DATA AUGMENTATION AND MODEL OPTIMIZATION FOR PIANO TRANSCRIPTION

*Sangeon Yong**

KAIST
Graduate School of Culture Technology
koragon2@kaist.ac.kr

*Changhyun Kim**, *Jiwon Kim*

SK Telecom
AI Center, T-Brain
changhyk@sktbrain.com, jk@sktbrain.com

ABSTRACT

We present two data augmentation methods that are suitable for an automatic piano transcription. After deep learning-based approaches are applied to the diverse music information retrieval problems, the performance of automatic piano transcription is also improved. However, a lack of dataset with various music genres causes the difficulty in model generalization. To solve this problem, we analyzed the two major piano transcription dataset and suggest data augmentation methods in both symbolic and audio domain based on Onsets-and-Frames architecture. Also, to maximize the performance, we tried the meta learner system to find out the best hyperparameters, and used model ensemble.

***Index Terms*—** Piano, piano transcription, onsets-and-frames, mixup, stochastic data augmentation, multipitch, music

1. INTRODUCTION

Automatic music transcription (AMT) is a task that find out multiple musical notes in polyphonic music with various kinds of instruments. It is considered as the one of the most difficult task in music information retrieval (MIR) because of the interference among notes.

Piano transcription is a specific subtask of AMT, which finds notes from audio with piano only. Among early researchers, nonnegative matrix factorization (NMF) was usually used for AMT [1]. After deep learning-based approaches started to get noticed, convolutional neural networks (CNN) [2] and recurrent neural networks (RNN) [3] are used to improve AMT performance. Recently, Google suggests Onsets-and-Frames model [4] that uses both convolutional neural networks and recurrent neural networks to find onsets and frames to predict musical notes.

2. DATASET

In piano transcription, we need a pair of audio and symbolic labels for each song. However, it is very difficult to make audio-midi paired data with real audio, and this is one of the major obstacles to make a fine piano transcriber. Recently, MAPS [5] and MAESTRO [6] datasets which contain aligned MIDI data with disklaviers to help MIR researchers, but there is still a problem that most songs are classical music.

2.1. MAESTRO

MAESTRO dataset [6] is a piano performance dataset with over 200 hours of recorded audio and aligned note labels. The songs in the MAESTRO dataset are from the International Piano-eCompetition, which recorded the performance with Yamaha Disklaviers that captures note onset, offset, velocity, and pedal information.

2.2. MAPS

MAPS database [5] is a piano dataset with real and virtual audio. There are four sets in MAPS, and one of them called the MUS set is a set with pieces of music recorded by disklavier. In the MUS set, there are about 238 pieces of classical and traditional music recordings.

Because the MUS set is recorded by disklavier, it provides an exact pair of MIDI and audio. However, because the MUS set is recorded by MIDI playing, the notes are quantized and all of them have the same velocity.

3. PROPOSED METHOD

To improve Onsets-and-Frames piano transcriber, we applied two different data augmentation methods for audio and symbolic domain. In the symbolic domain, we applied stochastic data augmentation method to create augmented scores, and synthesized audio through soundfont. Four sound fonts are used for synthesizing the piano sounds including YDP-Grand Piano, Kawai Stereo Grand, Kawai Upright Piano, and Steinway Grand Piano.

In the audio domain, we applied mixup [7] in melspectrogram domain while training to make generalized model in audio domain.

* Equally contributing authors.

Test Dataset	Best Model	Multi-domain				Single-domain			
		Single Model		**Ensemble-Mean**		**Single Model**		**Ensemble-Mean**	
		onset only	onset + offset	onset only	onset + offset	onset only	onset + offset	onset only	onset + offset
MAPS	onset best	0.837	0.603	**0.846**	0.652	0.843	0.628	0.845	**0.653**
	onset-offset	0.832	0.629			0.830	0.624		
MAESTRO	onset best	0.949	0.789	0.954	0.833	0.952	0.807	**0.956**	**0.836**
	onset-offset	0.951	0.816			0.955	0.819		
ALL	onset best	0.917	0.749	0.919	0.775	0.917	0.749	**0.920**	**0.777**
	onset-offset	0.912	0.753			0.914	0.756		

Table 1. Comparison results of single models and ensemble models.

3.1. Stochastic Data Augmentation

Three data augmentation methods are applied in the symbolic domain: key transition, key change, and tempo change. We selected the integer values between 0 and 11 for the first two data augmentation methods and tempo change values are selected in between 0.5 and 1.5. All these values having uniform distribution are randomly selected. Especially, the lower limit, 0.5 value, makes the original music slower. In this step, we could additionally consider the note frequency, the data augmentation probability map, and timing of the key change.

3.2. Mixup Data Augmentation

To increase the accuracy of the model and avoid overfitting, we applied a novel data augmentation approach called mixup [7], which simply mixes the two randomly sampled inputs and labels. It is quite similar to the method called between-class learning [8], which is originally used for audio classification. However, because between-class learning is not appropriate for the multi-label problem, we applied mixup instead of between-class learning for the project.

3.3. Meta Learner

Two processes, both hyper parameter-tunning and neural architecture search, are performed in the meta learner. Totally eight parameters, which are categorized into three areas, are grid-searched through meta learner. The first group includes functional parameters including the weight of the mixup data augmentation and the adversarial loss. The second category parameters contain learning related values such as learning rate, learning rate decay rate, learning rate decay steps, and input sequence length. The last categorical parameters are related with the neural network architecture: the number of output in convolutional layers and the number of recurrent network output in the long short term memory (LSTM) layers. Hyper parameter tunning is experimented with these eight parameters. All the combinations of the hyperparameter sets are fully trained up until 150,000 iterations. The best twelve parameter sets for both MAPS and MAESTRO datasets are selected.

	Non-optimized		Optimized	
	onset	onset-offset	onset	onset-offset
MAPS	0.845	**0.653**	**0.852**	**0.653**
MAESTRO	0.956	0.836	**0.967**	**0.842**
ALL	0.920	0.777	**0.929**	**0.781**

Table 2. The result of threshold change in single-domain ensemble model.

3.4. Ensemble

To increase the performance of the network, we chose the best models and applied the ensemble. For choosing the best models, we used two strategies: multi-domain and single-domain. For multi-domain, we chose models that shows the best performance in onset f1 score and onset-offset f1 score for MAPS, MAESTRO, and both. For single-domain, we chose top three models that shows the best performance each in the onset f1 score and onset-offset f1 score for both datasets.

3.5. Threshold Optimization

To find the best threshold for the onset network and frame network to get better results, we tried both threshold from 0.1 to 0.9 with interval size 0.1 in MAESTRO validation set. Most networks showed the best result with onset threshold 0.3 and frame threshold 0.5.

4. RESULTS

For evaluation, we measured f1 score for onsets and offsets of both MAPS and MAESTRO dataset with the best single model and ensemble model for each domain. As shown in Table 1, ensemble models show the best results for all 6 metrics, and single-domain ensemble model shows the best results for 5 results except onset f1 score for MAPS.

The result of models with optimized threshold is written in Table 2.

5. CONCLUSION

In this paper, we proposed the data augmentation method for piano transcription in both audio and symbolic domain. We showed that our approach is effective for Onsets-and-Frames based network to find more accurate musical notes. As future work, we aim to add more recent data augmentation methods like manifold-mixup [9], and have a plan to make own dataset contain non-classical music for model generalization for more diverse genres.

6. ACKNOWLEDGEMENTS

This research is supported by the SKT T-Brain Meta Learner team. We would like to thank Sungwan Kim and Youngsoon Lee for their careful coordinating the Meta Learner resources.

7. REFERENCES

[1] Paris Smaragdis and Judith C Brown, "Non-negative matrix factorization for polyphonic music transcription," in *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2003, pp. 177–180.

[2] Rainer Kelz, Matthias Dorfer, Filip Korzeniowski, Sebastian Böck, Andreas Arzt, and Gerhard Widmer, "On the potential of simple framewise approaches to piano transcription," in *Proceedings of the 17th International Conference on Music Information Retrieval (ISMIR)*, 2016.

[3] Siddharth Sigtia, Emmanouil Benetos, and Simon Dixon, "An end-to-end neural network for polyphonic piano music transcription," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 2016, vol. 24(5), pp. 927–939.

[4] Curtis Hawthorne, Erich Elsen, Jialin Song, Adam Roberts, Ian Simon, Colin Raffel, Jesse Engel, Sageev Oore, and Douglas Eck, "Onsets and frames: Dual-objective piano transcription," in *Proceedings of the 19th International Conference on Music Information Retrieval (ISMIR)*, 2018.

[5] Valentin Emiya, Roland Badeau, and Bertrand David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," in *IEEE Transactions on Audio, Speech, and Language Processing*, 2019, vol. 18.6, pp. 1643–1654.

[6] Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck, "Enabling factorized piano music modeling and generation with the MAESTRO dataset," in *International Conference on Learning Representations*, 2019.

[7] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*, 2018.

[8] Yuji Tokozume, Yoshitaka Ushiku, and Tatsuya Harada, "Learning from between-class examples for deep sound recognition," in *International Conference on Learning Representations*, 2018.

[9] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, Aaron Courville, David Lopez-Paz, and Yoshua Bengio, "Manifold mixup: Better representations by interpolating hidden states," in *Proceedings of the 36th International Conference on Machine Learning*, 2019.