ANALYSIS AND CORRECTION OF MAPS DATASET

Xuan Gong *
School of Electronic Information and Communications
Huazhong University of Science and Technology
Wuhan, China
M201771839@hust.edu.cn

Wei Xu †
School of Electronic Information and Communications
Huazhong University of Science and Technology
Wuhan, China
xuwei@hust.edu.cn

Juanting Liu
School of Electronic Information and Communications
Huazhong University of Science and Technology
Wuhan, China
M201871969@hust.edu.cn

Wenqing Cheng
School of Electronic Information and Communications
Huazhong University of Science and Technology
Wuhan, China
chengwq@mail.hust.edu.cn

ABSTRACT

Automatic music transcription (AMT) is the process of converting the original music signal into the digital music symbol. The MIDI Aligned Piano Sounds (MAPS) dataset was established in 2010 and is the most used benchmark dataset for automatic piano music transcription. In this paper, error screening is carried out through algorithm strategy, and three data annotation problems are found in ENSTDkCl, which is a subset of MAPS, usually used for algorithm evaluation: (1) there are 342 deviation errors of midi annotation; (2) there are 803 unplayed note errors; (3) there are 1613 slow starting process errors. After algorithm correction and manual confirmation, the corrected dataset is released. Finally, the better-performing Google model and our model are evaluated on the corrected dataset. The F values are 85.94% and 85.82%, respectively, and it is correspondingly improved compared with the original dataset, which proves that the correction of the dataset is meaningful.

1. INTRODUCTION

Automatic music transcription (AMT) is the process of converting acoustic music signals into digital music symbols, which is a challenging task in the field of music signal processing and music information retrieval (MIR) [1]. It consists of several subtasks, including multi-pitch estimation, onset offset detection [2], instrument recognition, beat and rhythm tracking [3]. Automatic music transcription systems can be used in music education, music creation, music production [4], music search [5] and so on. At present, AMT is still considered to be a challenging and open problem, especially for automatic piano music transcription [6]. The overlapping of sound events at the same time often shows harmonic overlap [7], which makes the identification task more difficult.

MIDI Aligned Piano Sounds (MAPS) [8] dataset was established in 2010 and is the most widely used benchmark dataset for piano transcription. MAPS dataset contains about 31GB of audio recordings. There are 9 types of audio recordings corresponding to different piano types and recording conditions, among which 7 types of audio are produced by software piano synthesizers, and 2 subsets ENSTDkAm and ENSTDkCl are recorded by an upright Disklavier piano. In general, ENSTDkCl in MAPS Disklavier dataset is adopted to evaluate. Since the MAPS dataset was established, only Ycart [9] updated it and added rhythm and key information in the annotation.

With the research on the AMT system, some researchers found that there were some problems in the MAPS Disklavier dataset for evaluation. Ewert [10] found in his study of studio piano transcription that in MAPS Disklavier dataset some midi-based annotation deviation exceeded the order of magnitude described in the document. Therefore, he used the greater temporal tolerance which might provide a more realistic impression of the transcription performance. Li [11] found several issues with the MAPS Disklavier dataset, including annotation deviations, omitted notes and recordings consisting entirely of percussive keybed and pedal noises. Hawthorne [12] suggested that some of the low-velocity notes in the annotations were not played during the Yamaha piano playing.

The development of machine learning has led many scholars to apply it to the piano transcription system. In [13]-[14], the authors demonstrated the potential of a single CNN-based acoustic model and an RNN model for polyphonic piano music transcription. The model proposed by Hawthorne [12], which we called the Google model, was a new method to predict the pitch using CNN and LSTM, and achieved the best system performance in 2018. We also designed a CNN-based model for piano transcription [15]. The model consisted of two networks, and an onset-event detector was used to align the pitch onset to a more accurate position. In the end, our model achieved an F1-measure score of 85.15% on the MAPS ENSTDkCl dataset, which is better than Google's system performance.

Data plays an important role in machine learning methods. After analyzing the results of errors in the ENSTDkCl set evaluation, we find that some errors are caused by the dataset itself instead of model identification. We believe that it is necessary to modify the dataset rather than directly modifying the criteria of the evaluation to a broader range. We use the algorithm preselection and manual check to correct the dataset. There are the deviation of midi annotation errors, unplayed note errors and slow start process errors in the data error, whose number is 342, 803 and 1613. The two models are evaluated using the modified dataset, with F1-measure scores 85.94% and 85.82%, respectively, and the modified dataset reflected the performance of the transcription system more accurately. In addition, applying

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

† Corresponding author.

these modified true piano data to the training process may help the network to learn the characteristics of true piano playing better.

2. THE ANALYSIS OF ERROR

The MAPS dataset includes midi files, txt files, and corresponding audio files. The txt file contains information in the midi, including the onset time, offset time and pitch. We analyze the data and find that the data set itself contains three problems: deviation errors of midi annotation, unplayed note errors and slow start process errors.

2.1. Method of analysis

We analyze the data of the ENSTDkCl set. Direct data analysis is a huge task, so we use the Google model [12] and our model [15] for transcription in the current research. We analyze the data based on the common error results of the two models. Common errors in the transcription of the two models include missed detection and surplus detection. According to the general evaluation criteria, the missed detection does not detect the corresponding pitch onset event within ±50 ms of the corresponding time of midi, and the surplus detection is that there is no matching midi annotation within ±50ms of the pitch onset event. There are few errors in surplus inspections and most of them are detection errors of the model. Therefore, the results of the missed detection are analyzed to observe the data problem.

2.2. Reasonable midi annotation

To analyze the data, we first describe the general performance of the note onset. The data of the MAPS subset ENSTDkCl is generated in the real environment. When there is a pitch onset event in the midi, the corresponding piano key that is tapped will generate energy, and the amplitude of the frequency corresponding to the pitch rises. Except the partial bass, the fundamental and second harmonics of the tone contain most of the energy produced by the tap [16]. So we use the fundamental and second harmonics in the figure to show the onset. The spectral transformation of the audio is more conducive to the presentation of its sound characteristics, so this analysis uses the CQT spectral transform.

As shown in Fig. 1 is a pitch onset event, Fig 1.a is a three-dimensional spectrogram, where the three axes are the time, frequency and amplitude of the CQT transform. The height and color represent the amplitude at the same time. The larger the amplitude, the closer the color is to yellow and the higher the height. The white dashed line represents the same time t1, and the white curve represents the case where the center frequency corresponding to the pitch changes with time. The red box in Fig 1.b represents the time and fundamental frequency range corresponding to the note event, and the blue box represents the frequency multiplication range of the pitch event. The abscissa is time, the ordinate is frequency, and the color depth indicates the magnitude of the corresponding time and frequency. The greater the degree of black, the larger the amplitude. The red curve in the lower of Fig 1.c shows the center frequency component of the fundamental frequency range, and the blue curve shows the center frequency component of the second harmonics. The three figures represent the same pitch event from different angles. In Fig 1.a, the midi is marked with time t1=177.574s, and the frequency amplitude of the pitch F5 has a large rise. From the Fig 1.b and the Fig 1.c, the fundamental frequency amplitude rises rapidly to 1.5 at t1, and the second harmonics also increases correspondingly. Therefore, when a pitch onset event occurs, the amplitude of the base frequency corresponding to the pitch will increase, and the amplitude of second harmonics will also have a certain upward trend.

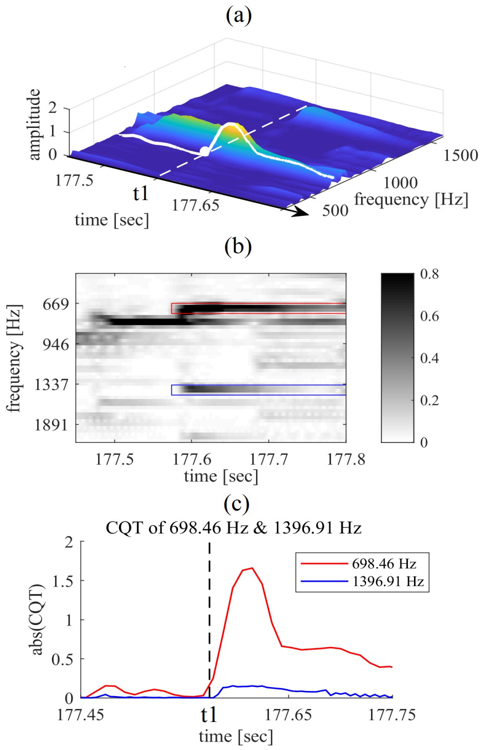

Figure 1: *The spectrum of reasonable midi annotation. (a) is a three-dimensional spectrogram, (b) is the colormap of spectrum, and (c) are the center frequency components of the fundamental frequency and second harmonics.*

2.3. The deviation error in midi

The MAPS subset ENSTDkCl is automatically played by the piano based on the information in midi, but we found that there is a case where the playing time in the audio is shifted from the midi annotation by more than 50 ms. The two time points have obvious inconsistency. This deviation in the midi annotation can lead to inaccurate final evaluation results.

As shown in Fig. 2 is a fragment of MAPS_MUS-schuim-1_ENSTDkCl, In Fig 2.a, t1 is the starting time of the midi pitch E4, 112.389s, and the time represented by t2 is 112.489s; in Fig 2.b the red box indicates the time and the fundamental frequency range of the corresponding pitch E4 in the midi; and the c is the curve of the two center frequencies corresponding to the pitch E4 with time. Fig 2.a shows that there is no amplitude increase at t1 and the frequency amplitude rises at t2. We think that t2 should be the onset time of this pitch event according to the general representation of onset event. Fig 2.b shows that all frequencies in the fundamental frequency range have no pitch starting characteristics at time t1. Observing the Fig 2.c, there is certain deviation between t1 and t2, and the distance between them is 0.1s. Therefore, we think that this is a deviation error in midi annotation.

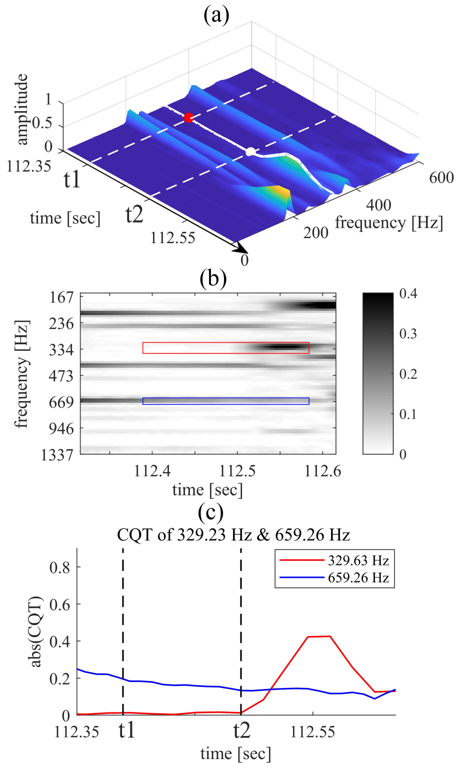

Figure 2: *The spectrum of the deviation error in midi. (a) is a three-dimensional spectrogram, (b) is the colormap of spectrum, and (c) are the center frequency components of the fundamental frequency and second harmonics.*

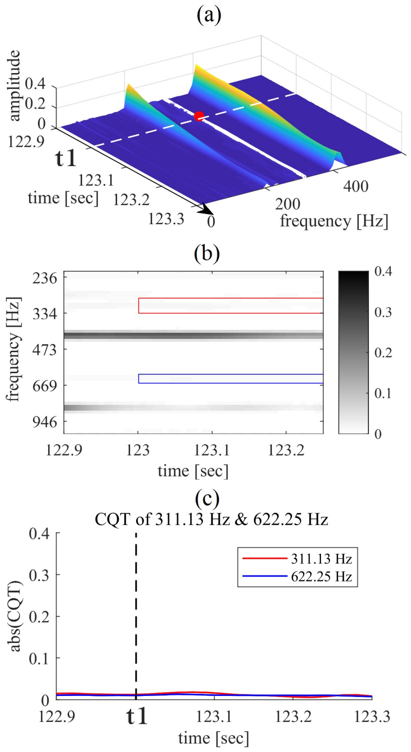

Figure 3: *The spectrum of unplayed note error which is a White Noise. (a) is a three-dimensional spectrogram, (b) is the colormap of spectrum, and (c) are the center frequency components of the fundamental frequency and second harmonics.*

2.4. The unplayed note error

When there is a note onset in the midi, the amplitude rises corresponding to the pitch frequency. It is not normal that there is a note onset event in midi but the frequency is degraded or absent in the spectrum. We believe that this is the case when the note is not played on the piano, and then we describe it with data.

In the first case, the note is not heard in the audio at the start time of the midi. By observing the frequency spectrum, it is found that the frequency component of the corresponding pitch at the labeling time is very small, which is the same as the noise level, so we consider that the pitch is not played.

As shown in the Fig. 3, it is a fragment of the MAPS_MUS-schuim-1_ENSTDkCl, the dotted line shown in Fig 3.a represents the midi annotation time t1, and its time is 123.001s. The white curve represents the change of the amplitude (311.13Hz) of the fundamental central frequency of pitch D#4 over time. In the Fig 3.b, the fundamental frequency and the second harmonics of the pitch D#4 are small and have no upward trend during the marked time period. It can be seen from Fig 3.c that the center frequency of the two frequency ranges are both below 0.1. This is not the performance of normal playing, so we think that the piano does not play D#4 at t1.

The second case is that the beginning of the same pitch marked at two very close times can only be heard once in the audio, so we don't think there is a new play in one of the labels. By analysing the spectrum, we find that this situation has a specific performance.

As shown in the Fig. 4, it is the track MAPS_MUS-mz_570_1_ENSTDkCl, t1 is the midi annotation time 436.823s, t2 is 436.920s, the white curve represents the curve of the center frequency component (466.16 Hz) of the pitch A#4, and Fig 4.b shows the spectral characteristics of the fundamental frequency and the frequency doubling corresponding to the pitch A#4, and two marked points are observed from the Fig 4.c. The characteristic is that compared with t1, the fundamental frequency center component of the pitch A#4 of t2 shows a downward trend, and there is no frequency rising process like the normal playing onset. We didn't hear the two playing while listening to the audio, so we thought that the piano did not play the pitch A#4 at t2.

2.5. Slow start process error

As a percussion instrument, when the piano key is pressed, there will be a process in which the frequency component of the key rises. When analysing the data, we find a pitch onset event, whose frequency component rises lasts longer than 100ms. This is not a common performance in piano playing, and the labelling itself is difficult to unify, which is a huge challenge for the recognition task.

As shown in the Fig. 5, the track MAPS_MUS-pathetique_1_ENSTDkCl, in Fig 5.a, the white curve is the change of the fundamental frequency of the pitch G4 with time, and the corresponding midi labeling time t1 is 292.26 s. It can be seen from the figure that this frequency component rises continuously. It can be seen from the Fig 5.c that the amplitude has been rising in the range of 292.15s-292.3s, and the rising process lasts for 150ms. The change process is long, which is difficult to accurately identify in specific tasks.

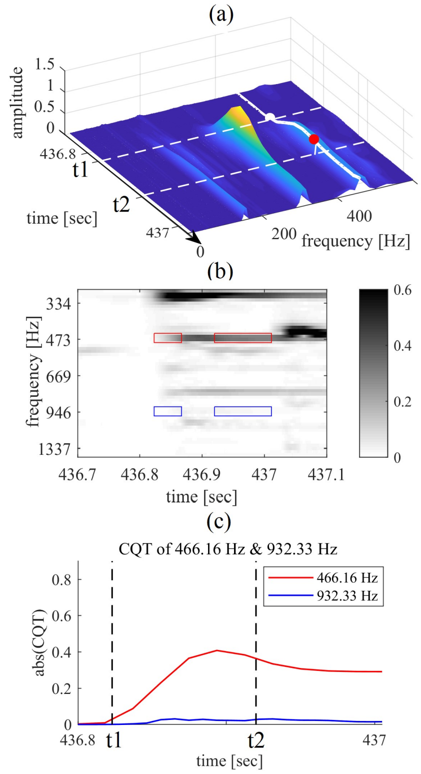

Figure 4: *The spectrum of unplayed note error whose frequency declines. (a) is a three-dimensional spectrogram, (b) is the colormap of spectrum, and (c) are the center frequency components of the fundamental frequency and second harmonics.*

3. JUDGEMENT AND CORRECTION METHOD

Based on the analysis in above chapter, we make the quantify judgment of three kinds of notes, there are 342 deviation errors, 803 unplayed note errors and 1613 slow starting process errors. We list these notes in Cl-correction-v1.0, the numbers of each kind of notes are listed in Table.1. The complete files are published on https://github.com/itec-hust/MAPS_ENSTDkCl-Dataset-Correction-v1.0.

3.1. Deviated notes correction methods

After data viewing and analysis, we find that deviated note whose label (t1) and true playing time (t2) have spacing over 50ms contain two features. Firstly, the time of spectrum's fastest growth spaces over 50ms from t1, and there is no obvious growth in t1-50ms to t1+50ms. At t2, the spectrum of the notes grows at a very fast rate. Secondly, the evaluation will change if we increase the evaluation tolerance. When the model detection result is consistent with t2, when we set the evaluation tolerance result as 50ms, the note is judged as multiple detection at t2 and miss detection t1. However, if we set the evaluation tolerance result as 150ms, the note will be judged as correct detection at t2.

Based on the first feature, we quantify the certain frequency component rising ratio of time point *i* as *St[i]*, which represents the amplitude rising degree within *10ms* around time point i. When the note event occurs, the corresponding frequency component will rise rapidly, therefore *St[i]* will be a large positive number; when the frequency component decreases, *St[i]* will be 0.

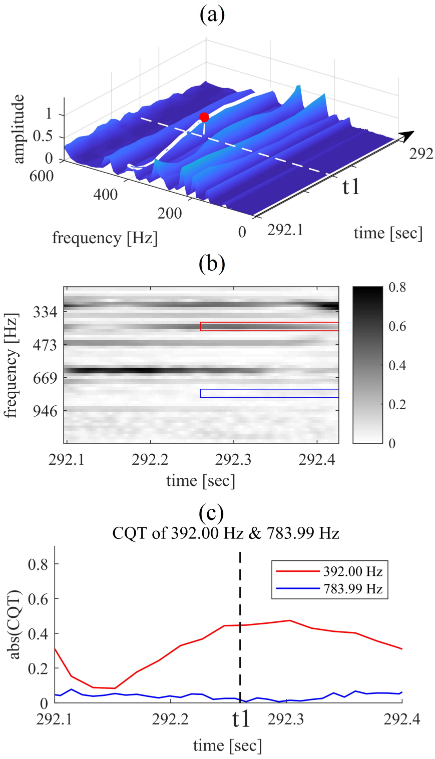

Figure 5: *The spectrum of slow start process error. (a) is a three-dimensional spectrogram, (b) is the colormap of spectrum, and (c) are the center frequency components of the fundamental frequency and second harmonics.*

According to the above analysis, we first extract the *St[t1-150ms:t1+150ms]* sequence by extracting the fundamental frequency component and the second harmonic component of all notes.

Based on the second feature, we find out the corresponding notes in the Google model's results and our model's results, calculate the intersection of the two sets. After the spectrum comparison and listening test, we find that the following characteristics exist:

$$\max(St[t1-150ms : t1+150ms]) \geq \max(St[t1-50ms \; : \; t1+50ms]) * \alpha \quad (1)$$

After statistical analysis, we set α as 4. When St sequence does not satisfy this condition, we believe that the detection error is due to slowly rising or some other reasons instead of deviated label. After the error is detected, we modify the onset time, select the two models at the detection start time of the same pitch event, average the two results as the modified onset, and the corrected end time as the corrected start time plus the duration of the note in the original midi.

In the deviatedNote directory of *Cl-correction-v1.0*, we list all the label deviated notes, and the data is arranged in the following order: (t1, t2, pitch, t3, t4), where (t1, t2, pitch) is the original midi, (t3, t4) are the corrected onset and offset.

Table 1: *Numbers of each kind of error notes in each song.*

Name	deviated	unplayed	slow	Name	deviated	unplayed	slow
alb_se2	0	0	11	mz_331_3	4	4	26
bk_xmas1	3	170	63	mz_332_2	32	41	91
bk_xmas4	2	45	16	mz_333_2	19	19	80
bk_xmas5	1	25	34	mz_333_3	0	2	22
bor_ps6	0	25	37	mz_545_3	1	1	26
chpn-e01	0	0	1	mz_570_1	28	71	238
chpn-p19	2	10	22	pathetique_1	7	25	87
deb_clai	24	24	35	schu_143_3	40	41	125
deb_menu	1	12	13	schuim-1	33	29	244
grieg_butterfly	5	2	14	scn15_11	3	9	16
liz_et6	1	6	25	scn15_12	12	15	33
liz_et_trans5	20	41	78	scn16_3	17	21	43
liz_rhap09	15	68	90	scn16_4	18	32	41
mz_311_1	16	22	41	ty_maerz	13	17	15
mz_331_2	20	18	32	ty_mai	5	8	14

3.2. Unplayed notes judgement

In the analysis in Section 2.3, it can be found that for unplayed note, the spectrum is completely degraded or basically white noise. Based on the first feature, we filter out all the notes whose spectrum is falling near the onset decision time, ie *max(St[onset-50ms : onset+50ms])=0*. We list them in set 1. Based on the second feature, the notes whose spectral components are extremely small and randomly change might not be played in the audio. Therefore, we count the spectral components of the quite period, and set the spectral median β as the threshold of the spectral component of the unplayed note, that means

$$max\left(cqt\left[onset-50ms\ :\ onset+50ms\right]\right)\leq\ \beta \tag{2}$$

We find all the notes corresponding to the second feature and list them in set 2. After the two sets are determined, we did listening test. In the unplayedNote directory of Cl-correction-v1.0, we list the notes that all tracks are not playing, and the data is arranged in the following order: (t1, t2, pitch), where (t1, t2, pitch) is the original midi.

3.3. Slow start note judgement

The slow-starting notes are characterized by a note that the duration of the note volume from 0 to the final value is significantly longer than other notes, and the spectrum rising process is slow and gentle. Therefore, we have filtered the spectrum growth value. When max(St)<γ and St[onset-50ms : onset+50ms]≥ max(St) * δ, it is defined as a slow start note. Different models may have different results on these notes, so we will list such detected notes, but do not make time correction or note existence judgement, the misjudgment of such notes may not be sufficient to indicate the validity of the research results. After the above judgment, we have screened a total of 155 slow-starting notes. In the slowStartNotes directory of Cl-correction-v1.0, we list these notes, and the data is arranged in the following order: (t1, t2, pitch), where (t1, t2, pitch) is the original midi.

4. EXPERIMENT

Based on the analysis in above chapter, we make the quantify judgment of three kinds of notes, and listing these notes in Cl-correction-v1.0, the numbers of each kind of notes are listed in Table.1, Our correction methods are as follows.

In order to verify the validity of the label correction, we compare the model evaluation results of the original dataset and the corrected dataset. If the corrected label is not detected as an error in the evaluation, and the unplayed notes are not judged to be missing, then this correction is worthwhile. In the corrected dataset, we modified the annotation of the note with the deviated label, deleted the unplayed notes, the data format is consistent with the original label. The corrected dataset is published in the correct_dataset directory of *Cl-correction-v1.0*, and the slow-start notes are listed in the slowStartNote directory.

The mir_eval library [17] is used to calculate the accuracy, recall, and F1 values that are widely used for AMT evaluation. The metrics are defined as follows:

$$P=\frac{N_{TP}}{N_{TP}+N_{FP}} \tag{3}$$

$$R=\frac{N_{TP}}{N_{TP}+N_{FN}} \tag{4}$$

$$F1=\frac{2*P*R}{P+R} \tag{5}$$

In Equation3, 4 and 5, 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. During the evaluation process, we set the time tolerance to 50ms.

Table 2: *Evaluation result of two models*

	Result of Google model [12]			Result of our model [15]		
	F	P	R	F	P	R
original dataset	84.34%	85.95%	83.05%	85.09%	87.8%	82.83%
corrected dataset	85.94%	89.77%	82.73%	85.82%	88.5%	83.59%

Based on the revised annotations, we performed the accuracy test as shown in Table 2. It can be found that the performance of the two models has increased in the evaluation results of the revised data set. In the process of data error review, we find that most of the errors detected by the model are errors caused by generalization errors, and the number of common errors is also reduced, so we think this data correction is effective.

5. SUMMARY AND FUTURE WORK

In the field of AMT, dataset construction has always been a difficult problem. Synthetic audio can't completely replace the audio of real piano performance. The authenticity of audio and the authenticity of labels cannot be well unified. Therefore, from the perspective of audio spectrum and human listening, we corrected some unreasonable labels, including deviation labels and unplayed labels, and listed the slow-starting notes in the audio. Finally, the validity of the corrected dataset was verified in two different model evaluation. We published our result on the Cl-correction-v1.0, providing a reference for the future work of subsequent researchers. In the future research, we will try to figure out the reasons why the spectrum will have slow rising of the slow-start notes, and summarize some other problems in the dataset.

6. ACKNOWLEDGEMENTS

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

7. REFERENCES

[1] A. P. Klapuri and M. Davy, Eds., “Signal Processing Methods for Music Transcription,” New York, NY, USA: Springer, 2006.

[2] C. Duxbury, M. Sandler, and M. Davies, “A hybrid approach to musical note onset detection”, *Proc. Digital Audio Effects Conf. (DAFX,02)*, Hamburg, Germany, pp.33–38, 2002.

[3] X. Shao, M.C. Maddage, C. Xu, and M.S. Kankanhalli, “Automatic music summarization based on music structure analysis,” *IEEE International Conference on Acoustics, Speech, and Signal Processing, Proceedings*, Philadelphia, Pennsylvania, USA, pp.1169–1172, 2005.

[4] M. Müller, D. P. Ellis, A. Klapuri, and G. Richard, “Signal processing for music analysis,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1088–1110, 2011.

[5] M. Schedl, E. Gómez, and J. Urbano, “Music information retrieval: Recent developments and applications,” *Foundations and Trends in Information Retrieval*, vol. 8, pp. 127–261, 2014.

[6] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri, “Automatic music transcription: challenges and future directions,” *Journal of Intelligent Information Systems*, vol. 41, no. 3, pp. 407–434, Dec. 2013.

[7] N. H. Fletcher and T. D. Rossing, “The Physics of Musical Instruments,” New York, NY, USA: Springer-Verlag, 1991.

[8] V. Emiya, R. Badeau, B. David, “Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1643–1654, 2010.

[9] A. Ycart, E. Benetos, “A-MAPS: Augmented MAPS dataset with rhythm and key annotations,” in *19th International Society for Music Information Retrieval Conference Late-Breaking Demos Session*, 2018.

[10] S. Ewert, M. Sandler, “Piano transcription in the studio using an extensible alternating directions framework,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no.11, pp. 1983-1997, 2016.

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

[12] C. Hawthorne, E. Elsen, J. Song, A. Roberts, I. Simon, C. Raffel, J. Engel, S. Oore, D. Eck, “Onsets and frames: Dual-objective piano transcription,” in *Proc. International Society for Music Information Retrieval Conference*, 2018

[13] R. Kelz, M. Dorfer, F. Korzeniowski, S. Böck, A. Arzt, “On the potential of simple framewise approaches to piano transcription,” arXiv preprint arXiv:1612.05153, 2016.

[14] S. Böck, M. Schedl, "Polyphonic piano note transcription with recurrent neural networks,” *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 121-124, 2012.

[15] K. Sicong, X. Wei, L. Wei, G. Xuan, L. Juanting, C. Wenqing, “Onset-aware polyphonic piano transcription: A CNN-based approach,” in *Proceedings of 2019 the 9th International Workshop on Computer Science and Engineering*, 2019. (In press).

[16] H. Tianqian, "Effect of striking point and striking dynamics on the amplitude spectrum of piano signals,” *Audio Engineering*, 2003.

[17] C. Raffel, B. McFee, “mir_eval: A transparent implementation of common MIR metrics,” In *Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR*, 2014.