Escaping from the Abyss of Manual Annotation: New Methodology of Building Polyphonic Datasets for Automatic Music Transcription

Li Su and Yi-Hsuan Yang

Center for Information and Technology Innovation, Academia Sinica
128 Academia Road, Section 2, Nankang, Taipei 115, Taiwan
{lisu,yang}@citi.sinica.edu.tw
https://sites.google.com/site/lisupage/

Abstract. While recent years have witnessed large progress in the algorithm of automatic music transcription (AMT), the development of general and sizable datasets for AMT evaluation is relatively stagnant, predominantly due to the fact that manually annotating and checking such datasets is labor-intensive and time-consuming. In this paper we propose a novel note-level annotation method for building AMT datasets by utilizing human's ability in following music in real-time. To test the quality of the annotation, we further propose an efficient method in qualifying an AMT dataset based on the concepts of onset error difference and the tolerance computed from the evaluation result. According to the experiments on five piano solos and four woodwind quintets, we claim that the proposed annotation method is reliable for evaluation of AMT algorithms.

Keywords: Automatic music transcription, multipitch estimation, note tracking, onset error difference.

1 Introduction

Evaluating the performance of an automatic music transcription (AMT) algorithm always requires a dataset containing music excerpts in which every note has annotations of onset and pitch. A sizable dataset covering various classes of music and with accurate note-level annotation would be very helpful for AMT research. However, by now such a dataset is virtually unattainable since annotating the notes in polyphonic music is a tedious process, and there is no straightforward and efficient way of either creating ground truth for multiple-instrument music or checking whether the ground truth is correct [1], [7]. Up to the present, most AMT algorithms perform experiments only on piano data because it is relative easy to generate them by a MIDI-controlled piano [1]. Without an efficient

1 In this paper we restrict the scope of AMT in the subtask of onset-only note tracking. Details can be found on the webpage of MIREX Multi-F0 Challenge [12].
2 In this paper the term “music excerpt” means a segment (usually 20-30 seconds) of audio content taken from a longer music composition.
method of note-level annotation, building sizable and general AMT datasets can be a boring job. Therefore, in comparison to the recent progresses of new AMT algorithms, the development of new and sizable datasets for AMT evaluation is quite slow and insufficient. The science on building an AMT dataset, including the scale, the efficiency and the quality of the annotations, is hardly questioned and investigated.

The purpose of this paper is not to introduce new AMT algorithms. In contrast, we concentrate on the methodology in building an AMT dataset. We expect that a good methodology should not only permit accurate note-level annotation, but also make it easy to build and extend the dataset. Moreover, its contents should be able to represent the music content encountered in real-world application. To discuss these requirements, we define the following criteria to evaluate the goodness of a dataset:

1. **Generality**: The form, genre or instrumentation of the music excerpts should be representative in the music universe and reveal the situations in real-world application. The dataset is not restricted to one class of instrument (e.g., piano solo only) or one specific form of music (e.g., woodwind quintet only). If we wish to apply the AMT algorithm for general online music contents, the dataset should contain real-world music played in real instruments by human and (sometimes) recorded in real environment rather than synthesized signals.

2. **Efficiency**: The annotation process should be fast rather than slow, knowledge intensive rather than labor intensive. Efficiency sometimes implies scalability, meaning that we can expand the dataset easily.

3. **Cost**: The cost of money and human resource for building the dataset should be minimized if possible.

4. **Quality**: The annotation should be accurate enough so as to evaluate the performance of an AMT algorithm in a proper manner. Although finding “correct” timing of the onsets and offsets is challenging, the annotation should be closed to the correct point so that the performance of an AMT algorithm can be evaluated accurately within a range of tolerance (e.g., in the standard of MIREX Multi-F0 Challenge, the tolerance is 50 ms.).

In terms of these criteria, the contributions of this paper are two-fold: A general, efficient and low-cost method for note-level annotation, and an efficient method in qualifying an AMT dataset. Given that the experience of annotating polyphonic music is actually similar to playing in an orchestra, the proposed annotation method utilizes the musician’s technique and behavior by setting an interface (in this paper it is an electric piano with MIDI output) transferring the musician’s playing into annotations.

Noticing that the evaluation procedure actually incorporates the concept of the tolerance to determine whether a note is correctly detected, we argue that the difference between detected note onset and the labeled note onset within the tolerance (named as onset error difference) could be a good indicator of the quality of an AMT dataset. Accordingly, we define criteria qualifying a dataset in terms of these quantities.
Section 2 reviews previous methods in building AMT datasets. Section 3 describes the proposed method, the technical details and experiment results on an AMT algorithm. Section 4 presents how we test the quality of this dataset. Section 5 concludes this paper.

2 Related Work

We categorize the previously proposed datasets available for AMT into 5 (overlapping) groups according to the method adopted to build the dataset: 1) the manual method, 2) the multi-track method, 3) the audio-score alignment method, 4) the autopiano method and 5) the synthesis method.

First, the manual method means that when building the dataset, we listen to the music excerpt and annotate every onset time and the corresponding note name of every note manually. This task is sometimes accomplished with the aid of audio visualization and musical signal analysis tools (e.g., Sonic Visualizer\(^3\)) and music scores. One example using this method is the annotations of MIREX 2007 dataset, where the onsets and offsets are determined by setting an amplitude threshold \(^1\). The manual method is the most direct method and is available for all real-world music without any restriction of recording environment, instrument, etc., therefore it completely satisfies the merit of generality. However, this method is notoriously labor intensive, inefficient and challenging. Similarly, checking the annotations and improving the quality of the dataset are equivalently insufficient. Previous studies have also commented on this \(^1\), \(^7\).

Up to the present, there is still no sizable and fully manually-built dataset providing complete information for AMT.\(^5\)

All the other four annotation methods are basically proposed to improve the efficiency of annotation at the cost of sacrificing generality. The multi-track method means that all the polyphonic music excerpts are mixed from separated parts of monophonic tracks. The monophonic tracks are easier to be labeled by the manual method or by single-pitch detection algorithm like YIN \(^14\). The greatest advantage of this method is that it downscales the complexity of annotating polyphonic music into monophonic music and is therefore relatively efficient. Another advantage is that we can produce a large variety of music excerpts by mixing the monophonic track in different ways and different polyphony levels. For example, when we have 5 separated parts of an excerpt of wood quintet, we can generate \(2^5 - 1 = 31\) excerpts with number of sources ranging from 1 to 5. The drawback of the multi-track method is its high cost, as it often requires musicians specialized in various instruments, professional recording equipments and studio, etc. Examples of AMT datasets built in this way are MIREX 2007 \(^12\) and Bach10 \(^3\), etc.

\(^3\) http://www.sonicvisualiser.org/
\(^4\) To paraphrase J.P. Bello et. al \(^7\): “Hand-marking is a painful and time-consuming task that leaves no room for the cross-validation of annotations.”
\(^5\) Some onset detection datasets are built fully in manual method, but the datasets do not provide pitch information \(^13\), \(^8\).
The audio-score alignment method utilizes an audio-score alignment algorithm to align a music excerpt and its corresponding MIDI. Note-level annotation is thereby obtained from the aligned MIDI. One example using this method is the set of syncRWC annotations based on the alignment framework proposed by Ewert et al. [6]. This method is undoubtedly efficient and low-cost, but its scope is available for only those music pieces having well-edited MIDI files.

The autopiano method typically utilizes a MIDI-controlled piano (e.g., Yamaha Disklavier) to generate the music excerpts (played mechanically with the piano) from known note-level annotation (MIDI contents). Since it is easy to generate music excerpts in this way, many AMT studies perform experiments in this way, such as the datasets created by Dessein et al. [9], Poliner and Ellis [10], and the MAPS dataset by Emiya et al. [4]. Similarly, the method is efficient but not general as it is limited to only one instrument.

The most efficient way to generate an AMT dataset is using a synthesizer to generate music excerpts directly from the MIDI files. In addition, the quality of the annotations are exact since the onset and pitch of the audio part is just identical in the MIDI part. An example of such kind of dataset for AMT is the TRIOS dataset [5], [11]. However, synthesized music excerpt is not able to represent lots of real-world audio contents. In most applications of AMT, such as the transcription in a live concert, the synthesized dataset is unsuitable for evaluation. It is worth mentioning that since the track information can be controlled in the multi-track and the synthesized methods, datasets built with these two methods can be used not only for AMT but also melody tracking and source separation.

From the discussion above, we found that current annotation methods cannot balance the tradeoffs among generality, efficiency and cost. Under some cases the quality can even not be controlled since we have to rely on the inefficient manual method to check the quality. Whenever we wish to verify whether a dataset is good, we always fall into “the abyss of manual annotation”.

3 Proposed Method

3.1 Overview

Our task of annotating a note is to find the onset time and the note name of the given note event in the music excerpt and on the score sheet. This process is actually quite similar to the case when a musician plays music in an orchestra: the musician follows the mixture of music played by all other members in the orchestra, and plays his or her part together with all other members. When the musician is doing this, he or she is actually annotating (producing onsets and note names) his or her own part along with time.

6 http://c4dm.eecs.qmul.ac.uk/rdr/handle/123456789/27
7 In our scenario we just need to find the “note name” rather than the accurate “fundamental frequency”. More specifically, we would not discriminate the fundamental frequencies of 440 Hz, 438 Hz or 442 Hz. Instead, we just need to identify “A4”. In other words, our pitch detection task allows an error up to a half semitone (±3%).
To leverage this to note-level annotation work, we need to find an interface which transfers the musician’s perception of note event and the behavior of playing into note-level annotation. An electric piano with MIDI output is the interface we want. A musician with both orchestra and piano playing experience would be able to do note-level annotation by “playing with the music excerpt”. Providing score sheets, the musician can further follow every parts of the music excerpts fluently and without significant errors. The detailed procedure of our method is described as follows (also see Fig. 2):

1. Find a pianist with band/orchestral experience.
2. Prepare an electric piano with MIDI output. Prepare the score sheets of the music excerpt we want to annotate (If the musician is confident to find all parts in the music then the score sheet in not necessary).
3. Let the musician listen to the music, preview the score sheet, transpose the scores of transposing instruments (e.g., clarinet and horn in woodwind quintet) for playing piano, and practice.
4. Let the musician play the piano with the music following the score sheet. The piano therefore automatically records the information the musician follows in the format of MIDI. When annotating piano solo, the musician just plays the same thing as the audio contents. When annotating multiple-instrument music, the musician imagines he or she as a member in the music ensemble/orchestra and plays his or her own part (not necessary piano) with the full group. This process is then repeated part by part. For example, to annotate a string quartet, the musician plays the part of first violin solely with the audio, then the second violin, viola, and cello. The audio is played four times and the musician follows the music part by part, thereby completing the annotation of each part by recording it in the MIDI file.
5. Finally, the generated MIDI file is manually calibrated and checked if needed.
Table 1. The music excerpts used for the experiments. PS1 to PS5 are piano solo, and WQ1 to WQ4 are wood quintet.

| No. | Composer         | Name                                                                 |
|-----|------------------|----------------------------------------------------------------------|
| PS1 | L.V. Beethoven   | Moonlight Sonata, Mov. 1, mm. 1–9                                    |
| PS2 | F. Chopin        | Nocturne No 9, Op. 32–1, mm. 1–8                                     |
| PS3 | W.A. Mozart      | Piano Sonata No. 16, KV545, Mov.1, mm. 1–12                          |
| PS4 | J. Brahms        | Waltz Op.39, No.15, mm. 1–16                                        |
| PS5 | R. Schumann      | Scenes from Childhood, Op. 15:7 (Dreaming), mm. 1–8                 |
| WQ1 | C. Nielsen       | Quintet, Op. 43, Mov. 1, mm. 13–20                                   |
| WQ2 | A. Schönberg     | Quintet, Op. 26, Mov. 4, mm. 18-27                                  |
| WQ3 | G.M. Cambini     | Quintet, No. 1 in B-flat, Mov. 1, mm. 7-17                           |
| WQ4 | F. Danzi         | Quintet, op. 56 No. 1 in B-flat, mm. 1-12                            |

The proposed method can be used to annotate any real-world music with score sheets. Therefore, it is more general than the autopiano method (which is only for audio contents of piano), the synthesized method (which is only for synthesized audio contents) and the audio-score alignment method (because we can get score sheets for many kinds of music which has no available MIDI). The cost of the method is expected to be lower than the multi-track method. Also, this method can be used to improve the efficiency of the multi-track method.

### 3.2 Discussion

There are some potential restrictions of this method. First, an electric piano is limited in annotating some special music contents such as unusual playing techniques (e.g., slides in guitar or trombone) or voices with pitch variation smaller than a semitone. Therefore, we only choose the music excerpts without pitch variation smaller than a half tone. At the same time we also need to notice that using piano we can only annotate the pitch up to a level of half tone.

Second, in comparison to the scenario of playing in a real orchestra, our annotation scenario (Figure 2) still lacks some critical information, such as the verbal communication or the gestures among the performers. Therefore, the musician may not be able to track the beats instantaneously and perfectly for tempo variation, such as tempo rubato, *accelerando* (gradually accelerating), *ritardando* (gradually slow down), and fast note sequences such as *trill* and *arpeggiato*. For the above cases, the annotation is more likely to incorporate the musician’s own interpretation. To investigate what the above cases influences the musician's annotation, we consider some music excerpts with tempo variation and fast note sequences in the experiment.

### 3.3 Experiment

The musician working at the creation of the dataset has more than 20 years of piano playing and teaching. The musician is also a professional clarinet player, and had more than 15 years of experience in orchestra. The musician is asked to
play 9 musical excerpts including 5 piano solos and 4 woodwind quintets (flute, oboe, clarinet, horn and bassoon). Each excerpt has a length of 20 to 30 seconds. Detailed information is listed in Table 1. The music excerpts are selected arbitrarily from the recordings of professional classical music players (information of the CDs will be listed in the supplementary materials). Some of the excerpts are relatively challenging for annotation due to the aforementioned situation: PS2, PS5 and WQ2 have significant tempo variation, PS4 has arpeggiato in the left-hand part, and WQ3 has some fast note groups including trill. PS3, WQ1 and WQ4 have relatively stable tempo. Details of these music excerpts and the annotations are publicly available online.⁸

The musician is asked to play with the audio as accurate as she can. Each excerpt is then checked by three amateur piano players with band experiences. The checking method is to play the audio excerpt and the annotated MIDI synchronously, and carefully find mismatch between the two. Mismatches are fixed and the audio and the MIDI are replayed and checked, until all mismatched have been fixed. Each excerpts is checked twice. Although it is hard to record accurate time spent on annotation and checking, in our experiment we found that, starting from previewing the scores to finishing the recordings, it takes about 75 minutes for the musician to finish a piano solo excerpt and 1.5 hour to finish a woodwind quintet excerpt. Checking one music excerpt takes roughly from 15 minutes to more than 2 hours.

3.4 Note-level Evaluation

To get a brief picture of this new dataset, we use an AMT algorithm, the constrained non-negative matrix factorization (C-NMF) algorithm proposed by Vincent et al. [2]⁹ for our experiments. By default, it uses the NMF algorithm to learn one template per pitch adaptively from the signal being analyzed on-the-fly, with constraints that enforce harmonicity and spectral smoothness of the learned templates. Because the performance of C-NMF appears to be sensitive to the value of β for computing the β-divergence¹⁰ and the value of θ for thresholding the activation patterns, we optimize the values of β and θ for each music excerpt independently and empirically by running a grid search.

We compare three different cases:

1. **Original**: The basic case, where the AMT algorithm is performed on the real-world music excerpts, and the F-scores are computed by taking the unchecked annotations generated by our musician as the ground truth.

2. **Checked**: The algorithm is also processed on the real-world music excerpts, but the ground truth is taken from the checked annotation.

---

⁸ https://sites.google.com/site/lisupage/research/new-methodology-of-building-polyphonic-datasets-for-amt
⁹ http://www.irisa.fr/metiss/members/evincent/multipitch_estimation.m
¹⁰ The family of β-divergence includes the Euclidean distance (β = 2), Kullback-Leibler divergence (β → 1) and Itakura-Saito divergence (β → 0).
3. **Oracle**: This is an ideal but unrealistic case, where the music excerpts are not the real-world ones, but the ones synthesized by the MIDI files of the unchecked annotation. The ground truth is also taken from the unchecked annotations. In this case, all annotations are considered perfectly matched to the audio contents, and the audio contents are clean and without noise or any real-world artifact. The synthesized audio contents are obtained from Edirol Hyper Canvas.\(^{11}\)

These three cases will also be useful in the discussion in the next section.

### Table 2. Note-level evaluation results (F-scores in %).

| No. | Original | Checked | Oracle | No. | Original | Checked | Oracle |
|-----|----------|---------|--------|-----|----------|---------|--------|
| PS1 | 41.18    | 42.62   | 73.96  | WQ1| 30.25    | 28.69   | 64.29  |
| PS2 | 40.51    | 41.35   | 88.31  | WQ2| 23.85    | 22.12   | 32.73  |
| PS3 | 76.16    | 76.16   | 85.45  | WQ3| 39.35    | 39.35   | 28.24  |
| PS4 | 30.99    | 33.82   | 69.61  | WQ4| 40.87    | 38.79   | 41.47  |
| PS5 | 32.67    | 35.33   | 70.31  |     |          |         |        |

We evaluate the performance of the algorithm in terms of note-level F-score. Here, a ground truth note is assumed to be correctly transcribed if the algorithm returns a note that is within a half semitone of that note, and the returned onset of the note is within ±δ ms of the onset of the ground truth note. Here δ is called the tolerance; in MIREX δ = 50 ms \([12]\). Note-level F-score is defined as \(F = 2PR/(P+R)\), where \(P = N_{TP}/(N_{TP}+N_{FP})\) and \(R = N_{TP}/(N_{TP}+N_{FN})\) (\(N_{TP}, N_{FP}, N_{FN}\) are the number of true positives, false positives and false negatives, respectively).

Table 2 shows the results of each music excerpt by setting δ = 50 ms. Each F-score is obtained from optimal β and ϑ. We can see that, when there is only one instrument (piano solo), the Oracle results outperform other cases from about 10% (PS3) to even 47% (PS2). For wind quintets the difference is smaller due to higher complexity of the audio signal, but the Oracle case is still better on average. The good performance of the Oracle reveals how good (and also how unrealistic) it would be if we use synthesized data in AMT evaluation.

Table 2 also indicates that the checked annotations do not necessarily lead to better F-scores. Specifically, the Checked case of the four wind quintets are even worse than the Original case, perhaps due to the systematic bias of human. In general, there is marginal improvement in F-scores after checking.

Two issues emerge. First, the F-score listed here still answer nothing specific about the quality of the dataset. High F-scores can be resulted from multiple factors: perhaps the annotation is accurate, perhaps the audio content is clean, perhaps the algorithm is good, or perhaps the specific excerpt is simple. More specifically, given one music excerpt, the F-score that can be obtained is jointly

\(^{11}\) [http://www.roland.com/products/hq_hyper_canvas/](http://www.roland.com/products/hq_hyper_canvas/)
influenced by the algorithm and the annotation. Before we identify the errors stemming from these two sources we cannot make any conclusion. Second, from the experiment we also learn that checking is not surely to make the data more reliable. It is likely that we are careless somewhere and we should have spent more time on the manual checking process so as to make them better. But should we keep spending time to improve the annotation manually and stay in the abyss of manual annotation?

These issues will be addressed in the next section.

4 Evaluating the Quality of an AMT Dataset

4.1 The Issue

Studies on AMT mostly concentrate only on the performance of the AMT algorithm rather than the quality of the AMT dataset. This is due to the fact that building a dataset of real-world music with exact annotation is nearly impossible, even when we use the tedious manual method – this is already not what we expect. In this section, we would like to answer the following two questions: 1) How to qualify a dataset in a general sense without manual checking? and 2) for the proposed method in Section 3, can we use the musician’s original annotation for evaluation of AMT without manual checking?

The purpose of pursuing high quality annotation is not necessarily to make it as perfect as possible. Instead, our main purpose is to make the dataset able to reveal the performance of the AMT algorithm. Although we cannot build a dataset with exact annotation, we can argue that as the labels are closer to the exact ones, we can use smaller tolerance to evaluate the AMT algorithm. The conception of “good annotation” is inexact; it depends on the value of tolerance. The measure the quality of the dataset is a function of the tolerance.

Therefore, we are more interested in the relationship between the labeled onset and the detected onset, rather than the performance of the algorithm. For convenience in discussion, we define the onset error difference (denoted by $\epsilon$) as the difference between the detected onset and the labeled onset given a finite tolerance window. Notice that $\epsilon$ can only be defined for TPs since FPs and FNs are not “detected” onsets.

4.2 Meanings of the Onset Error Difference

To investigate the meanings of $\epsilon$, firstly we need to clarify where $\epsilon$ stems from. In general, $\epsilon$ is contributed by two sources: annotation error and algorithmic error, where the former is caused by the musician and the latter is the caused by the engineer who designs the algorithm. The two sources cannot be clearly separated. To simplify the problem, we assume that short-time onset error differences (e.g., $\epsilon < 50$ ms) are mostly caused by bad annotation, whereas long-time onset error differences (e.g., $\epsilon > 200$ ms) are caused mostly by bad AMT algorithm.\textsuperscript{12} This

\textsuperscript{12} Also, we assume that FPs and FNs are also caused by bad AMT algorithm.
assumption is reasonable since we do not expect a professional musician to make $\epsilon > 200$ ms. If $\epsilon$ is small for small tolerance $\delta$ then we are confident about the quality of the annotation.

The criteria of the quality of the dataset is described as follows by comparing the average onset error difference of the original case ($\epsilon_R$), the checked case ($\epsilon_C$) and the oracle case ($\epsilon_O$) for all TPs in every music excerpt:

1. In the worst case, assuming $\epsilon$ is uniformly distributed in $[0, \delta]$,\textsuperscript{13} then, for usable annotation, the averaged $\epsilon_R$ of a music excerpt should be smaller than $\delta/2$, especially for large tolerance window. Conversely, if $\epsilon_R > \delta/2$ for large $\delta$, then we claim the annotation is surely unusable.

2. In the ideal case, $\epsilon_R$ should be as close to $\epsilon_O$ as possible. Here we further propose two criterions to test the closeness. The weak criteria states that, by defining $\Delta = \epsilon_R - \epsilon_O$, $\Delta$ should be less than 10 ms for $\delta = 50$ ms. The strong criterion states that through a two-tailed t-test, $\epsilon_R$ should not be significantly larger than $\epsilon_O$ for $\delta = 50$ ms. If the weak criterion is satisfied then we claim the annotation is good. If the strong criterion is satisfied then we claim the annotation is high-quality.$^{14}$

Table 3. $\epsilon_R$, $\Delta$, p-value and (d.f.) The unit of $\epsilon_R$ and $\Delta$ is in milliseconds.

| No. | $\epsilon_R$ | $\Delta$ | p-value (d.f.) | No. | $\epsilon_R$ | $\Delta$ | p-value (d.f.) |
|-----|--------------|----------|----------------|-----|--------------|----------|----------------|
| PS1 | 27.1         | 2.12     | 1.63E-1 (171)  | WQ1 | 26.4         | 6.39     | 2.51E-2 (106)  |
| PS2 | 26.7         | 10.8     | 5.15E-8 (148)  | WQ2 | 25.3         | 6.90     | 3.09E-2 (57)   |
| PS3 | 21.8         | 8.18     | 4.66E-8 (278)  | WQ3 | 21.8         | 9.99     | 5.26E-3 (81)   |
| PS4 | 23.3         | 5.48     | 4.29E-3 (212)  | WQ4 | 24.5         | 0.69     | 4.19E-1 (90)   |
| PS5 | 28.8         | 6.78     | 8.72E-4 (137)  |     |              |          |                |

4.3 Results

Figure 2 shows the results of the average onset error differences with various lengths of the tolerance under the three cases for each music excerpt.

For most excerpts, although the average $\epsilon_R$ and $\epsilon_C$ are slightly higher than $\delta/2$ (the diagonal thin line) for $\delta < 50$ ms, they are constantly smaller than $\delta/2$ for $\delta$ larger than 100 ms. As discussed earlier, $\epsilon_O$ mainly stems from the AMT algorithm rather than annotation for larger $\delta$. Therefore, we may say that the proposed method is reliable. Table 3 lists the values of $\epsilon_R$ for $\delta = 50$ ms.

The other important issue is to see how close $\epsilon_R$ (and also $\epsilon_C$) and $\epsilon_O$ are. Figure 2 shows that this depends on the music excerpts. For the excerpt with

\textsuperscript{13} Obviously, the onset error difference is larger than zero and smaller than the tolerance $\delta$.

\textsuperscript{14} These criteria are arbitrary and are depending on how accurate we need for the annotation in real application.
Fig. 2. Results of average $\epsilon_R$, $\epsilon_C$, and $\epsilon_O$ with varying $\delta$ for each music excerpt. Grey thin line: baseline $(\delta/2)$. Thick black line: $\epsilon_R$. Thin black line with ‘+’ marks: $\epsilon_C$. Dash-dot line: $\epsilon_O$.

stable tempo (e.g., PS03), $\epsilon_R$, $\epsilon_C$, and $\epsilon_O$ are generally small; $\epsilon_R$ is even smaller than 50ms for $\delta = 300$ms. Not surprisingly, PS04 (arpeggiato in the left-hand part) and WQ02 (unstable tempo) have relatively high $\epsilon$. In addition to the brief view of Figure 2, Table 3 lists $\Delta$ and the p-value at $\delta = 50$ ms of every music excerpt for more detailed and rigorous comparison. Most music excerpts satisfy the weak criterion that $\Delta < 10$ ms excerpt for PS02. This is because the use of tempo rubato, a fermata and the term stretto in the sixth measure, and the term poco ritenuto in the seventh measure may have caused the musician unable to follow the music perfectly without any personal interpretation.

We also see that only PS1 and WQ4 satisfy the strong criterion, unfortunately. In fact, PS1 and WQ4 are both of slow and relative stable tempo (PS2 and PS5 are also slow but their tempo varies over time). Besides, the p-values of WQ01 and WQ02 are both near to 0.05.

Finally, comparing $\epsilon_C$ and $\epsilon_R$ in Figure 2 we found marginal improvement of checking. This result may stem from two reasons: the first is that the original annotation already “imprints” a position to the checking person and makes he or she believe the original annotation is correct. The second reason is that we do not pay enough labor and time in checking. However, we argue that the first is unavoidable while the second is unwanted. Therefore, to improve the quality, we opt for improving the technical details of our annotation scenario rather than pay more labor and time in calibration.
In summary, our proposed method can produce usable annotation for all music excerpts in the experiment. For music with slight tempo variation and medium speed, our method can produce good annotation. Moreover, for slow music excerpts with stable tempo, our method can produce high-quality annotation.

5 Conclusion

In this paper, we have firstly addressed the issue on how to build and qualify a sizable dataset for AMT without manual work. By clarifying the criteria for a good dataset, we have proposed a general, efficient and low-cost method for note-level annotation, and an efficient way in qualifying the quality of an AMT dataset. We have demonstrated the experiments, reported technical details and results, and verified its feasibility. For future work, we see two directions for improvement: the first is to test the quality of the offsets in addition to onsets and note names, and the second is to consider some scenarios for improving quality, e.g., collecting annotations from more than one musician. It is hoped that the presented methodology can help open the door of AMT research to more diverse training and test sets to cross the gap between scholarly curiosity to commercial applications.

Acknowledgments. The authors would like to thank Patricia Hsu for performing the experiment of the proposed annotating method. Also the authors thank Che-Yuan Kevin Liang, Yuan-Ping Chen and Pei-I Chen for checking the annotations. This work was supported by a grant from the Ministry of Science and Technology under the contract MOST 102-2221-E-001-004-MY3 and the Academia Sinica Career Development Program. Dr. Li Su was further supported by the postdoctoral fellowship from the Academia Sinica.

References

1. Benetos, E., Dixon, S, Giannoulis, D., Kirchhoff, H., Klapuri, A.: Automatic Music Transcription: Challenges and Future Directions. J. Intelligent Information Systems, vol. 41, no. 3, pp. 407–434 (2013)
2. Vincent, E., Bertin, N., Badeau, R.: Adaptive Harmonic Spectral Decomposition for Multiple Pitch Estimation. IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 3, pp. 528–537 (2010)
3. Duan, Z., Pardo, B., Zhang, C.: Multiple Fundamental Frequency Estimation by Modeling Spectral Peaks and Non-peak Regions. IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 8, pp. 2121–2133 (2010)
4. Emiya, V., Badeau, R., David, B.: 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)
5. Fritsch, J.: High Quality Musical Audio Source Separation. Master Thesis, Queen Mary Centre for Digital Music (2012)
6. Ewert, S., Müller, M., Grosche, P.: High Resolution Audio Synchronization Using Chroma Onset Features. In: Proc. IEEE International Conference on Acoustics Speech, and Signal Processing, pp. 1869–1872 (2009)

7. Bello, J.P., Daudet, L., Sandler, M.B.: Automatic Piano Transcription Using Frequency and Time-domain Information. IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 6, pp. 2242–2251 (2006)

8. Holzapfel, A., Stylianou, Y., Gedik, A., Bozkurt, B.: Three Dimensions of Pitched Instrument Onset Detection. IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 6, pp. 1517–1527 (2010)

9. Dessein, A., Cont, A., Lemaitre, G.: Real-time Polyphonic Music Transcription with Nonnegative Matrix Factorization and Beta-divergence. In: 11th International Society for Music Information Retrieval Conference, pp. 489–494 (2010)

10. Poliner, G., Ellis, D.: Adiscriminative Model for Polyphonic Piano Transcription. EURASIP Journal on Advances in Signal Processing, vol 8, pp. 154–162 (2007)

11. Benetos, E., Cherla, S., Weyde, T.: An efficient shift-invariant model for polyphonic music transcription”. In: 6th International Workshop on Machine Learning and Music (2013)

12. MIREX 2014 Multiple Fundamental Frequency Estimation and Tracking Challenge: http://www.music-ir.org/mirex/wiki/2014:Multiple_Fundamental_Frequency_Estimation\_\&\_Tracking

13. MIREX 2014 Audio Onset Detection Challenge, http://www.music-ir.org/mirex/wiki/2014:Audio_Onset_Detection

14. Cheveigné, D.A., Kawahara, H.: YIN, a Fundamental Frequency Estimator for Speech and Music. The Journal of the Acoustical Society of America, vol. 111, no. 4, pp. 1917–1930 (2002)