ISBN 978-981-14-1684-2
Proceedings of 2019 the 9th International Workshop on Computer Science and Engineering
(WCSE 2019 SUMMER)
Hong Kong, 15-17 June, 2019, pp. 454-461
doi: 10.18178/wcse.2019.06.068

Onset-Aware Polyphonic Piano Transcription: A CNN-Based Approach

Sicong Kong [1], Wei Xu [1 +], Wei Liu [1], Xuan Gong [1], Juanting Liu [1], Wenqing Cheng [1]

[1] School of Electronic Information and Communications, Huazhong University of Science and Technology, China

Abstract. Automatic music transcription (AMT) transforms the musical audio content into symbolic notations, including onsets, offsets and pitches. In this paper, we designed a polyphonic piano transcription system based on Convolutional Neural Network (CNN), and it improves the note-level results. Our proposed method has two advantages: Firstly, A CNN model is used to detect the onset event and align the onsets of the notes into more accurate position. Secondly, the other CNN model is used to detect the onsets of 88 notes. And we improve the model's performance by using dual-channel spectrogram as input, appropriate number of convolution layers and the weights for the positive samples in loss function. The public dataset of MAPS is adopted to train and evaluate. Finally, in the 'ENSTDkCl' subset, our proposed solution achieves 85.15% on note-level F1-measure. To the best of our knowledge, the result is highest F1-measure scores in the state of art.

Keywords: polyphonic piano transcription, convolutional neural network, onsets detection, onset alignment

1. Introduction

Automatic music transcription(AMT) transcribes the musical audio signal into the symbolic representation, mainly including onsets, offsets and pitches. It is a fundamental problem in Music Information Retrieval(MIR) and has widespread applications, such as providing feedback to a piano learner in music education, searching songs with a similar baseline in content-based music search, improvising Jazz in musicological analysis of non-notated music, and visualizing the music content in music enjoyment. However there are challenging problems for the polyphonic AMT[1]. The parallel notes overlap in the time domain and interact in the frequency domain, which increase the complexity of the polyphonic signals. And piano which contains 88 keys or pitches is a typical polyphonic instrument, so there are also many comprehensive researches in the polyphonic piano transcription[2]. In this paper, we also focus on the polyphonic piano transcription.

The approaches to AMT can be divided into frame-based methods and note-based methods. The frame-based approaches will estimate the pitches in each time frame and then post-process the pitches' information of all frames to the note-level results. Spectrogram factorization techniques are the most popular methods. For example, non-negative matrix factorization(NMF)[3], probabilistic latent component analysis(PLCA)[4]. With the development of the machine learning(ML) and deep learning(DL), using discriminative approaches for spectrogram factorization and AMT has also become a common method. Poliner and Ellis use support vector machines(SVM)[5] to classify normalized magnitude spectra. Sigtia et al[6] compared the performance of three kinds of neural networks and proposed a Recurrent Neural Network(RNN) language model for music transcription. And Kelz et al.[7] described the potential of simple frame-wise approaches to piano transcription by using the Convolutional Neural Network(CNN). And the frame-based methods will only extract features for every time frame and ignore the context information, which is disadvantageous.

[+] Corresponding author: Wei Xu Tel: +86 18907152769
E-mail address: xuwei@hust.edu.cn.

The note-based approaches directly analyze the audio data and estimate the notes, mainly including the onsets and pitches and can be divided into two solutions. One solution of the note-based approaches is the time-series-based method, which will make use of the time-series information of the audio data. In early, Kameoka et al.[8] used harmonic temporal structured clustering(HTC) to estimate the attributes of the notes. Then Bock and Schedl[9] used an RNN with bidirectional long short-term memory(LSTM) to directly estimate the notes. And Hawthorne et al.[10] combined the CNN and RNN to directly estimate the onsets and improve the note-level results. The other solution is onsets-based method, which will directly detect the onsets and the corresponding pitches of the notes. The onsets-based approaches are usually used to improve the note-level results. And among the onsets-based methods, the most advanced methods are based on deep learning. Wang et al.[11] designed a two-stage CNN model to detect the common onset time and estimate the pitches at the onset time, respectively, while their CNN is too simple to recognize the pitches well. Liu et al.[12] used the multi-tasking CNN model to simultaneously detect the onsets and pitches, but the onsets are scattered and offset largely from the ground-truth.

In order to improve the note-level results of the AMT, we proposed a CNN-based framework which is the onsets-based method. Compared to existing onsets-based AMT approaches, the proposed method has the following two advantages:(1) We proposed an ***Onset-event*** model to detect the onset event. The two-class *Onset-event* is simple and have perfect performance, so it can select precise onset time range and align the scattered onsets into more precise position, which greatly improves the note-level performance. And the *Onset-event* makes the final note-level results aware of the accuracy of the onsets. (2) We designed a CNN model named *Onsets* which has 88 outputs to detect the onsets of the piano's 88 notes and we tried to improve the model's performance by using dual-channel spectrogram inputs, proper number of convolution layers and the weights for the positive samples in loss function. And finally our proposed method achieved better performance compared to the state-of-the-art approach and the optimal CNN-based approaches.

2. Proposed Model

The proposed transcription framework is shown in Fig. 1, and it can be divided into four parts: Extracting input features; *Onsets* detection; *Onset-event* detection and *Onsets alignment*. Firstly, the raw audio signal will be extracted into dual-channel spectrogram slice. Secondly, *Onsets* model will be used to detect the piano-rolls, including onsets and pitches. Its output is a 88-dimensional vector, and each dimension value represents the probability that a note onsets now, so the model can detect the pitches and onsets simultaneously or the piano-rolls. Thirdly, we added an *Onset-event* model, as shown in the *Onset-event* detection. The model output is a 1-dimensional vector, which indicates whether an onset event occurs. The time range selected by *Onset-event* can be used to align the onsets into a more accurate position. Finally, through our post-processing method called *Onsets alignment*, the results of the two CNN models are considered together, and the final piano-rolls with aligned onsets are obtained.

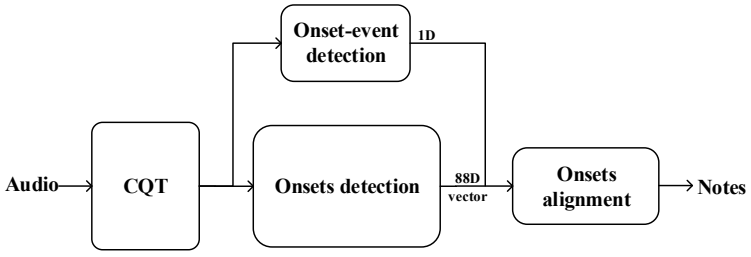

Fig. 1: Overview of the proposed framework. The extracted CQT features are fed to *Onset-event*, *Onsets* models. The piano-rolls can be obtained by *Onsets alignment* operations.

The following three subsections describe the operations to extract input features, the implementation details of the *Onsets* and *Onset-event* models and the details of the post-processing named *Onsets alignment*.

2.1. Input Features

In order to obtain the time-frequency representation, we perform constant Q transform (CQT)[13] on input audio signal. The pitch of the piano ranges from $A_0 \sim C_8$, and the corresponding fundamental

frequency range is 27.5 ~ 4186Hz. So we use bins ranging from $A_0 \sim C_8$ with a spacing of 48 bins per octave, for a total of 356 frequency bins. And our spectrogram frames are spaced 512 audio samples apart; For the audio of the input files, which has a sample rate of 44.1kHz, this leads to a time resolution of about 11.61ms.

The following operations will be used to extract the features for CNN models. The original audio is read into the left and right channels. Then we perform CQT and max-normalization on the two channel audio signals to obtain the dual-channel CQT spectrogram. Finally, we use a sliding window of 9 to take a spectrogram slice of several frames as a single input, yielding a total of 9 * 356 * 2 = 6408 input values.

2.2. CNN Models

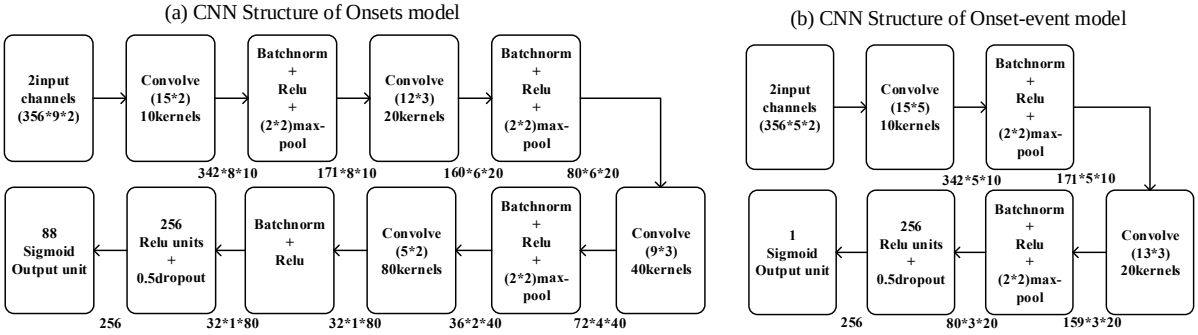

Fig. 2: CNN structure of the models

The *Onsets* model is used to detect the onsets of 88 notes, and the output of the 88-dimensional vector represents 88 notes of $A_0 \sim C_8$ of the piano. In the design of the input data, the dual-channel CQT spectrogram slice is used, just like the input data in subsection 2.1. In the design of the model structure, in order to fully extract the input data features and enhance the recognition ability of the model, a 4-layer convolution is selected. The model structure is shown in Fig. 2(a). In the loss function design, we use the sigmoid cross entropy loss function. However, considering the imbalance of the positive and negative samples of the training data, the positive sample data is weighted in the loss function, as shown in Equation 1:

$$L_{\text{onsets}} = \sum -l(i)logp(i) * pos\_\text{weight} - (1 - l(i))\log(1 - p(i)) \quad (1)$$

Where $\text{l} = l_{onsets}$ represents the label of the data, $\text{p} = predictions_{\text{onsets}}$ represents the predicted output of the *Onsets* model and pos_weight is the weights of the positive samples, which is used to alleviate the imbalance of the positive and negative samples. And pos_weight is proved to be 4 in the experiments.

The *Onset-event* model is used to detect onset event and is a two-class model. The onset event can be understood as the action of tapping the piano. The core function of the *Onset-event* is to align the note onsets from the *Onsets*. During the piano performance, multiple note keys are pressed at nearly the same time. With this feature, we can align the scattered onsets to a smaller time range, thereby improving the onsets detection results. We stipulate that the output of the *Onsets* is meaningful only when the *Onset-event* determines that there is an onset event, so that the wide range onsets of the notes will be aligned to a smaller time range. The input data is also dual-channel spectrogram slice. But the length of the sliding window has a significant influence on recognizing the onset event. Through experiments, we chose the optimal 5-frame window. In terms of model structure, due to the simple task, only 2 convolutional layers are used, as shown in Fig. 2(b). And the loss function uses the sigmoid cross entropy loss, as shown in Equation 2:

$$L_{onset-event} = \sum -l(i)logp(i) - \big(1 - l(i)\big)\log\big(1 - p(i)\big) \quad (2)$$

Where $\text{l} = l_{onset-event}$ represents the label of the data, $\text{p} = predictions_{\text{onset-event}}$ represents the predicted output of the *Onset-event* model.

2.3. Onsets Alignment

After getting the output of *Onsets* and *Onset-event* model, we get the piano-rolls with the aligned onsets by a post-processing strategy called *Onsets alignment*. The post-processing is mainly divided into three steps:

Step 1, selecting candidate data by *Onset-event* model:

$$candidates = \text{datas}(P_{onset-evt} > evt\_th) \tag{3}$$

In Equation 3, $P_{onset-evt}$ represents the predicted probability of the *Onset-event*, and evt_th represents the threshold. The input data whose output by the *Onset-event* is greater than the evt_th is the active candidate data. And the candidates will contribute to determine the pitches and onsets. As shown in Fig. 3(a), the horizontal axis represents time, and the vertical axis represents the probability value of $P_{onset-evt}$, thrA is the threshold. During time t1 to t3, $P_{onset-evt}$ > thrA, so the candidates are the data ranging from t1 to t3.

Step 2, determining the pitches of newly played notes in the candidate data:

$$P_{onsets} = \max\left(P_{onsets}^{candidates}\right) \tag{4}$$

$$pitches = P_{\text{onsets}} > onset\_th \tag{5}$$

In Equation 4 and 5$P_{onsets}^{\text{candidates}}$, $P_{onsets}^{candidates}$represents the output value of the *Onsets* with the candidate data as input, which is a matrix of N * 88. For each note, we select the maximum output probability as the onset probability, and then P_{onsets} represents the onset probabilities of 88 notes. $onset\_th$ represents the threshold for selecting pitches, and the pitches of notes whose onset probabilities P_{onsets} are greater than the threshold are the pitches of newly played notes in the candidate data. As shown in Fig. 3(b), horizontal axis is as same as the horizontal axis in Fig. 3(a), represents the time, while the vertical axis represents the outputs of the *Onsets*, and the three curves represents three pitches A, B, C. In time t1 to t3, the maximum probabilities, P_A, P_B, P_C represents the onset probabilities of their corresponding notes, and the threshold is thrB. From Equation 4 and 5, we can determine that the newly played pitches are A and B because P_A > thrB and P_B > thrB, while P_C < thrB.

Step 3, determining the onsets of the notes within the time range of the candidate data:

$$onset_{time} = max(where\_max(P_{onset-event}^{candidates}), where\_max\ (P_{onset}^{candidates})\ ) \tag{6}$$

In Equation 6, $where_$max represents the function used to find the position of the maximum value of the array. $P_{onset-event}^{candidates}$ represents the output values of the *Onset-event* in the candidate data and $P_{onset}^{candidates}$ represents the output values of the *Onsets* in the candidate data. The note onset is the maximum one between two positions. As shown in Fig. 3(b), *Onset-event* has maximum output value at time t2, while the curves of the newly played pitches A, B have the maximum value at time ta and tb respectively. From the Fig. 3(b), we can know ta < t2 and tb > t2, so the onset of the note A is t2 while the onset of the note B is tb.

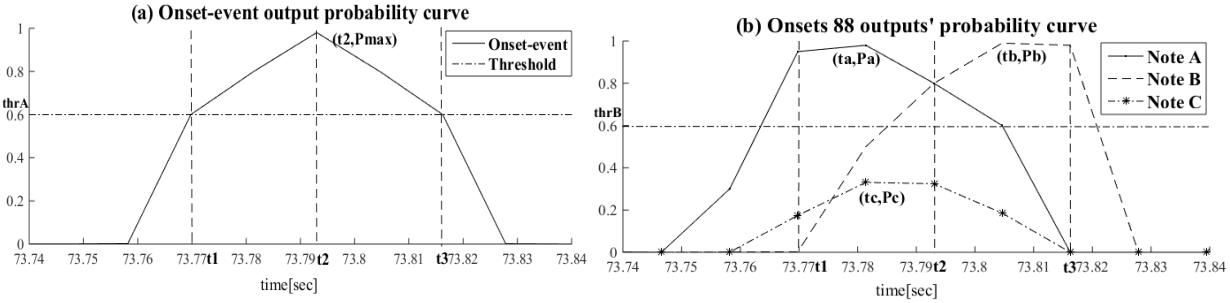

Fig. 3: Overview of *Onsets alignment* process

3. Experiment

In this Section, we will introduce the practical performance of our approach on polyphonic transcription. And the dataset used in the experiments, evaluation metrics will also be described in detail. All the following experiments are carried out using Tensorflow[14] and the spectrum transformation is carried out using

Librosa[15] library. Finally, we presented the results from the different experiments and analyze the performance of the proposed approach.

The MIDI aligned piano sounds(MAPS)[16] is the dataset on which the transcription experiments were conducted. According to the different piano types and recording conditions, the dataset can be divided into nine categories. Seven categories of audio are produced by software piano synthesizers while the other two('Cl' and 'Am') are obtained from a real Yamaha Disklavier upright piano. Sigtia et al.[6] introduced two kinds of configuration setting for the MAPS dataset. In configuration II, 210 tracks created using synthesized pianos are used to train and validation, and 60 audio recordings obtained from Yamaha Disklavier piano recordings are used to test. And this configuration is used by many subsequent researches which used the MAPS dataset[11][12][17]. In this paper, we will also use the configuration II setting for the experiments.

The precision, recall and f1-measure are both used for both frame and note based evaluation[18]. And the metrics are defined as:

$$P = \frac{N_{TP}}{N_{TP}+N_{FP}} \tag{7}$$

$$R = \frac{N_{TP}}{N_{TP}+N_{FN}} \tag{8}$$

$$F1 = \frac{2*P*R}{P+R} \tag{9}$$

In Equation 7, 8 and 9, P is precision, R is recall and F1 is the f1-measure which is a comprehensive score that considers the precision and recall. And N_{TP} is the number of true positives, N_{FP} is the number of false positives and N_{FN} is the number of false negatives. And mir_eval[19] library was used to calculate the note-based precision, recall and f1-measure.

3.1 Individual Model Results Analysis

For the *Onsets* model, it will detect the onsets and pitches of the notes simultaneously and the performance of the model needs to be improved. In order to explore the optimal *Onsets* structure, we explore the three aspects, the number of the input data channel, the number of the convolution layers and the positive sample weighting of the loss function. The results of the experiments for the *Onsets* is shown in Table 1, Fig. 4.

Table 1: Evaluation results of the onsets model under mono and dual channel conditions.

Channel	Precision	Recall	F1-measure
Single	78.32%	78.34%	78.33%
Dual	81.1%	80.95%	81.02%

As shown in Table 1, The dual-channel audio outperforms 2.69% over the mono-channel audio on the F1 metric. This is because dual-channel audio has more information than mono-channel audio, allowing the neural network to learn more useful features, thereby enhancing recognition results.

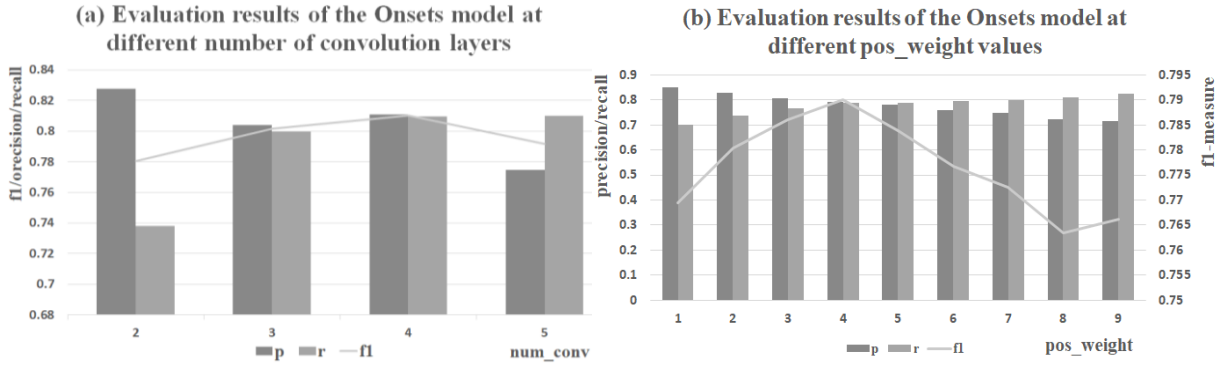

Fig. 4: Evaluation results of the *Onsets* model at different number of convolution layers and pos_weight value

As shown in Fig. 4(a), where the x-axis represents the number of the convolutional layers of the *Onsets* model, y-axis represents the values of precision, recall and f1-measure. The dark gray histogram p represents the precision, the light gray histogram r represents the recall and the gray curve f1 represents the f1-measure. From Fig. 4(a), when the number of convolution layers is 4, the *Onsets* model can obtain the optimal recognition result. When the number of layers is too small, the CNN is too simple and prone to under-fitting.

When the number of layers is too large, the model is too complicated and prone to over-fitting. As shown in Fig. 4(b), where the x-axis represents the the pos_weight value in Equation 1, left y-axis represents the precision and recall values and right y-axis represents the f1-measure. The dark gray histogram p represents the precision, the light gray histogram r represents the recall and the gray curve f1 represents the f1-measure. From the curve in Fig. 4(b), it can be known that the *Onsets* has the optimal recognition result when pos_weight=4. When the value is gradually increased, the recall is gradually increased, and the precision is gradually decreased. And when pos_weight=4, the f1-measure reaches the maximum. The pos_weight value mainly affects the balance between the precision and the recall of the results during training. When pos_weight=4, the optimal balance can be achieved.

Through the exploration of the structure of the *Onsets* model, the optimal structure is dual-channel audio as input, 4-layer convolution CNN structure, and pos_weight=4 in loss function.

Table 2: Evaluation results of the *Onset-event* under different window length.

Win-length	Precision	Recall	F1-measure
Win3	91.67%	89.78%	90.63%
Win 5	91.19%	91.36%	91.18%
Win 7	91.35%	90.60%	90.89%
Win 9	91.42%	90.15%	90.65%

For the *Onset-event* model, we mainly searched for the relation between the model performance and the length of the sliding window. Since the middle frame's time is on behalf of the input data and it will cause the ambiguity if the window length is even, we conducted the experiments with the window length 3, 5, 7, 9. And Table 2 show the results of the *Onset-event* in different window length. As shown in Table 2, the window length 5 performs best. The *Onset-event* uses CNN to detect the edges of the spectrogram and therefore it needs enough long frames of spectrums. When the window length is 3, it is too small and easy to omit the onset event, so the recall is a little smaller than the others. However, when the length is increased from 5 to 9, the precision remains similar, but the recall is gradually reduced. This is because the adjacent onset events is easy to be merged if the window length is too large. And from the experiment results, we concluded that the 5 is the optimum window length.

3.2. System Results and Comparative Approaches

First of all, we will show the import role of the *Onset-event* in detecting the piano-rolls. In Table 3, we show the results of the notes evaluation ignoring offsets. The median-filter uses only *Onsets* model's results and the same post-process strategy in Liu[12] to obtain the piano-rolls while the alignment method will use the *Onset-event* model to align the onsets. And it shows that the alignment method outperforms the median-filter method by 9.05%. Because the time range selected by the *Onset-event* is accurate and narrow, the alignment has excellent effects to align the onsets into precise position and can also exclude some wrong results.

Table 3: Evaluation results of median-filter post-processing and alignment post-processing.

	Precision	Recall	F1-measure
median-filter	75.15%	77.70%	76.10%
alignment	87.92%	82.82%	85.15%

Secondly, to illustrate the superiority of our approach, we compare with state-of-the-art approach and some optimal CNN-based methods. The requirement for evaluation is that the onsets must be within ±50ms of ground truth but ignoring offsets. All of the following methods are trained on MAPS synthesized audio, tested on real piano recordings, and final metric is the mean of scores calculated per piece in test dataset. And during inference, the threshold for *Onset-event*, *Onsets* is 0.5, 0.73 respectively. As shown in Table 4, our method outperforms the Wang[11] by 6.93% on f1-measure in the first 30s of the Cl dataset. This is due to the improvement of the performance of the *Onsets* by dual-channel inputs, optimal network structure and the appropriate pos_weight in loss function, and the carefully designed sliding window length for the *Onset-event* also plays an important role. Our method yields a relative improvement of 13.13% over Liu[12] in the Am and Cl dataset. Liu used only one CNN to detect the onsets and pitches simultaneously and it is a

complex multi-task model. Detecting the onsets of the multi-notes leads to the lack of constraints on the onset of each note, and so the deviation of the onset result is large. However, our model is more focused on onsets detection, by using the *Onset-event* model to align the onsets of multi-notes, resulting in a significant improvement on the results. On the Cl dataset, compared to Hawthorne[10], the state-of-the-art method, we also have an increase of nearly 1% on the f1-measure. Although both methods focus on the onsets detection, we use the alignment operation of the *Onset-event* model to align the onsets to a smaller time range, which is more suitable for the piano playing situation and leads to the slightly improvements.

Table 4: Evaluation results of our method, the state-of-the-art note-level method, and several optimal CNN methods

Method	Conditions	Precision	Recall	F1-measure
Ours	Am+Cl	86.32%	77.30%	81.36%
	Cl only	87.92%	82.82%	85.15%
	Cl only, first 30s	89.32%	85.48%	87.16%
Hawthorne[10]	Cl only	85.95%	83.05%	84.34%
Liu[12]	Am+Cl	-	-	68.23%
Wang[11]	Cl only, first 30s	85.93%	75.24%	80.23%

4. Conclusions

In this paper, we propose a framework for polyphonic piano transcription that improves the onsets detection performance. Firstly, a CNN model named *Onset-event* is designed to detect the onset event and align the onsets results into precise position, which greatly improve the performance. And this model makes our note-level results aware of the accurate onsets. Secondly, We improve the *Onsets* model by using dual-channel spectrogram inputs, appropriate number of convolution layers and the weights of positive samples in the loss function. Finally, our method improves the note-level results and achieves state-of-the-art performance.

5. Acknowledgements

This work is supported by The National Natural Science Foundation of China (No. 61877060).

6. References

[1] Klapuri, Anssi. "Introduction to music transcription." Signal Processing Methods for Music Transcription. Springer, Boston, MA, 2006. 3-20.

[2] Cogliati, Andrea, Zhiyao Duan, and Brendt Wohlberg. "Piano transcription with convolutional sparse lateral inhibition." IEEE Signal Processing Letters 24.4 (2017): 392-396.

[3] Gao, Lufei, et al. "Polyphonic piano note transcription with non-negative matrix factorization of differential spectrogram." Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA. 2017.

[4] Smaragdis P, Raj B, Shashanka M. A probabilistic latent variable model for acoustic modeling[J]. Advances in models for acoustic processing, NIPS, 2006, 148: 8-1.

[5] Poliner, Graham E., and Daniel PW Ellis. "A discriminative model for polyphonic piano transcription." EURASIP Journal on Advances in Signal Processing 2007.1 (2006): 048317.

[6] Sigtia, Siddharth, Emmanouil Benetos, and Simon Dixon. "An end-to-end neural network for polyphonic piano music transcription." IEEE/ACM Transactions on Audio, Speech, and Language Processing 24.5 (2016): 927-939.

[7] Kelz, Rainer, et al. "On the potential of simple framewise approaches to piano transcription." arXiv preprint arXiv:1612.05153 (2016).

[8] Kameoka H, Nishimoto T, Sagayama S. A multipitch analyzer based on harmonic temporal structured clustering[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2007, 15(3): 982-994.

[9] Böck, Sebastian, and Markus Schedl. "Polyphonic piano note transcription with recurrent neural networks." ICASSP. 2012.

[10] Hawthorne, Curtis, et al. "Onsets and frames: Dual-objective piano transcription." arXiv preprint

arXiv:1710.11153 (2017).

[11] Wang, Qi, Ruohua Zhou, and Yonghong Yan. "A two-stage approach to note-level transcription of a specific piano." Applied Sciences 7.9 (2017): 901.

[12] Liu, Shuchang, Li Guo, and Geraint A. Wiggins. "A Parallel Fusion Approach to Piano Music Transcription Based on Convolutional Neural Network." 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018.

[13] Brown, Judith C. "Calculation of a constant Q spectral transform." The Journal of the Acoustical Society of America89.1 (1991): 425-434.

[14] Girija, Sanjay Surendranath. "Tensorflow: Large-scale machine learning on heterogeneous distributed systems." (2016).

[15] Brian McFee, “librosa: v0.4.3,” May 2016.

[16] Emiya, Valentin, Roland Badeau, and Bertrand David. "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle." IEEE Transactions on Audio, Speech, and Language Processing18.6 (2010): 1643-1654.

[17] Li, Samuel. "Context-Independent Polyphonic Piano Onset Transcription with an Infinite Training Dataset." arXiv preprint arXiv:1707.08438 (2017).

[18] Bay, Mert, Andreas F. Ehmann, and J. Stephen Downie. "Evaluation of Multiple-F0 Estimation and Tracking Systems." ISMIR. 2009.

[19] Raffel, Colin, et al. "mir_eval: A transparent implementation f common MIR metrics." In Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR. 2014.