TRANSITION-AWARE: A MORE ROBUST APPROACH FOR PIANO TRANSCRIPTION

Xianke Wang *, *Wei Xu* †, *Juanting Liu*, *Weiming Yang*, *Wenqing Cheng*

Smart Internet Technology Lab
School of Electronic Information and Communications
Huazhong University of Science and Technology
Wuhan 430074, China
{M202072113, xuwei, juanting, M202072117, chengwq}@hust.edu.cn

ABSTRACT

Piano transcription is a classic problem in music information retrieval. More and more transcription methods based on deep learning have been proposed in recent years. In 2019, Google Brain published a larger piano transcription dataset, MAESTRO. On this dataset, Onsets and Frames transcription approach proposed by Hawthorne achieved a stunning onset F1 score of 94.73%. Unlike the annotation method of Onsets and Frames, Transition-aware model presented in this paper annotates the attack process of piano signals called atack transition in multiple frames, instead of only marking the onset frame. In this way, the piano signals around onset time are taken into account, enabling the detection of piano onset more stable and robust. Transition-aware achieves a higher transcription F1 score than Onsets and Frames on MAESTRO dataset and MAPS dataset, reducing many extra note detection errors. This indicates that Transition-aware approach has better generalization ability on different datasets.

1. INTRODUCTION

Piano transcription is the process of inferring onset time, offset time, and pitch of notes from the piano audio. Due to the limited modeling ability, traditional methods [1, 2, 3] based on nonnegative matrix factorization (NMF) undergoes slow development. Since 2012, the emergence of new transcription approaches represented by deep learning and a larger dataset, MAESTRO, has further promoted piano transcription technology evolution.

Before MAESTRO dataset [4] became available, MAPS dataset [5] had been widely used for piano transcription studies. MAPS dataset includes 210 pieces of synthetic audio and 60 real performances recorded from Yamaha Disklavier piano. There are some difficulties with studies based on MAPS dataset. The 210 training performances on MAPS dataset are synthesized using synthesizer software, and only 60 performances for testing are played on the Yamaha Disklavier piano. This results in a large audio characteristics gap between the training set and the test set. As a result, models trained on the training set often cannot be well generalized to the test set. In addition, as mentioned by Gong [6], there are still some errors in MAPS dataset annotations, which affects the evaluation of different models. MAESTRO dataset proposed by Google Brain in 2019 solves these problems. It contains 1,184 real performances of the International Piano-e-Competition, with 6.18 million notes and a total of 172.3 hours of audio. The entire audios of MAESTRO dataset are recorded from Yamaha Disklavier piano. Onsets and Frames transcription approach [4] achieves an F1 score of 94.73% on this dataset.

Kong has proposed the High-resolution model [7], which integrates the velocity information into the onset branch and frame branch of Onsets and Frames model. Meanwhile, High-resolution model uses a probability model to annotate the frames near onset time, which theoretically overcomes the time resolution limitation of hop_length in transcription systems. Besides, High-resolution model also takes the pedal prediction into account. On MAESTRO dataset, High-resolution model achieves an F1 score of 97.38%. Apart from the model's innovation, Kong has proposed GiantMIDI-Piano dataset [8], which is beneficial for music information retrieval and musical analysis.

The above methods have achieved good performance on MAESTRO dataset, and they are verified on MAPS dataset and OMAPS (Ordinary MIDI Aligned Piano Sounds) dataset in this paper. OMAPS dataset contains audio and video for the purpose of our further audio and video multimodal transcription research. More details about OMAPS dataset will be given in the experiment section. Compared with performance on MAESTRO dataset, the accuracy of Onsets and Frames model and High-resolution model decreases a lot on MAPS dataset and OMAPS dataset, revealing the lack of generalization ability. Therefore, the goal of this paper is to enhance the robustness of transcription models.

Onsets and frames annotates onset time with a single frame (Figure 1a). In contrast, High-resolution annotates onset with multiple frames, and the label value of each frame is subject to a specific probability distribution (Figure 1b), which implies that transition has been considered in the annotation of High-resolution model. The basic idea of attack transition was used in the segment boundary blurring method[9] to improve music segmentation, as well as frequency-based smearing[10] to improve fundamental frequency estimation. In both of those cases, the approach can be understood as accounting for uncertainty or lack of precision in the annotation process. But in this paper, the concept of transition is used to model the spectrogram evolution process, which overcomes the problem that traditional annotation methods only consider the center frame in the attack stage, like the spectrogram envelope of note's attack, decay, sustain, release(ADSR) proposed by Kelz[11]. The key idea of transition is to describe piano audio from a process rather than a single moment. High-resolution model adopts the method of multi-frame onset annotation and achieves better results than single-frame annotation.

Despite its good performance on MAESTRO dataset, the tran-

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

† Corresponding author

scription accuracy of High-resolution model on MAPS dataset and our OMAPS dataset declines much more than that of Onsets and Frames model. It demonstrates that both the complex multi-frame onset annotation and the simple single-frame onset annotation will affect the robustness of transcription models. Therefore, a compromise approach of multi-frame onset annotation, called attack transition annotation, which uses constant values instead of probability model to annotate multiple frames around onset moment (Figure 1c), is proposed in this paper. Furthermore, the three methods above all annotate a complete evolution process of the note spectrogram called whole transition annotation (Figure 1d).

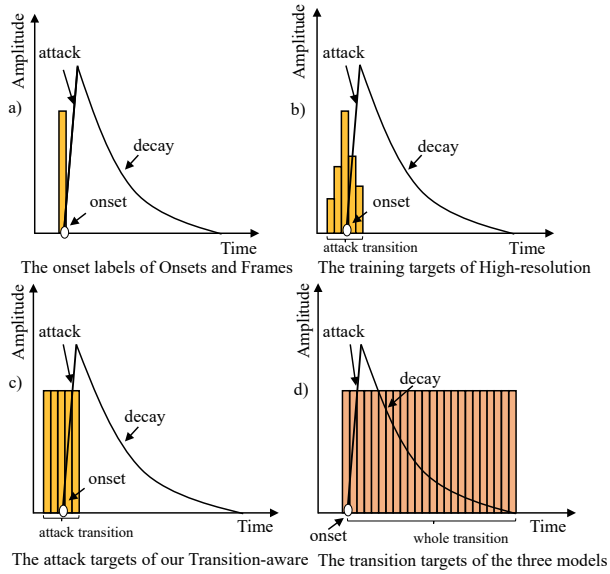

Figure 1: *The annotation method of attack transition and whole transition. The rectangle represents the label value.* ***a)*** *Onsets and Frames only annotates the onset frame.* ***b)*** *High-resolution annotates attack transition with a probability model.* ***c)*** *Transition-aware annotates attack transition with a constant value, one.* ***d)*** *The three models annotate the whole transition in the same way.*

Overall, this paper's main contribution is that the transition concept is used to model the spectrogram evolution process. Transition-aware model is divided explicitly into attack transition and whole transition. Combined with joint training, the attack transition can more accurately predict onset, and the whole transition can also reduce the interference of spectral noise on onset detection. The main contributions are as follows:

- A more robust piano transcription model: it proposes the concepts of attack transition and whole transition and enhances the generalization ability of onset detection through the joint training of two branches, which is superior to the existing methods on MAPS dataset and OMAPS dataset.
- Open source: code of our model and OMAPS dataset will be all available on Github[1] when the paper is published. The inference code and part of OMAPS dataset are now available.

[1]https://github.com/itec-hust/transition-aware

2. RELATED WORK

2.1. The Effect of Time-frequency Representations

Different time-frequency representations have a great influence on the performance of transcription models. Kelz [12] and Cheuk [13] studied the impact of various transformations like CQT and mel on CNN-based models. It indicated that spectrograms with logarithmic frequency axis could get better onset detection results than that of linear frequency axis, for the internal components of the music signal are presented logarithmically. Gao [14] found that using differential spectrograms as input could enhance spectral changes at onset time, and significantly improve the recall of transcription models. Bittner [15] used the spectrograms of harmonic constant Q transform (HCQT) to capture odd harmonics. Cwitkowitz [16] found that the spectrogram features extracted by learning-based filters were not as good as those by fixed-parameter filters, revealing that the end-to-end automatic music transcription systems were not satisfactory.

2.2. Traditional Piano Transcription Methods

Traditional transcription models usually use NMF, SVM, Adaboost, and other machine learning methods. These methods cost less training time but generally could not achieve high accuracy. Cheng [3] used non-negative matrix factorization (NMF) to model piano signals as attack components and decay components, reaching an F1 score of 81.80% on ENSTDKCl, a subset of MAPS dataset. Cogliati [17] utilized convolutional sparse lateral inhibition to reduce piano onset detection errors. Valero-Mas [18] used multi-pitch estimation and onset detection modules to complete note segmentation, and then tried Decision Tree, AdaBoost, Support Vector Machine (SVM), etc., to accomplish pitch classification, finally achieving an F1 score of 73% on MAPS dataset. Deng [19] sent the fundamental and harmonics of 88 pitches in the spectrogram into 88 Adaboost binary classifiers, respectively, so that the feature redundancy and interference of classifiers were reduced.

2.3. Transcription Methods Based on Deep Learning

The development of computer vision has brought much inspiration to piano transcription. Generally, time-frequency transformation is used to convert the one-dimensional audio sequence into a two-dimensional spectrogram. Then convolutional networks are employed to process the spectrogram to complete transcription [20, 21, 22]. For example, Kelz [11] used three parallel CNN networks to predict onset, frame status, and offset, respectively, followed by HMM in the late-stage according to the envelope model of attack, decay, sustain, release (ADSR), achieving an F1 score of 81.38% on MAPS dataset. Pedersoli [23] found that the transcription accuracy of baseline convolutional neural networks could be improved by pre-stacking a U-Net. From another perspective, audio is a sequential signal with strong temporal correlation, so many models are based on the CRNN framework for transcription. Ullrich [24] used CNN's framework with Seq2Seq to model the time correlation and long-term signals dependency. Hawthorne [4] and Kong [7] used CNN with bidirectional LSTM to accomplish onset detection and time correlation learning and achieved F1 score of 94.73% and 97.38% on MAESTRO dataset, respectively. Besides, transcription researches using musical language

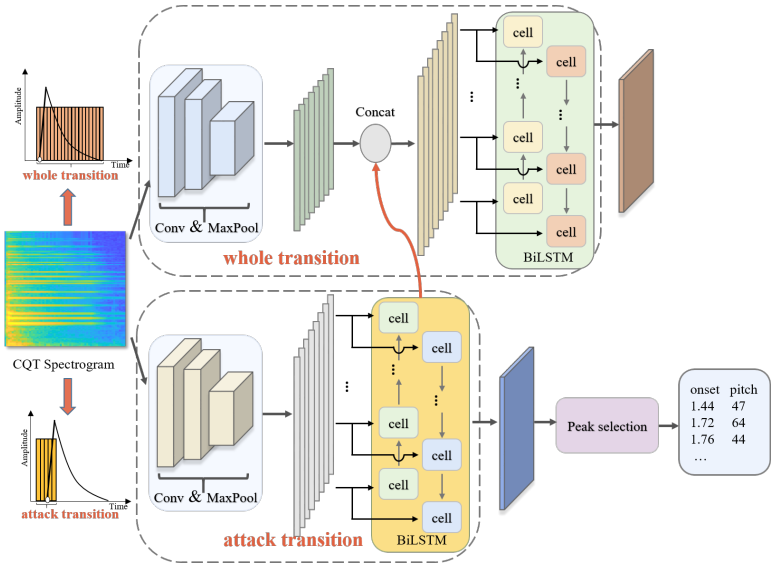

Figure 2: *Overall structure of Transition-aware model.*

models [25, 26, 27] and adversarial learning [28, 29] also have been conducted.

3. METHOD

To improve the generalization ability of piano transcription models, Transition-aware model is proposed in this paper to model the spectrogram evolution of played notes, with details described in the following paragraphs.

3.1. Problem Definition

In the proposed model, constant Q transform (CQT) is adopted to convert piano audio into CQT spectrogram, and then the CQT spectrogram is cut into segments along the time axis. The onset time and pitch[2] corresponding to the notes played in each segment are detected. The complete (onset, pitch) sequence of a track is obtained as the transcription result.

3.2. Overall Network Structure

Transition-aware model consists of three parts: attack-transition, whole-transition, and peak selection. Whole-transition branch integrates probabilities of attack-transition branch, as shown in Figure 2. Attack-transition and whole-transition are trained jointly, but the transcription results are only extracted from attack-transition branch through peak selection. Transition-aware framework firstly uses a multi-layer convolutional neural network to extract features from the CQT spectrogram, then uses BiLSTM for time-dependent modelling, and finally uses fully connected layers as the classifier to complete onset detection and frame estimation. The idea of Transition-aware model is as follows:

- Attack-transition branch realizes the note attack transition process's modelling by annotating multiple frames around onset and completing note onset detection. Attack-transition branch can focus on the spectral attack process of notes, which is more stable than the model that only annotates a single frame.
- Whole-transition branch annotates all frames from onset time to offset time to realize the modelling of the whole transition process and complete the notes' activation prediction. Probabilities from attack-transition branch are fused into whole-transition branch to enhance the stability of the model.
- Peak selection takes the threshold for the probabilities of attack transition branch, selects the peak values in the period above the threshold, and finally obtains each of 88 notes' onset time.

The most significant difference of Transition-aware model is that attack-transition and whole-transition are used to model the spectrogram evolution process, and peak selection is used for post-processing. Compared with annotating only the onset frame, Attack-transition enables the model to focus on the spectrogram evolution of notes in the attack stage, making the onset detection more stable. Whole-transition branch allows the model to focus on the complete spectral evolution process of notes, and it also plays an auxiliary role in stabilizing onset detection. The implementation details of the Transition-aware model are described in the following sections.

3.3. Constant Q Transform

Constant Q transform (CQT) is a standard time-frequency transformation method in music signal processing, which realizes converting the one-dimensional audio sequence to the two-dimensional spectrogram. Compared with short time Fourier transform (STFT), different frequency resolution is used for high and low frequency signals in CQT, which is more suitable for music signal processing. CQT in this paper is realized by the CQT function in Librosa library [30]. The specific parameters are referred to [31].

[2]The value range of 88 pitches is 21-108

The sampling rate is 44100Hz, the frame hop_length is 512, the frequency bins of each octave are 48, and the total frequency bins are 356.

3.4. Attack-transition and Whole-transition

Attack-transition branch uses a five-layer convolutional network to extract the features of CQT spectrogram segments and then uses BiLSTM to conduct time-dependent modelling. Finally, through fully connected layers, an 88-dimensional vector is output, representing the probabilities of note onsets, and the process modelling of note attack transition is accomplished.

As shown in Table 1, each feature map's size can be expressed in the form of $time\_length \times freq\_bins \times channels$. For example, the size of the input is $(T+8) \times 356 \times 2$, where $(T+8)$ represents the number of the input spectrogram frames. The optimal value of T will be obtained in section 4.3. Since convolution operation will reduce the size, to ensure the predictions are T frames, 8 frames of the spectrogram are padded around the T frames segment in the center. 356 is the CQT frequency bins, as described in 3.3. Since dual-channel audios are used, the number of input spectrogram channels is 2.

To ensure the size of the predictions is T frames, the size of the feature map on the time axis can only be reduced by 8. So a small convolution kernel size, 3, is used on the time axis. To get a slightly large receptive field, 5 is used as the convolution kernel size on the frequency axis. Thus the size of the convolution kernel is 3×5. In order to learn the time dependence between different frames, BiLSTM is used in the model. And to speed up training convergence, batch normalization layer is used behind each convolutional layer. To prevent over-fitting, dropout layers are added to the fully connected layers. Specific parameters of attack-transition branch are shown in Table 1. The convolutional layer parameters, $H \times W@C$, refer to the height of the convolutional kernel as H, width as W, and the number of kernels as C. The max-pooling layer parameters, $pH \times pW/pSH \times pSW$, indicate that the height of the pooling area is pH, the width is pW, the moving step in the high direction is pSH, and the moving step in the wide direction is pSW.

Table 1: *Attack-transition network parameters.*

Input	Layer & Parameter	Output
$(T+8) \times 356 \times 2$	Conv:3×5@8	$(T+6) \times 352 \times 8$
$(T+6) \times 352 \times 8$	Conv:3×5@16	$(T+6) \times 352 \times 16$
$(T+6) \times 352 \times 16$	Pool:$2 \times 2/1 \times 2$	$(T+6) \times 176 \times 16$
$(T+6) \times 176 \times 16$	Conv:3×5@32	$(T+4) \times 172 \times 32$
$(T+4) \times 172 \times 32$	Conv:3×5@64	$(T+2) \times 168 \times 64$
$(T+2) \times 168 \times 64$	Pool:$2 \times 2/1 \times 2$	$(T+2) \times 84 \times 64$
$(T+2) \times 84 \times 64$	Conv:3×5@128	$T \times 80 \times 128$
$T \times 80 \times 128$	Reshape	$T \times 10240$
$T \times 10240$	Dense:1024	$T \times 1024$
$T \times 1024$	BiLSTM:512	$T \times 1024$
$T \times 1024$	Dense:88	$T \times 88$

Whole-transition branch also uses a five-layer convolutional network to extract the features of the CQT spectrogram segments and then uses fully connected layers to reduce the dimension. Then, the features fused with attack-transition branch are sent to BiLSTM for time-dependent modelling. Finally, an 88-dimension vector is output, representing the probabilities of the activation status of notes, realizing the whole-transition process modelling of notes. The specific parameters of whole-transition branch are shown in Table 2.

Table 2: *Whole Transition network parameters.*

Input	Layer & Parameter	Output
$(T+8) \times 356 \times 2$	Conv:3×5@8	$(T+6) \times 352 \times 8$
$(T+6) \times 352 \times 8$	Conv:3×5@16	$(T+6) \times 352 \times 16$
$(T+6) \times 352 \times 16$	Pool:$2 \times 2/1 \times 2$	$(T+6) \times 176 \times 16$
$(T+6) \times 176 \times 16$	Conv:3×5@32	$(T+4) \times 172 \times 32$
$(T+4) \times 172 \times 32$	Conv:3×5@64	$(T+2) \times 168 \times 64$
$(T+2) \times 168 \times 64$	Pool:$2 \times 2/1 \times 2$	(T+2)$\times 84 \times 64$
$(T+2) \times 84 \times 64$	Conv:3×5@128	$T \times 80 \times 128$
$T \times 80 \times 128$	Reshape	$T \times 10240$
$T \times 10240$	Dense:88	$T \times 88$
$T \times 88$	Concat	$T \times 176$
$T \times 176$	BiLSTM:128	$T \times 256$
$T \times 256$	Dense:88	$T \times 88$

Transition-aware model combines the training of the attack-transition branch and whole-transition branch and simultaneously completes the modelling of the spectral process of notes in the attack stage and the whole evolution, improving the detection probabilities of note onset. The loss function l_{note} of the joint training can be expressed as:

$$l_{note} = l_{attack} + l_{whole} \tag{1}$$

Where, l_{attack} and l_{whole} are the cross entropy loss functions of attack-transition branch and whole-transition branch respectively.

3.5. Peak Selection

Peak selection is the post-processing part of Transition-aware model. Peak selection is performed on the probabilities of note onsets output by attack-transition branch to obtain the final (onset, pitch) sequence as the transcription result. For each of the 88 notes, peak selection considers only the period in which the onset probabilities are greater than the threshold of 0.5. Peak selection then selects the peak values within each period and sets the corresponding moment of peak value as the related note's onset.

4. EXPERIMENT

4.1. Dataset

The principal datasets of piano transcription include MAPS dataset [5] and MAESTRO dataset [4]. MAPS dataset has a total of 270 complete piano performances. Generally, 210 synthetic audio recordings are used as the training set, and 60 real performances are used as the test set. MAESTRO dataset has been created by Google Brain in collaboration with the International Piano-e-Competition. Besides, MAESTRO-v2 has also been proposed, added 396 audio tracks from the 2018 contest compared to MAESTRO-v1.

The OMAPS (Ordinary MIDI Aligned Piano Sounds) dataset established in this paper contains complete playing videos, audios, and corresponding annotations, primarily for our study of audio-visual fusion transcription. The OMAPS dataset was recorded from Yamaha electric piano P115 by a piano player. The Logitech C922 Pro HD stream webcam was used to record video and audio simultaneously. The Logitech camera is available in both 1080p/30fps and 720p/60fps video configurations. To ensure the

resolution of the video, we used the 1080p/30fps configuration. The Logitech camera audio module's sampling rate is 44100Hz. Since the recorded videos and piano MIDI files were out of sync, we manually aligned the exported MIDI files as annotations. The OMAPS dataset contains 106 different pieces for a total of 216 minutes, with an average two minutes per piece. The number of notes played per second is used to measure the playing speed. According to the playing speed, the OMAPS dataset is divided into a training set and a test set. The training set and the test set have the same playing speed distribution. The training set contains 80 videos, and the test set contains 26 videos, as shown in Table 3.

Table 3: *Statistics of the OMAPS dataset.*

Split	Performance	Duration, minutes	Size, GB	Notes
Train	80	123	3.18	60,589
Test	26	53	1.03	19,135
Total	106	176	4.22	79,724

Although OMAPS dataset contains videos and audios, and the dataset is divided into training part and test part, only the audios and annotations of the dataset are used in this paper. Besides, since the total time of the audios is short, OMAPS dataset is not used to train the model. Instead, both the training set and test set of OMAPS are used for evaluation.

4.2. Evaluation Metrics

Precision, recall and F1 score are used to evaluate the performance of piano transcription models. Precision represents the extra note detection errors, recall represents the missed note detection errors, and F1 score represents the model's all-around performance. Precision, recall, and F1 score calculation formulas are as follows:

$$P = \frac{TP}{TP + FP} \tag{2}$$

$$R = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = \frac{2 \times P \times R}{P + R} \tag{4}$$

Where TP is the number of correct detected notes, FP is the number of extra detected notes, and FN is the number of missed detected notes. At present, the evaluation algorithm implemented in mir_eval library [32] is commonly used to evaluate transcriptional models, and the time tolerance of onset is set to ±50ms.

4.3. Hyperparameters Selection

As shown in Figure 1c, Transition-aware models the note attack transition by successively annotating multiple frames around onset to improve the detection accuracy of note onset. Meanwhile, whole-transition branch is used to model the whole transition process to assist attack-transition. However, different annotation lengths of attack-transition will make the model consider attack-transition of different time scales. Spectrogram segments of different lengths brings different context information to whole-transition modelling, which will affect the detection accuracy of note onset.

To discuss the impact of the time scale, we defines the blurred length and frame length. Blurred length represents the annotation length of attack-transition process. Frame length describes the length of the input CQT spectrogram segments in the time dimension. Frame length is the T in section 3.4. To simplify the experiment, it is assumed that the effects of blurred length and frame length on the model are not coupled. First, the blurred length is fixed at 5 to find the optimal frame length. Then fix the frame length and search for the best blurred length. We train on MAESTRO dataset and test on OMAPS dataset. The experimental results are shown in Figure 3. It can be seen that the optimal frame length, or T, is 101, and the corresponding time scale is 1.173s. The optimal blurred length is 5, corresponding to a time scale of 0.058s.

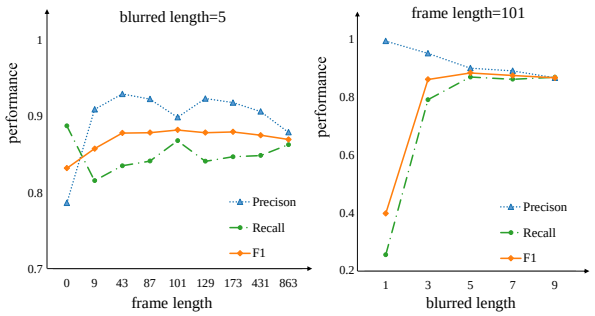

Figure 3: *Performance of each model with different blurred length and frame length on OMAPS dataset.*

4.4. Different network structures

The structure of Transition-aware is shown in Figure 2, where the features of the attack-transition branch are integrated into the whole-transition branch but the whole-transition branch is not used during prediction. This section will expand on this consideration in detail. We separately studied the results of using only the attack-transition branch, the attack features integrated into the whole-transition branch, the whole-transition features integrated into the attack-transition branch, and the fusion of the attack-transition features and whole-transition branch features, as shown in Figure 4. Figure 4b corresponds to Figure 2.

We use MAESTRO-v1 dataset to train the four models above, and test them on MAESTRO-v1 test set, MAESTRO-v2 test set, MAPS test set, and the whole OMAPS dataset respectively. The results are shown in Table 4. The *only attack* model structure only considers the modeling of the attack transition. Although good results have been achieved on MAESTRO-v1 dataset and MAESTRO-v2 dataset, the generalization ability is weak, and the performance on MAPS dataset and OMAPS dataset is not very good. *Attack into whole* model structure is slightly better than the *whole into attack* model structure, and has achieved the best performance on MAPS dataset and OMAPS dataset. Although the *attack-whole fused* model structure performs best on MAESTRO-v1 dataset and MAESTRO-v2 dataset, it does not perform well on the MAPS dataset and OMAPS dataset. Maybe this fusion method is more complicated, resulting in overfitting. So finally we choose the *attack into whole* model structure as our optimal model structure.

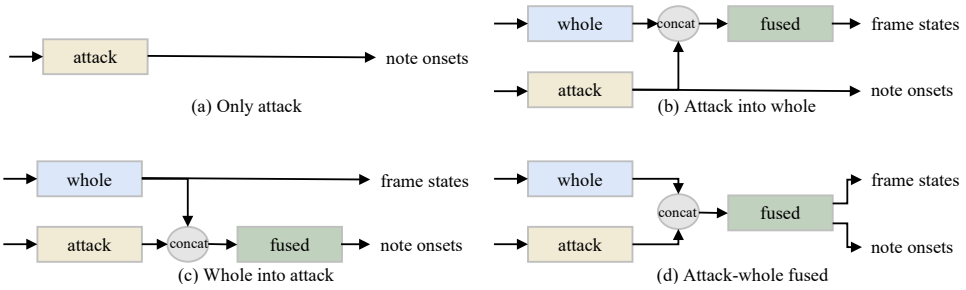

Figure 4: *Four network structures.*

Table 4: *Performance of the four model structures.*

	MAESTRO-v1			MAESTRO-v2			MAPS			OMAPS		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Only attack	98.78	94.33	96.48	98.70	94.65	96.59	85.91	87.04	86.41	77.87	89.73	83.24
Attack into whole	99.10	92.17	95.45	98.96	92.37	95.47	89.64	85.62	**87.52**	89.87	86.80	**88.21**
Whole into attack	98.69	93.33	95.89	98.49	93.42	95.82	88.32	86.73	87.47	86.65	88.96	87.70
Attack-whole fused	98.41	94.20	**96.24**	98.41	94.51	**96.38**	84.38	86.98	85.60	77.50	90.25	83.24

4.5. Comparison with the State-of-the-Art Methods

As described in the respective papers, MAESTRO-v1 dataset was used to train and validate Onsets and Frames model [4], and MAESTRO-v2 dataset was used to train and validate High-resolution model [7]. Although Onsets and Frames model and High-resolution model have used different training sets for training, the impact on the final performance of the models can be ignored, for the difference between MAESTRO-v1 dataset and MAESTRO-v2 dataset is very small. MAESTRO-v1 dataset is used to train and validate Transition-aware model in our experiment. In this paper, MAESTRO-v1 dataset, MAESTRO-v2 dataset, MAPS dataset and OMAPS dataset are used to test the above three models. Test sets in MAESTRO-v1 and MAESTRO-v2, the 60 real performances in MAPS dataset, and all recordings in OMAPS dataset are used for testing. To illustrate clearly, the dataset used for training and validation of each model is shown in Table 5. All models are testing on MAESTRO-v1 test set, MAESTRO-v2 test set, MAPS test set (i.e., 60 real performances) and the whole OMAPS dataset (i.e., training set and test set).

Table 5: *Dataset configuration of the three models.*

	Training	Validation
Onsets and frames	MAESTRO-v1 training set	MAESTRO-v1 validation set
High-resolution	MAESTRO-v2 training set	MAESTRO-v2 validation set
Transition-aware	MAESTRO-v1 training set	MAESTRO-v1 validation set

The performance of the three models on the above four datasets is shown in Table 6. High-resolution model has the best performance on MAESTRO-v1 dataset and MAESTRO-v2 dataset, with the onset F1 score reaching 97.38% and 96.77%, respectively. However, High-resolution model's performance is the worst on MAPS test set and OMAPS dataset, which indicates that High-resolution model has the problem of over-fitting and weak generalization ability.

The F1 score of Transition-aware model is significantly higher than that of High-resolution model on MAPS test set and OMAPS dataset (4.35% and 10.31%, respectively), and slightly lower on MAESTRO-v1 test set and MAESTRO-v2 test set (1.93% and 1.3%, respectively). Besides, Transition-aware model is better than Onsets and Frames model on the four datasets[3]. This shows that the annotation method of Transition-aware is more effective than that of Onsets and Frames. In conclusion, Transition-aware model is a better choice in terms of piano transcription performance and robustness.

To analyze the performance of each model in detail, the three models' error distribution on MAPS test set and OMAPS dataset is shown in Figure 5. Compared with Onsets and Frames model and High-resolution model, Transition-aware model reduces many extra note errors, increasing the stability of onset detection, which is consistent with the design goal of Transition-aware model in this paper.

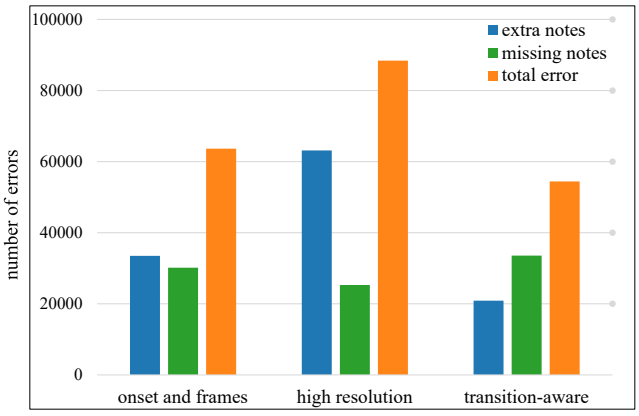

Figure 5: *Error distribution of the three models on MAPS test set and OMAPS dataset.*

[3]Onsets and Frames achieved 94.80% of F1 score on MAESTRO-v1 in [4], but we found that only 94.73% of F1 score was achieved by repeating the prediction process.

Table 6: *Performance of the three models.*

	MAESTRO-v1			MAESTRO-v2			MAPS			OMAPS		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Onsets and frames[4]	97.91	91.82	94.73	97.88	92.26	94.93	87.42	85.58	86.43	81.03	90.22	85.17
High-resolution [7]	98.52	96.29	**97.38**	98.16	95.46	**96.77**	79.57	87.35	83.17	68.31	91.58	77.90
Transition-aware	99.10	92.17	95.45	98.96	92.37	95.47	89.64	85.62	**87.52**	89.87	86.80	**88.21**

High resolution model used complex annotation methods, which might lead to the model fitting the noise pattern in the piano signal, resulting in many extra note errors. Onsets and Frames model is robust, and its extra note errors and missing note errors are few. However, its modeling of attack stage is not enough, and its onset detection performance can not reach the optimal. Transiton-aware model uses the method of annotating continuous multiple frame around onset time combined with peak selection post-processing to obtain the most robust onset detection results. On the one hand, the continuous annotation of multiple frames overcomes the over fitting problem of High resolution; on the other hand, the peak selection is different from the method of directly taking threshold of output probabilities proposed by Onsets and Frames model and High resolution model, which makes the model more insensitive to interference.

5. DISCUSSION AND CONCLUSION

In this paper, a simple and effective Transition-aware piano transcription model is proposed. Joint training of attack-transition branch and whole-transition branch enhances the perception ability of transcription model on the evolution process of piano signals and improves the stability and robustness of onset detection. OMAPS dataset using the ordinary electric piano in the general recording environment is established to expand the discussion on the models' generalization ability. Transition-aware model proposed in this paper has a comprehensive optimal performance on MAESTRO-v1 dataset, MAESTRO-v2 dataset, MAPS dataset and OMAPS dataset.

Of course, there are many limitations to Transition-aware model. First of all, the piano signal is a combination of fundamental and harmonics, and different playing methods and pianos will lead to various combination forms of fundamental and harmonics. Such complexity brings difficulties for the perception of Transition-aware model. Secondly, most of the current piano transcription models only complete transcription from a single auditory or visual [33, 34]. In the future, transcription systems based on audio and video multi modes will accomplish this task better.

6. ACKNOWLEDGMENTS

This work is supported by The National Natural Science Foundation of China (No. 61877060). We also thank the musician, Tianyue Yu, for recording OMAPS dataset.

7. REFERENCES

[1] E. Vincent, N. Bertin, and R. Badeau, "Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 109–112.

[2] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 528–537, 2009.

[3] T. Cheng, M. Mauch, and E. Benetos, "An attack/decay model for piano transcription," in *Proceedings of the International Society for Music Information Retrieval Conference*, 2016, pp. 584–590.

[4] C. Hawthorne, A. Stasyuk, and A. Roberts, "Enabling factorized piano music modeling and generation with the maestro dataset," in *Proceedings of International Conference on Learning Representations*, 2019.

[5] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1643–1654, 2009.

[6] X. Gong, W. Xu, and J. Liu, "Analysis and correction of maps dataset," in *Proceedings of the 22th International Conference on Digital Audio Effects (DAFx-19)*, 2019.

[7] Q. Kong, B. Li, , and X. Song, "High-resolution piano transcription with pedals by regressing onsets and offsets times," in *arXiv preprint arXiv:2010.01815*, 2020.

[8] Q. Kong, B. Li, , and J. Chen, "Giantmidi-piano: A large-scale midi dataset for classical piano music," in *arXiv preprint arXiv:2010.07061*, 2020.

[9] J. Schlüter K. Ullrich and T. Grill, "Boundary detection in music structure analysis using convolutional neural networks," in *Proceedings of the International Society for Music Information Retrieval Conference*, 2014, pp. 417–422.

[10] P. Li J. W. Kim, J. Salamon and J. P. Bello, "Crepe: A convolutional representation for pitch estimation," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 161–165.

[11] R. Kelz, S. Böck, and G. Widmer, "Deep polyphonic adsr piano note transcription," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 246–250.

[12] R. Kelz, M. Dorfer, and F. Korzeniowski, "On the potential of simple framewise approaches to piano transcription," in *Proceedings of the International Society for Music Information Retrieval Conference*, 2016, pp. 475–481.

[13] K. W. Cheuk, K. Agres, and D. Herremans, "The impact of audio input representations on neural network based music transcription," in *Proceedings of IEEE International Joint Conference on Neural Networks*, 2020, pp. 1–6.

[14] L. Gao, L. Su, and Y. H. Yang, "Polyphonic piano note transcription with non-negative matrix factorization of differential spectrogram," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 291–295.

[15] R. M. Bittner, "Deep salience representations for f0 estimation in polyphonic music," in *Proceedings of the International Society for Music Information Retrieval Conference*, 2017, pp. 63–70.

[16] F. C. Cwitkowitz, *End-to-End Music Transcription Using Fine-Tuned Variable-Q Filterbanks*, Ph.D. thesis, Rochester Institute of Technology, 2019.

[17] A. Cogliati, Z. Duan, and B. Wohlberg, "Piano transcription with convolutional sparse lateral inhibition," *IEEE Signal Processing Letters*, vol. 24, no. 4, pp. 392–396, 2017.

[18] J. J. Valero-Mas, E. Benetos, and J. M. Iñesta, "A supervised classification approach for note tracking in polyphonic piano transcription," *Journal of New Music Research*, vol. 47, no. 3, pp. 249–263, 2018.

[19] K. Deng, G. Liu, and Y. Huang, "An efficient approach combined with harmonic and shift invariance for piano music multi-pitch detection," in *Proceedings of International Workshop on Pattern Recognition*, 2019.

[20] Q. Wang, R. Zhou, and Y. Yan, "A two-stage approach to note-level transcription of a specific piano," *Applied Sciences*, vol. 7, no. 9, pp. 901–901, 2017.

[21] S. Liu, L. Guo, and G. A. Wiggins, "A parallel fusion approach to piano music transcription based on convolutional neural network," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 391–395.

[22] R. Kelz, S. Böck, and G. Widmer, "Multitask learning for polyphonic piano transcription, a case study," in *Proceedings of IEEE International Workshop on Multilayer Music Representation and Processing*, 2019, pp. 85–91.

[23] F. Pedersoli, G. Tzanetakis, and K. M. Yi, "Improving music transcription by pre-stacking a u-net," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 506–510.

[24] K. Ullrich, "Music transcription with convolutional sequence-to-sequence models," in *Proceedings of the International Society for Music Information Retrieval Conference*, 2017.

[25] S. Sigtia, E. Benetos, and N. Boulanger-Lewandowski, "A hybrid recurrent neural network for music transcription," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 2061–2065.

[26] 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.

[27] A. Ycart, A. McLeod, and E. Benetos, "Blending acoustic and language model predictions for automatic music transcription," in *Proceedings of the International Society for Music Information Retrieval Conference*, 2019, pp. 454–461.

[28] A. Ycart, D. Stoller, and E. Benetos, "A comparative study of neural models for polyphonic music sequence transduction," in *Proceedings of the International Society for Music Information Retrieval Conference*, 2019, pp. 470–477.

[29] J. W. Kim and J. P. Bello, "Adversarial learning for improved onsets and frames music transcription," in *Proceedings of the International Society for Music Information Retrieval Conference*, 2019, pp. 670–677.

[30] C. Schörkhuber and A. Klapuri, "Constant-q transform toolbox for music processing," in *Proceedings of Sound and Music Computing Conference*, 2010, pp. 3–64.

[31] S. Kong, W. Xu, and W. Liu, "Onset-aware polyphonic piano transcription: A cnn-based approach," in *Proceedings of the International Workshop on Computer Science and Engineering*, 2019, pp. 454–461.

[32] C. Raffel, B. McFee, and E. J. Humphrey, "mir_eval: A transparent implementation of common mir metrics," in *Proceedings of the International Society for Music Information Retrieval Conference*, 2014, pp. 367–372.

[33] J. Li, W. Xu, and Y. Cao, "Robust piano music transcription based on computer vision," in *Proceedings of the High Performance Computing and Cluster Technologies Conference*, 2020, pp. 92–97.

[34] A. S. Koepke, O. Wiles, and Y. Moses, "An end-to-end approach for visual piano transcription," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 1838–1842.