Explore the Feasibility of Acoustic Model Augmented Transformer for Multi-Task Music Transcription

Taehyeon Kim
Gwangju Institute of Science and Technology
Cheomdangwagi-ro, Buk-gu, Gwanju, Republic of Korea
esteban12@gist.ac.kr

Chang Wook Ahn*
Gwangju Institute of Science and Technology
Cheomdangwagi-ro, Buk-gu, Gwanju, Republic of Korea
cwan@gist.ac.kr

ABSTRACT

In Automatic Music Transcription, Neural Networks have made significant progress over the nonnegative matrix factorization. Then, a model in the form of a combination of CNN-based acoustic model and RNN-based language model has been actively used. Recently, the Text-to-Text approach enables the Transformer model to do multi-task learning for multi-instruments transcription. Inspired by the CNN-RNN structure, we propose a combination of the CNN-based acoustic model and the Transformer-based language model. By combining the acoustic model estimating probabilities for pitches locally and the language model evaluating the global correlation between the pitch combinations, we expect the additive effect of joining both global information and local information. In this work, we demonstrate the performance of the Acoustic-Transformer by comparing its performance with the pre-existing model MT3 on the Slakh2100 Dataset. We also propose a musical instruments integrated evaluation method to evaluate the output from each model and calculate their note metrics score.

KEYWORDS

Automatic Music Transcription, Multi-Task Learning, Text-to-Text Approach, Acoustic Model, Language Model

1 INTRODUCTION

Automatic Music Transcription(AMT)[1, 2] is a task that generates symbolic representations of music from acoustic signals. AMT is considered a core task in Music Information Retrieval(MIR) tasks. This is because AMT connects audio-based music representations with symbolic-based music representations and extends music information to be computationally addressed. In one acoustic signal, plenty of sounds with various musical instrument timbres are overlapped simultaneously to form a harmony. For these reasons, AMT differs from Automatic Speech Recognition(ASR) in that it must be capable of Multi-Tracking and Multi-Tasking. At this time, Multi-Tracking refers to the ability to simultaneously transcribe various instruments, and Multi-Tasking refers to the ability to translate various genres of music.

Due to the lack of datasets and the limitation of domain-specific models, Multi-Task and Multi-Track Learning in AMT tasks have been a challenging problem for a long time. However, a large number of datasets such as Slakh2100[3] and MAESTROv3[4] have recently come out, and a Text-to-Text approach that does not require specific domain knowledge using Transformer has been proposed[5]. Therefore, the MT3[6] applies a multi-taskable Transformer to AMT, and multi-tasking became easier in AMT tasks. On the other hand, while transformers have the advantage of obtaining the global context over the attention mechanism, CNN has the advantage of obtaining the local information with shared position-based kernels over a local window.

In this work, we propose an Acoustic-Transformer model that consists of the CNN-based acoustic model and the Transformer-based language model. It concurrently obtains local information for estimating probabilities for pitches and global information for the evaluating the correlation between the pitch combinations. Then, we demonstrate the effectiveness of the acoustic model and the performance of the proposed models by comparing their performance with the MT3.

2 Related Work

2.1 Previous Approach in Automatic Music Transcription

Early AMT research was studied on a single musical instrument, led by Piano Transcription. Non-negative matrix factors(NMFs) were used as a popular method for performing music transcription by decomposing spectrograms into polyphonic notes[7]. Recently, various neural networks such as fully connected neural networks (FCNN), convolution neural networks (CNN), and recurrent neural networks (RNN) have been applied to AMT[8-11]. As AMT Tasks became considered a sequence-to-sequence tasks, the RNNs began to be applied to AMT. CNNs were also recognized for their performance in MIR tasks. Therefore, a model that integrates RNN to CNN has been attempted to be used as an AMT model namely CNN-RNN structures. Complex domain knowledge was applied and

*corresponding author

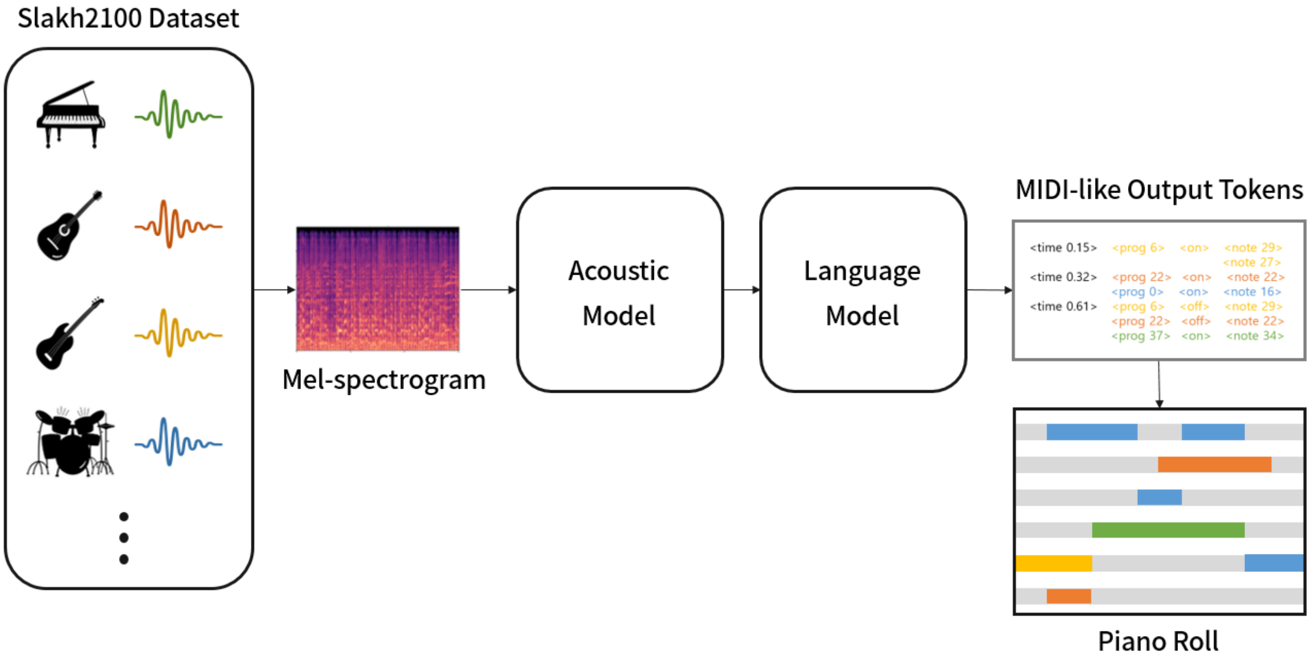

Figure 1: Acoustic-Transformer is possible to transcribe an arbitrary number of instruments from Mel spectrograms of raw audio. It uses both the acoustic model and language model to output MIDI-like output tokens. Shown here is the overall process of transcription with Slakh2100 which is composed of arbitrary instrument types.

developed to deep neural network architectures and decoding processes[10-14]. Onsets and Frames model[12] is a further advanced model that is CNN-LSTM-based dual objective system to predict onset and frame output.

2.2 Acoustic Model and Language Model in Automatic Music Transcription

The Acoustic Model is a model to estimate probabilities for pitches present in each frame of the audio waveforms. When converting audio signals to input representations during the AMT task's process, the audio sequence is often converted to the form of two-dimensional time-frequency representations. Therefore, CNN started to be applied to AMT Tasks. Until recently, it has been actively used as an Acoustic Model in combination with the Language Model[12-14].

Language models are actively used in NLP tasks such as Machine Translation, Spell Correction, etc. Language models assign probability $P(w_1, w_2, \ldots, w_m)$ to a word sequence of length m. On the other hand, AMT Tasks may be regarded as a sequence-to-sequence tasks with sequences of music audio frames as input and sequences of symbolic representations of music as output[5, 8, 12]. Therefore, since the pitches occur highly correlated with the progression of harmonies or chords in music sequence, the language model plays a role in sequentially evaluating the correlation between pitch combinations. Language Models such as RNN, LSTM, and Transformers were used as core components in AMT tasks and led to the development of AMT tasks.

2.3 Text-to-Text Approach in Automatic Music Transcription

The Transformer[15] has recently demonstrated its high performance in sequence-to-sequence problems across multiple domains. After the advent of the Transformer, various models using Transformer's Encoder and Decoder have been proposed[16, 17]. Among them, T5[18] is a single text-to-text encoder-decoder model that uses the basic structure of Transformers. Text-to-Text approach defined all text-based language problems in text-to-text format and achieved SOTA on multiple NLP tasks.

By applying the T5 model to AMT tasks, many changes were made in AMT tasks[5]. Symbolic representations of music were converted to text, and a method to perform AMT tasks with simple encoding/decoding methods without any applications of domain knowledge was proposed. Furthermore, by converting input data to text, a simple labeling process can easily be extended for various musical instruments. Based on the multilingual pre-trained text-to-text transformer MT5[19], the MT3[6] model, a unified framework capable of performing Multi-Track and Muli-Task learning on various datasets, has been proposed.

3 Model

An overview of model architecture is show in Fig.1. Our model uses a pre-trained T5-small model, and a CNN-based acoustic model from [12]. For the mel-spectrogram of input

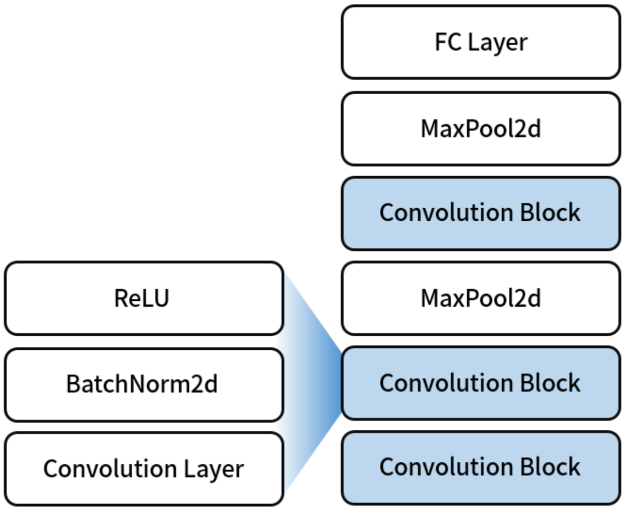

Figure 2: The architecture of Acoustic Model used in our model. Each Convolution Block is composed with Conv2d, BatchNorm2d, ReLU function.

audio, our models concurrently estimate probabilities for pitches and the correlation between the pitch combinations. It enables our model to obtain local information and global information in the time-frequency representations of the input data. Then, it produces the MIDI-like outputs tokens as outputs.

3.1 Model Architecture

The structure of the CNN-based acoustic model applied in this work is shown in Fig.2. Each convolution layer has a kernel of $3 \times 3 \times 21$ sizes in common, and the length of both output and input are adjusted equally through zero padding. We integrated the convolution layer, batch normalization, and rectified linear unit(ReLU) and defined them as one 'convolution block' left for future works. Stacking convolution blocks with max pooling layer and fully connected layer, the acoustic model was completed. The acoustic model used in our model consists of 3 convolution blocks and has a total of 4.2M parameters.

Recently, there is a tendency to use very large models in various domain studies. However, in the case of AMT Task, sufficient learning was achieved with a T5-small model. Besides, [5] showed that using a larger T5 models has the riskiness of overfitting. Therefore, we used T5-small model with 1024 feed-forward output dimensionality(d_{ff}), 64 key/value dimensionality(d_{kv}), 512 embedding size(d_{model}), 6 head attentions, 8 layers in each encoder and decoder, and 0.1 dropout ratio for sub-layer outputs and embedded inputs same as [6]. Additionally, absolute positional embedding was applied to ensure the same resolution for all locations.

3.2 Model Inputs and Outputs

The model receives time-frequency representations in the form of Mel spectrogram as the input. In the process of converting audio to Mel spectrogram, we used a sample rate of 16,000Hz, FFT window length of 2048 samples, and hop length of 128 samples. 512 mel scale bins were applied to the model according to the models' embedding size. As a result, input sequences with total of 512 positions are composed of 511 spectrogram frames and one EOS embedding.

Our model produces a MIDI-like outputs token which contains a subset of the original MIDI specification[20]. Each segment is calculated independently by dividing the audio by non-overlapping segments quantized with an interval of 10ms. [6] proposed this method to improve the limitation of too large to fit in memory that has the existing transformer-based sequence model. The vocabularies of MIDI-like outputs token consists of the following token types.

Table 1: Token types composing the vocabularies of MIDI-like outputs token

Token types	Explanation
Instrument (128 values)	Represents distinct values that matched with General MIDI specifications for designating specific instruments.
Note (128 values)	Represents note events for the MIDI pitches
On/Off (2 values)	Represents changes of subsequent note events.
Time (205 values)	Represents absolute time location of events within a segment.
Drum (128 values)	Represents a drum onset distinct values that matched with General MIDI standard.
End Tie Section (1 value)	Represents to end tie section.
EOS (1 value)	Represents the end of a sequence.

In Table 1, General MIDI is standardized specification for the 128 distinct programs used to designate specific instruments. The 'End Tie Section' is a token to prevent notes spanning multiple segments. It is located at the beginning of each segment and indicates which notes are already active. By using token types in Table 1, vocabularies are defined to enable general learning of different musical instruments or datasets. It was the key contribution of the [6], we also actively adopted these segmentation methods and vocabularies.

4 Experiments

We used AdamW optimizer with a batch size of 4. In the [5], they trained the T5 model using 32 TPU v3 cores, while we used 4 RTX 2080Ti GPUs only. By allocating 8 batches for each core, they were possible to use a total of 256 batch sizes. However,

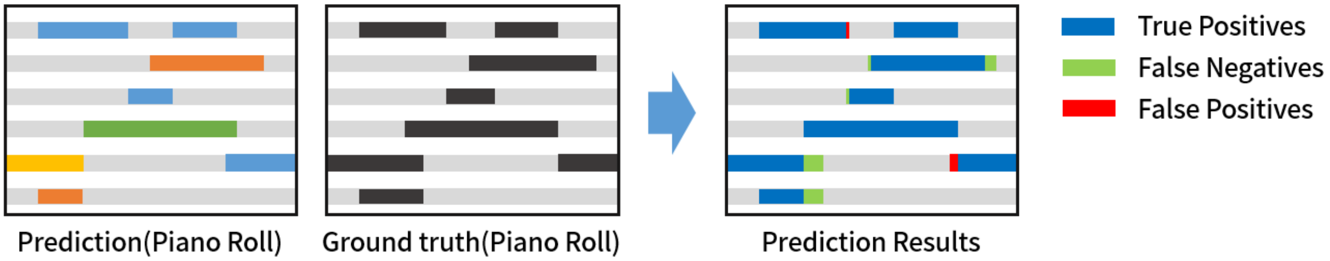

Figure 3: Estimation of the piano roll results of the model predictions with TP(True Positives), FN(False Negatives) and FP(False Positives). TP(True Positives) are the number of the correct predictions of the positive class and TN(True Negatives) are the number of the correct predictions of negative class. Similarly, FP(False Positives) are the number of incorrect predictions of the positive class and FN(False Negatives) are the number of incorrect predictions of the negative class.

considering the huge size of the dataset and the limitations of computing resources, we set hyperparameter values that can be stably trained within the limited RAM size of the training environment. We trained a model in 500K steps with a learning rate of 1e-3 in the main experiment. We used the Slakh2100 dataset to evaluate our model's multi-task performance for multiple instruments and the ddsp library to preprocess the input spectrogram. Further, we did a additional experiment with 50K and 150K training steps to find out the appropriate value of the learning rate.

4.1 Datasets

We selected the Slakh2100 dataset for evaluating the multi-tasking and multi-tracking performance of our model with a single dataset. The Slakh2100 dataset is a dataset provided for music source separation(MSS) and multi-instrument AMT. It has 2100 automatically mixed-track audios with instruments set arbitrarily consisting of musical instruments, totaling 145 hours of audio data with 104.3GB. It consists of a training set of 1500 tracks, a validation set of 375 tracks, and a test set of 225 tracks.

4.2 Evaluation

$$Precision = \sum_{t=0}^{T-1} \frac{TP[t]}{TP[t] + FP[t]} \quad (1)$$

$$Recall = \sum_{t=0}^{T-1} \frac{TP[t]}{TP[t] + FN[t]} \quad (2)$$

$$F1\ score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (3)$$

For numerical comparison between the inference results and ground truth values, we calculated the precision, recall, and F1 score for the piano roll. We used the sklearn library, and the formulas of each metrics score are represented above (1), (2), (3). Precision is the number of true positive results divided by the number of all positive results, Recall is the number of positive results divided by the number of all positive results, and F1 score is the harmonic mean of the precision and recall. In this work, each frame-based score was evaluated integrally for all musical instruments without distinguishing them. Through this method, it was intended to conduct multi-taskable performance evaluation, not transcription performance evaluation for a specific instrument. The overall evaluation progress is shown in Fig.3.

4.3 Results

4.3.1 Main Experiment

Table 2: Transcription metrics scores(F1 score, Precision, Recall) for main experiment with 500K iterations

Model	Acoustic-Transformer(ours)	Original-Transformer(MT3)
F1 score	**0.300**	**0.598**
Precision	0.382	0.701
Recall	0.258	0.526

The main experiment was conducted with a learning rate of 1e-3 and 500K training steps. In Table 4, the results of our model were lower than previous model. Contrary to expectations, the acoustic model used in our model did not show a performance improvement. The value of F1 score of our model was 0.300, which was half of the 0.598. It became curious whether the learning rate of 1e-3 would be an appropriate value for our model's learning.

4.3.2 Supplementary Experiment

Because the complexity of the model was increased by adding a CNN-based acoustic model, we considered the lower learning rate for stable loss descent in training progress. Therefore, we conducted a additional experiment to determine whether the learning rate should be maintained at 1e-3 or adjusted to 1e-4. We trained the models with 50K steps and 150K steps, and compared the results of each model with learning rates of 1e-3 and 1e-4.

Table 3: Transcription metrics scores(F1 score, Precision, Recall) for supplementary experiment with learning rate 1e-4

Model	Acoustic-Transformer(ours)			Original-Transformer(MT3)		
Iteration	0K	50K	150K	0K	50K	150K
F1 score	0.483	**0.556** (↑)	0.478	0.484	**0.691** (↑)	0.678
Precision	0.624	0.728	0.654	0.615	0.781	0.776
Recall	0.399	0.456	0.385	0.404	0.624	0.607

Table 4: Transcription metrics scores(F1 score, Precision, Recall) for supplementary experiment with learning rate 1e-3

Model	Acoustic-Transformer(ours)			Original-Transformer(MT3)		
Iteration	0K	50K	150K	0K	50K	150K
F1 score	0.483	**0.172** (↓)	0.344	0.484	**0.619** (↑)	0.601
Precision	0.624	0.271	0.453	0.615	0.711	0.702
Recall	0.399	0.129	0.282	0.404	0.552	0.529

The Table 3 shows the metrics score evaluation results according to the number of training iterations of our model and the MT3 with learning rate 1e-4. When the models were trained with 50K steps, the F1 score was the highest. It means that the training steps of 150K was too much for the models, so the models were overfitted. Comparing the F1 scores of each model, the original transformer still has better performance.

The Table 4 shows the metrics score evaluation results according to the number of training iterations of our model and the MT3 with learning rate 1e-3. The original transformer still has better performance, but the most noticeable point is that our model's F1 score values are exceptionally low. It is understood that lower learning rate were required for the augmented acoustic model to be fitted into a pre-trained transformer. As a results, since 1e-4 works well for the models, we determined to use 1e-4 as future experiment's learning rate.

4.4 Discussions

In the main experiment, the performance of our model is lower than the previous model. We took a supplementary experiment to determine whether the learning rate should be maintained or adjusted. In the supplementary experiment, we compared the learning rate according to 50K and 150K training steps. Since a learning rate of 1e-4 works well not only for the previous model but also for our model, we decided to conduct the future experiment with a learning rate of 1e-4.

Moreover, [6] showed that overall onset+offset F1 score with the mixed instuments dataset was relatively low to the single musical instrument dataset. The reason why the results trained with mixed instrument datasets are low might be related to the contents of the dataset. Since each track of the Slakh2100 is consists of random musical instruments, the 2100 samples in the dataset are distributed for each musical instrument. Therefore, some specific musical instruments might have a deficient number of tracks. For these musical instrument, it causes the model to be overfitted, but not to be generalized.

Thus, it will be better to compare the performance for the evaluating effectiveness of the acoustic model with the specific instrument such as Piano transcription. Then it will be possible to use a sufficient dataset and get more accurate results for each model. After that, It could be extended to the multi-task music transcription with several datasets.

Furthermore, a new structure of Accoustic-transformer model can be explored. In our model, the acoustic model located in front of the pre-trained language model subsampled the input spectrogram of audio data. However, it seems that the extracted features from the acoustic model do not help the language model match the input spectrogram to midi-like outputs. Even if the acoustic model is fully trained for the dataset, our model's performance is lower than the previous model. From the information theory perspective, it can be understood with a Information Bottleneck(IB) problem[21]. Since the language model processes inputs from the acoustic model, it can be considered as a succesive Markov chain and related to refinement of information. For the nueral networks, each layer can increase the IB distortion level and compress its inputs. In our model, the acoustic model increased the IB distortion level and compressed the input information. Therefore, the acoustic model rather hindered the learning of our model.

In order to continue to use the acoustic model, we should apply another training method for the acoustic model and language model. To avoid the interruption of the acoustic model to the language model, training both models concurrently and concatenating them can be a candidate for the solution. Also, using pre-trained MT3 from [6] can be another candidate for the solution. Then, our model will be better to match the input and the MIDI-like output tokens than using T5 model pre-trained on language dataset. On the other hand, there is a method for performance improvement by combining CNN and self-attention in the transformer[22-26]. [22] proposed the method which does not just attach the CNN module before the transformer, but added a CNN module inside the encoder of the transformer. Unlike other methods, this method is expected to be applicable in our work, because it achieves performance improvement in the speech recognition task which is a similar task to AMT.

5 Conclusion

In this work, we introduced Acoustic-Transformer for end-to-end multi-task AMT by integrating the CNN-based acoustic model and Transformer-based language model. We checked the effectiveness of the CNN-based acoustic model in performing multi-task AMT through comparison with the MT3. We also proposed a musical instrument integrated evaluation method on our model, using the metrics scores such as F1 score, Precision, and Recall. Additionally, we checked the appropriate learning rate for the models with the supplementary experiments.

Although this study failed to achieve a performance improvement, there are many previous studies that achieved performance improvement by combining convolutional layers and transformers. We suggested some problem solutions for our model, and it is expected that the performance of transformer could also be improved through various attempts with CNN modules. Thus, this work will be continued to combine CNN to self-attention in Transformer used in [5, 6] and improve the performance of transcription.

ACKNOWLEDGMENTS

This work was partly supported by the Technology development Program of MSS (S2962806) and the IITP grant funded by the Korea government (MSIT) (No.2019-0-01842, AI Graduate School Program (GIST)).

REFERENCES

[1] C. Raphael, “Automatic transcription of piano music,” in International Society for Music Information Retrieval (ISMIR), 2002.

[2] E. Benetos, S. Dixon, Z. Duan, and S. Ewert, “Automatic music transcription: An overview,” IEEE Signal Processing Magazine, vol. 36, no. 1, pp. 20–30, 2018.

[3] Ethan Manilow, Prem Seetharaman, and Bryan Pardo. Simultaneous separation and transcription of mixtures with multiple polyphonic and percussive instruments. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 771–775. IEEE, 2020.

[4] 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. arXiv preprint arXiv:1810.12247, 2018.

[5] Curtis Hawthorne, Ian Simon, Rigel Swavely, Ethan Manilow, and Jesse Engel. Sequence-tosequence piano transcription with Transformers. arXiv preprint arXiv:2107.09142, 2021..

[6] Josh Gardner, Ian Simon, Ethan Manilow, Curtis Hawthorne, Jesse Engle. MT3: Multi-Task Multitrack Music Transcription. arXiv preprint arXiv:2111.03017, 2022

[7] K. O’Hanlon and M. D. Plumbley, “Polyphonic piano transcription using non-negative matrix factorisation with group sparsity,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 3112–3116.

[8] S. Böck and M. Schedl, “Polyphonic piano note transcription with recurrent neural networks,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012, pp. 121–124.

[9] S. Sigtia, E. Benetos, and S. Dixon, “An end-to-end neural network for polyphonic piano music transcription,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 5, pp. 927–939, 2016.

[10] R. Kelz, M. Dorfer, F. Korzeniowski, S. Böck, A. Arzt, and G. Widmer, “On the potential of simple framewise approaches to piano transcription,” in International Society for Music Information Retrieval (ISMIR), 2016.

[11] B. Liang, G. Fazekas and M. Sandler, "Piano Sustain-pedal Detection Using Convolutional Neural Networks," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 241-245

[12] 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. arXiv preprint arXiv:1710.11153, 2017.

[13] Q. Kong, B. Li, X. Song, Y. Wan, and Y. Wang, “Highresolution piano transcription with pedals by regressing onsets and offsets times,” arXiv:2010.01815, 2020.

[14] J. Engel, R. Swavely, L. H. Hantrakul, A. Roberts, and C. Hawthorne, “Self-supervised pitch detection by inverse audio synthesis,” in ICML Workshop on SelfSupervision in Audio and Speech, 2020.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in neural information processing systems, 2017.

[16] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners, 2019.

[17] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108, 2019.

[18] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text Transformer. arXiv preprint arXiv:1910.10683, 2019.

[19] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. arXiv preprint arXiv:2010.11934, 2020.

[20] MIDI Manufacturers Association and others, “The complete midi 1.0 detailed specification,” Los Angeles, CA, The MIDI Manufacturers Association, 1996.

[21] Naftali Tishby, Noga Zaslavsky “Deep Learning and the Information bottle neck principle” arXiv preprint arXiv: 1503.02406

[22] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented Transformer for speech recognition. arXiv preprint arXiv:2005.08100, 2020.

[23] I. Bello, B. Zoph, A. Vaswani, J. Shlens, and Q. V. Le, “Attention augmented convolutional networks,” in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 3286– 3295.

[24] B. Yang, L. Wang, D. Wong, L. S. Chao, and Z. Tu, “Convolutional self-attention networks,” arXiv preprint arXiv:1904.03107, 2019.

[25] A. W. Yu, D. Dohan, M.-T. Luong, R. Zhao, K. Chen, M. Norouzi, and Q. V. Le, “Qanet: Combining local convolution with global self-attention for reading comprehension,” arXiv preprint arXiv:1804.09541, 2018.

[26] Z. Wu, Z. Liu, J. Lin, Y. Lin, and S. Han, “Lite transformer with long-short range attention,” arXiv preprint arXiv:2004.11886, 2020.