BWV871-01	17.1	16.78
Bach	BWV871-02	14.07	13.72
Bach	BWV875-01	19.01	17.23
Bach	BWV875-02	16.33	14.4
Beethoven	Op027No1-01	11.13	11.13
Beethoven	Op027No1-02	12.25	9.08
Beethoven	Op027No1-03	12.88	12.57
Beethoven	Op031No2-01	16.2	16.34
Beethoven	Op031No2-02	17.17	17.72
Beethoven	Op031No2-03	16.6	16.94
Brahms	Op010No1	11.21	11.94
Brahms	Op010No2	11.59	12.92
Chopin	Op010-03	11.27	8.81
Chopin	Op010-04	12.97	13.42
Chopin	Op026No1	13.22	12.31
Chopin	Op026No2	12.74	11.1
Chopin	Op066	10.83	8.76
Haydn	Hob017No4	15.4	15.52
Rachmaninov	Op039No1	20.07	19.24
Skryabin	Op008No8	10.44	8.13

Table 1. The mean and standard deviation (SD) of estimated MIDI velocity error towards ground truth on note onset level.

4. CONCLUSION AND FUTURE WORK

In this research, we looked into a novel method to estimate MIDI velocity by DNN using FiLM conditioning. This research has wide applications such as performance visualisation, music education, expressiveness markings transcription as in f, p, mf, $crescendo$, $decrescendo$, etc.

When it comes to music education applications, this model allows for visualising students' performance in terms of loudness. This gives students an objective way to see their performance. 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. In this use case, we need to take into account the unaligned case between MIDI/score and audio, considering not only tempo misalignment but also wrongly played notes such as missing notes and extra notes [13, 14].

When it comes to expressiveness marking transcription problems, we must consider a map between MIDI velocity to perceptual loudness since dynamic markings are relative loudness and perceptual to some extent contrary to MIDI velocity which is directly related to 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 [2, 15]. However, this area of research needs interdisciplinary knowledge by collaborating musicologists since this is relative mapping seeing the context of loudness of performance.

5. 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.

6. 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] R. B. Dannenberg, "The interpretation of MIDI velocity," p. 4.

[3] W. G. Busse, "Toward objective measurement and evaluation of jazz piano performance via midi-based groove quantize templates." *Music Perception:An Interdisciplinary Journal*, vol. 19, no. 3, p. 443–461, 2002.

[4] 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

[5] S. Ewert and M. Müller, *Estimating note intensities in music recordings*, pages: 388.

[6] 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/

[7] D. Jeong and J. Nam, "Note intensity estimation of piano recordings by score-informed NMF," p. 9.

[8] 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

[9] 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 Trans. Audio, Speech and Lang. Proc.*, vol. 29, p. 3707–3717, oct 2021. [Online]. Available: https://doi.org/10.1109/TASLP.2021.3121991

[10] O. Slizovskaia, G. Haro, and E. Gómez, "Conditioned source separation for music instrument performances," vol. 29, pp. 2083–2095. [Online]. Available: http://arxiv.org/abs/2004.03873

[11] 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

[12] M. Müller, V. Konz, W. Bogler, and V. Arifi-Müller, "Saarland music data (SMD)," 2011.

[13] S. Wang, S. Ewert, and S. Dixon, "Identifying missing and extra notes in piano recordings using score-informed dictionary learning," vol. 25, no. 10, pp. 1877–1889. [Online]. Available: https://ieeexplore.ieee.org/document/7971931/

[14] S. Ewert, S. Wang, and M. Sandler, "SCORE-INFORMED IDENTIFICATION OF MISSING AND EXTRA NOTES IN PIANO RECORDINGS," p. 7.

[15] K. Kosta, R. Ramirez, O. F. Bandtlow, and E. Chew, "Mapping between dynamic markings and performed loudness: A machine learning approach," p. 24.