IN-DEPTH PERFORMANCE ANALYSIS OF THE ADTOF-BASED ALGORITHM FOR AUTOMATIC DRUM TRANSCRIPTION

Mickaël Zehren
Umeå Universitet
mzehren@cs.umu.se

Marco Alunno
Universidad EAFIT Medellín
malunno@eafit.edu.co

Paolo Bientinesi
Umeå Universitet
pauldj@cs.umu.se

ABSTRACT

The importance of automatic drum transcription lies in the potential to extract useful information from a musical track; however, the low reliability of the models for this task represents a limiting factor. Indeed, even though in the recent literature the quality of the generated transcription has improved thanks to the curation of large training datasets via crowdsourcing, there is still a large margin of improvement for this task to be considered solved. Aiming to steer the development of future models, we identify the most common errors from training and testing on the aforementioned crowdsourced datasets. We perform this study in three steps: First, we detail the quality of the transcription for each class of interest; second, we employ a new metric and a pseudo confusion matrix to quantify different mistakes in the estimations; last, we compute the agreement between different annotators of the same track to estimate the accuracy of the ground-truth. Our findings are twofold: On the one hand, we observe that the previously reported issue that less represented instruments (e.g., toms) are less reliably transcribed is mostly solved now. On the other hand, cymbal instruments have unprecedented relative low performance. We provide intuitive explanations as to why cymbal instruments are difficult to transcribe and we identify that they represent the main source of disagreement among annotators.

1. INTRODUCTION

Automatic Music Transcription (AMT) is a particularly important task in music information retrieval because it provides access to many high-level features of a musical track, such as its structure, melody, and rhythm. A subtask of AMT is automatic drum transcription in the presence of melodic instruments (DTM), which focuses on the estimation of the onsets of drum sounds and the identification of what drum instruments play them. In this article, we focus on DTM and specifically on the transcription of drum and cymbal sounds.

Recently, Zehren et al. presented a DTM algorithm, which we refer to as "ADTOF-based" algorithm, based on supervised learning from abundant crowdsourced annotations [1]. Thanks to the size and diversity of the datasets, the algorithm surpasses the accuracy of the previous state-of-the-art [2]. However, the resulting models are still not perfect, as their estimations contain mistakes. In this work, we carefully investigate these state-of-the-art algorithms, aiming to identify the most common sources of errors.

We evaluated the models in two distinct conditions: (i) when the training and the testing take place on different datasets (out-of-domain), and (ii) when they take place on the same dataset (on-domain). In the first case, the model is not expected to achieve perfect accuracy because of generalization errors that can be attributed to differences between testing and training data. In the second case, testing on-domain, the errors are more concerning as they suggest flaws in the algorithm; in fact, if the dataset were large enough, the model would be expected to learn the data distribution and therefore achieve nearly perfect accuracy. Thus, in this study we focus specifically on the most common errors that arise in the latter case. This was done in three steps, as described in the following.

First, in order to identify the most difficult instruments to transcribe, we independently evaluated the performance of the models on the different instrument classes. When trained and evaluated on (a different split of) the crowdsourced datasets, we observed that the models can reliably transcribe those instruments that play less often, something that in previous studies was arguably problematic to achieve. On the flip side, we also observed that the models do not transcribe cymbals as precisely as drums.

Second, to understand why cymbals are problematic, we employed both a new metric, which we named "octave F-measure", and a pseudo confusion matrix. Through the new metric, we identified that the models often mistook the beat subdivision at which cymbals are played. Specifically, the rhythm estimated is often half or double the speed of the ground truth (e.g., eighth notes are estimated instead of quarter notes). Through the pseudo confusion matrix, we showed that different kinds of cymbals are hard to discern.

Finally, we assessed how much the quality of crowdsourced annotations affected the evaluated performance of the models. Due to discrepancies in the labels, some of the correct estimations from the models could have been mistakenly reported as errors. To estimate the accuracy of the ground truth itself, we quantified the agreement among

different annotators of the same tracks. Any difference in the annotations of two or more annotators indicates that at least one of them made a mistake; this, in turn, leads to a harsher evaluation of the models than needed.

The remainder of this article is organized as follows: Previous works on the evaluation of DTM algorithms is presented in Sec. 2; in Sec. 3 the transcription accuracy of each class is evaluated, and in Sec. 4 different sources of errors are quantified; finally, the accuracy of the annotations is estimated in Sec. 5, conclusions are drawn in Sec. 6.

2. RELATED WORKS

Automatic drum transcription has evolved from a single, complex task into a series of intermediary steps of increasing difficulty. This evolution facilitated the development of new and more efficient algorithms [3]. Previously, the transcription was limited to simplified audio tracks or constrained vocabulary sizes. However, recent progress made through approaches based on supervised deep learning (DL) has been so significant that it becomes realistic to tackle such a complex task as the non-simplified DTM. The development of DL algorithms focused on two aspects: First, better and more complex architectures are exploited to improve the capabilities of the models, most recently with the introduction of the self-attention mechanism by Vaswani et al. [4] which has been adapted for DTM (e.g., [1, 5]). Second, better training procedures are employed to tune the models, e.g., with the creation of new datasets [1, 2, 6–9].

The de facto method to measure the accuracy of these DTM algorithms, as suggested by the Music Information Retrieval Evaluation eXchange (Mirex) [10], is the F-measure (also known as F1 score). This metric is computed with the harmonic mean between precision and recall of the drum onsets: An onset is considered correct when its estimation is within a small distance from the ground truth. A distance between 20 ms to 50 ms is what is generally used, but it can be tuned depending on the precision of the ground truth [1, 11]. Moreover, the F-measure can be computed at different levels of granularity: from a single class and track to the overall result for a whole dataset. To average multiple tracks and classes, the F-measure can be either computed as the mean value (mean F-measure) or by joining tracks and classes as if they were part of the same file and instrument (sum F-measure). In this study, we rely on the latter because it is more robust to rare edge cases (e.g., a track or class with very few onsets) [12, p.23]. However, since the F-measure gives the same importance to all onsets regardless of their position (i.e., strong or weak beats) or dynamics (loudness), this metric does not necessarily capture the opinion of human listeners [6].

Besides the F-measure, other tools are also used to assess a transcription. For example, Callender et al. used listening tests "where raters compared synthesized transcriptions to original recordings" to estimate the perceived quality of the transcriptions [6]. Vogl et al. relied on confusion matrices adapted to multi-label classification to identify the errors performed by their model [2]. Ishizuka et al. proposed a "tatum-level error rate based on the Levenshtein distance" [5].

Besides questions related to metric issues, the results of an evaluation are also heavily impacted by the datasets used for testing. There are a handful of datasets suitable for DTM which we group into the following three categories. A thorough description of the datasets is provided by Zehren et al. [13].

- Small but accurate datasets, which have been mostly annotated by hand by their creators, such as RBMA [14], ENST [15], or MDB [16].
- Large but synthetic datasets, synthesized from MIDI files to generate the input audio, such as TMIDT [2].
- Large but inaccurate datasets, which have been annotated by a crowd of people and refined algorithmically, such as ADTOF-RGW [17], ADTOF-YT [1], or A2MD [9].

To choose which datasets to use for testing, we singled out two criteria: First, the characteristics of the datasets (their data distribution) constitute the distribution in which the model is evaluated and should ideally be representative of a real-world situation. For example, testing can be done on different musical genres [1,2], real-world or synthesized audio [5], or different mixtures of instruments (e.g., audio containing four or five sound sources) [18,19]. Second, the test dataset may be part of the training dataset or be a new one, never used during training. The latter is known as an out-of-domain evaluation and, although more challenging, gives a better approximation of the true performances of the model (i.e., its generalization capabilities) [1, 20].

3. CLASS-SPECIFIC RESULTS

In this section, we compare the F-measures for different classes (set of instruments), to identify the most difficult instruments to transcribe for a model when trained in different ways. For this purpose, we selected the "Frame self-att" deep-learning architecture, as it has been recently employed for drum transcription [1], and compared three existing training procedures: 1) training on TMIDT with refinement on ENST, MDB, and RBMA [2]; 2) training on ADTOF-RGW [17]; and 3) training on ADTOF-RGW and ADTOF-YT [1]. [1]

We evaluated the resulting models on a set of five datasets (RBMA, ENST, MDB, ADTOF-RGW, and ADTOF-YT), to be representative of a real-world situation and to include both on-domain and out-of-domain evaluations. These datasets were carefully mapped to a common vocabulary containing five classes: bass drum (BD), snare drum (SD), toms (TT), open and closed hi-hat (HH), and other cymbals (CY). For the sake of brevity, in Fig. 1 we only present the results of the tests on ENST and ADTOF-YT, as these are representative of the tests on all five datasets.

[1] The models are available at `github.com/MZehren/ADTOF`

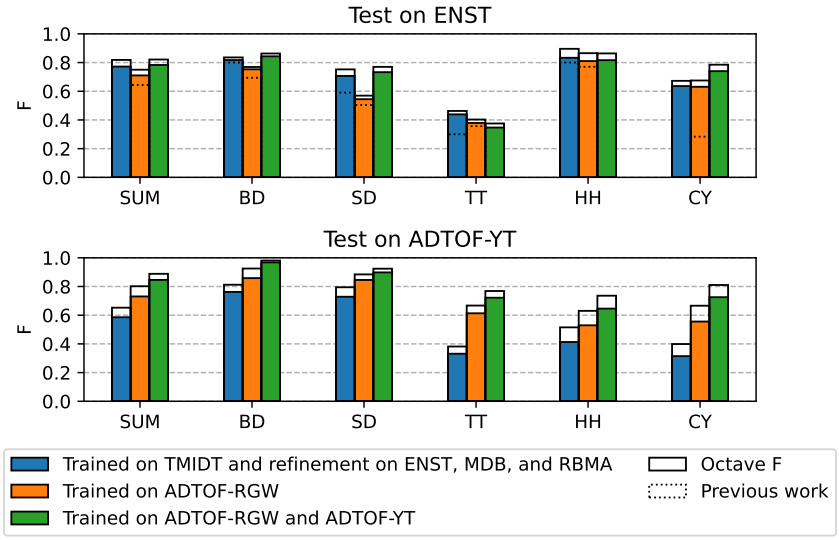

Figure 1: F-measure for the individual classes when testing on ENST (top) and ADTOF-YT (bottom).

First, we analyzed the results on ENST (top) to assess if and to what extent our results are close to those of the original authors of the three existing training procedures. Notably, our reproduction of pre-training on TMIDT with refinement on the non-crowdsourced datasets by Vogl et al. (blue bars) is slightly ahead when compared to their original work (whose results are indicated by the dotted lines inside the blue bars) [2, p.6]. Due to the fact that the evaluation was performed on a vocabulary larger than what we used in our test, a comparison for all instruments was not possible. The little improvement of our model in transcribing the instruments that can be directly compared (BD, SD, TT, and HH) is an indication that we were successful in reproducing the original algorithm. The reproduction of training on only ADTOF-RGW (orange bars) is also slightly better than the results reported [17, p.823] (indicated by the dotted lines inside the orange bars). We attribute this improvement to adopting a more random sampling procedure, something that helped train the models; indeed, compared to the previous work where consecutive (back to back) sequences were drawn between multiple occurrences of the same track, we sampled randomly the datasets (i.e., random track and position, without replacement), thus creating a more homogeneous training. Finally, although the reproduction of training on ADTOF-RGW and ADTOF-YT (green bars) cannot be compared on the class-specific results since they were not previously reported, our model achieved virtually the same sum F-Measure. Namely, when testing on ENST, the model trained on the two ADTOF datasets matches the performance of the model trained on ENST. Thus, ADTOF-based training, as it allows generalization towards ENST, does not overfit models.

Second, we analyzed the results achieved on ADTOF-YT (bottom) to highlight the potential of this dataset. The model achieved a very high F-measure on ADTOF-YT when training on both ADTOF datasets [1] (green bars), and almost a perfect score for BD, which is surprising considering that this dataset includes the fastest tempi and the densest sequences of onsets, which intuitively are features that hinder transcription. However, we noted that such a high performance is achieved only when ADTOF-YT is part of the training data, which means that the other datasets generalize poorly to it. The fact that only ADTOF-YT attains such a high accuracy both on-domain and out-of-domain may be explained by its large size and the homogeneity of its acoustic and drum patterns (due to the bias toward the metal music genre).

Third, although training and testing on ADTOF-YT yields the highest on-domain performance, we observed that the model has an atypical distribution of performance. In contrast to the usual result where the less represented instruments are less reliably transcribed (e.g., TT when training and testing on ENST), which is due to a lack of training examples, here, the model performs worse on frequently playing instruments. In fact, most of the mistakes of the model concern the transcription of cymbals (HH and CY).

4. ERRORS IN THE ESTIMATIONS

To identify why the transcription of cymbals is prone to mistakes, we quantified the errors made by the model when training and testing on ADTOF-YT. Note that this part of the study is not interested in the generalization capabilities of the model, but in assessing how well it can learn the target data distribution when training on it. We analyze the errors of the model with two tools: First, we approximated the number of errors due to quiet notes, also known as ghost notes, in the dataset with the octave F-measure. Second, we quantified the confusion between the instruments with a pseudo confusion matrix.

4.1 Octave F-measure

We attribute the low performance for cymbals, after conducting a preliminary inspection of the estimations, to both a specific characteristic of their timbre —long sustain that may mask the next onset and the presence of many quiet notes. [2] Both features make cymbals very challenging to transcribe and their transcription suffers from false negatives and false positives.

As a first step toward solving this problem, we created a new metric meant to quantify how often the presence of quiet notes leads to transcription mistakes. Unfortunately, because ADTOF-YT does not contain reliable velocity information in the annotations, we could not estimate the presence of quiet notes through velocity. Therefore, we started with an assumption from the expert knowledge according to which, in many music genres, cymbals are commonly played in alternation between loud (accentuated) and quiet notes. [3] This insight is in agreement with our observation that errors in the estimations are often rhythms that are half or double the speed of the ground truth, so that the algorithm would transcribe a sequence of quarter notes where it should be an eighth note or vice versa. Assuming that this mismatch is due to quiet notes, we created the octave F-measure to allow rhythms that are exactly half or double the speed of the annotations (white bars in Fig. 1). A parallel can be drawn with tempo estimation that uses the "accuracy2" metric which is defined to accept estimations that have a double or triple relationship with ground truth, disregarding ipso facto the so-called octave tempo errors [21]. The octave F-measure gives us an upper bound of the performance of the models if these mistakes were not present in the estimation and the ground truth, and helps us quantify the issues yet to be solved in the algorithms.

When looking at the octave F-measure on ADTOF-YT, we confirm the presence of undetected annotations exactly at the middle point between two estimations, and the presence of extra estimations exactly at the middle point between two annotations. This phenomenon is more common in cymbals than in any other instrument of the drum kit and it is observed in most of the datasets. In other words, the models are mistaking the beat subdivision at which cymbals occur, a problem we attribute to their specific timbre and alternation between loud and quiet notes.

4.2 Confusion Between Classes

To identify typical errors made by the model on ADTOF-YT, we employed the pseudo confusion matrix represented in Fig. 2. Compared to a standard confusion matrix, ours differs in two aspects: First, since in AMT any time position may contain multiple labels—different instruments play simultaneously—the possible sets of labels, instead of each single class, are uniquely listed in the rows and

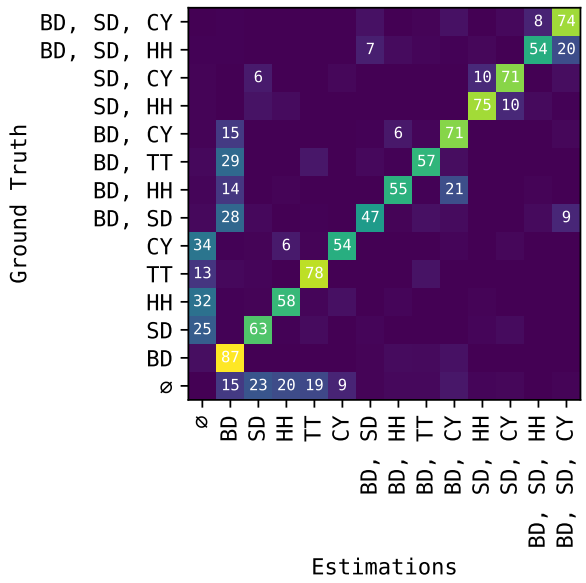

Figure 2: Pseudo confusion matrix on ADTOF-YT. The numbers represent a percentage of the ground truth.

columns of the matrix. [4] As uniquely identifying sets of labels leads to a large matrix (2^5 columns and rows with five classes), we truncated the figure for readability to show only the most frequent columns and rows. Second, to remove the imbalance between classes, we normalized the rows (i.e., the rows sum up to 100%). Thus, rather than displaying the count of each set of labels, we represented their proportion relative to the number of occurrences of the ground truth.

We categorize the errors (i.e., the cells outside of the diagonal) in three types following Vogl's approach [2]: i) confusion, when the onset is detected but the label is wrong (false positives with false negatives); ii) masking, when an onset is missing, presumably because of another one, correctly detected, hides it (false negatives with true positives); iii) excitement where an extra onset is detected, presumably because another one, correctly detected, generates excitement (false positives with true positives). Additionally, another cause of mistakes, which we do not consider in this study, might be related to the low number of occurrences of some combinations of labels (e.g., BD and SD played at the same time), which makes them more difficult to estimate for the model. In Fig. 2, we identified three trends.

First, the left-most column shows that CY, HH, and SD are missing $\approx 30\%$ of the time when they play alone; at the same time, the bottom row highlights that they are incorrectly estimated $10 - 20\%$ of the time when there should be no instrument playing. Surprisingly, this common issue cannot be categorized as due to confusion, masking, or

[2] SD also contains many quiet notes, but not to the same extent as HH and CY in this dataset.

[3] An illustration of this phenomenon can be viewed in the ENST dataset: Notice how every second HH onsets sounds quieter in the example video "Drummer 3, Angle 1" https://perso.telecom-paristech.fr/grichard/ENST-drums/

[4] In practice, since onsets that slightly deviate from the correct position are considered simultaneous, the confusion matrix was created by using an agglomerative clustering that group onsets with a tolerance of 50 ms. As a side effect, we do not count the true negatives (i.e., positions without onsets: $\emptyset$ in the ground truth and estimation).

excitement. Instead, we attribute these two phenomena to the presence of quiet notes for those three instruments, as already commented in Section 4.1.

Second, when looking at the intersection of the rows containing HH with the equivalent columns containing CY, such as the cell in the row "BD, HH" and column "BD, CY", we notice that HH is often confused with CY. Similarly for the intersection of the rows containing CY with the equivalent columns containing HH, we observe that CY is often confused with HH. Again, this highlights that similar-sounding instruments are misinterpreted, as already identified by Vogl [2]. However, looking specifically at the rows "CY" and "HH", we notice that CY and HH are less often confused with each other when they occur in isolation. Thus, we conclude that confusion is exacerbated by the presence of other instruments, likely because of their masking effects.

Lastly, the second column highlights that TT and SD are both missed $\approx 30\%$ of the time when they appear with BD. Presumably because BD's wide spectral range masks the other instruments' spectra, this illustrates that masking seems to be very prevalent on ADTOF-YT. Excitation, on the other hand, is not a common issue compared to masking or confusion. We only notice the presence of an extra CY onset 9% of the times that "BD, SD" occurs.

In addition to the Octave F-measure that illustrates how the model misjudges the beat subdivision at which cymbals are played, the confusion matrix shows that the model does not differentiate well the cymbals. However, one might wonder why these issues are prevalent on ADTOF-YT and not the other datasets.

5. ANNOTATIONS ACCURACY

To understand why the model is prone to make mistakes specifically with ADTOF-YT, we took a closer look at the accuracy of its annotations. Both the datasets ADTOF-RGW and ADTOF-YT are crowdsourced; since human annotations are not perfect and annotators do not always agree with each other, we expect mistakes in the datasets. While a cleansing/cleaning procedure was employed to improve the time position of the annotations and to remove label ambiguity [1, p.784], it is not realistic to expect that all mistakes will be corrected. Although DL is generally robust to label noise [22], incorrect labeling might affect the models during training and testing, especially with crowdsourced datasets that likely contain more mistakes than non-crowdsourced ones. Specifically during testing, any error in the annotations is indistinguishable from wrong estimations of the models and impacts their evaluation. Therefore, by assessing the accuracy of the annotations, it is possible to estimate an upper bound of the performance of the models tested on the dataset. As this bound corresponds to the score achieved by a perfect classifier, it can show how far the current models are from this ideal.

To estimate the annotations' accuracy on a dataset and create a ground truth of high confidence, it is common to compare labels provided by independent annotators on the same data and measure the confidence of the annotations,

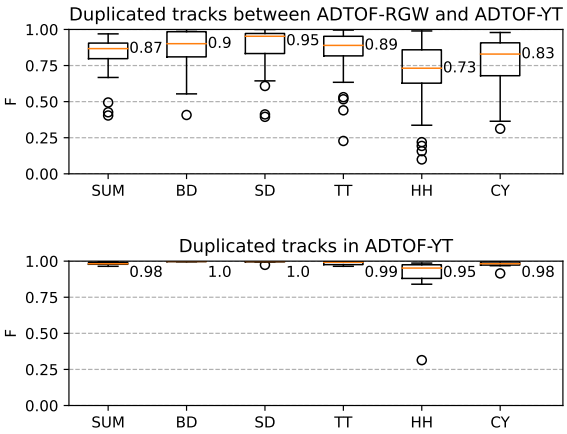

Figure 3: Box plots representing the distribution of the F-measure on tracks present both in ADTOF-RGW and ADTOF-YT (top) and on duplicated tracks in ADTOF-YT (bottom).

for example by grouping multiple independent annotations into a single set (e.g., [23, p.7], [24], [25, p.255]). Further, by comparing this ground truth with a new (group of) annotator(s), one can estimate either the ground truth accuracy or the human-level accuracy depending on how much one trusts the reference group (e.g., [23, p.7 and 31], [26]). In our context, similarly to what Flexer and Grill [27] did in their work, we aim to estimate an upper limit of the score achievable on the datasets by assessing the agreement among human annotators.

To do so, we rely on the tracks that appear multiple times and are annotated by different persons in the datasets. After aligning two instances of the same track according to their annotations, we were able to compute the agreement between the annotators the same way we evaluate any algorithmic estimation: By taking either set of annotations as the reference and the other one as the estimation, we can then compute the F-measure. Note that the results do not depend on which annotator is used as the reference: By switching the annotator used as the reference, precision and recall are also switched, without impacting the F-measure. The distribution of the F-measure for all the tracks found both in ADTOF-RGW and ADTOF-YT (34 couples, 4h28min) and duplicated in ADTOF-YT (7 couples, 34min) is shown in Fig. 3. There are no duplicated tracks within ADTOF-RGW.

On the one hand, the annotations in duplicated tracks of ADTOF-YT are almost identical, whereas they differ between ADTOF-RGW and ADTOF-YT. This is an indication that the annotators of ADTOF-YT agree more often with themselves than with the annotators of ADTOF-RGW. In turn, this is a sign that the annotations of ADTOF-YT are very accurate. If that is the case, then the models we evaluated on ADTOF-YT are far from perfect, as they do not achieve results close to the inter-rater agreement. However, we acknowledge that this trend is only supported by seven couples of tracks and further investigation is required to claim that this dataset contains so few errors.

On the other hand, the median agreement between annotators of ADTOF-RGW and ADTOF-YT is very similar to the best model's sum F measure on ADOTF-YT (0.87 for the annotators Fig. 3 compared to 0.85 for the model Fig. 1). This suggests that the model performs as well as the annotators of ADTOF-RGW on ADTOF-YT. Moreover, this trend holds for the majority of the classes. Most notably, both the annotators and the model manifest difficulties with the discrimination between HH and CY (lowest agreement and performance). Similarly to the models in the previous sections, we attribute these human errors both to the fact that one instrument is mistaken for the other because of their similar timbre, and to the use of different rhythms because of the presence of quiet notes. See Fig. 4 (top) showing the disagreement between two annotators as an illustration of both phenomena. Although it is not clear if these discrepancies are part of ADTOF-RGW, ADTOF-YT, or both, they impact negatively the measure of the model performance. However, we noticed that the agreement between annotators is much lower than the performance of the model on BD (0.90 for the annotators compared to 0.97 for the model). This is due to the presence of simplified annotations in ADTOF-RGW. As represented in Fig. 4 (bottom), these simplifications are meant to ease the gameplay when a double bass drum technique is required (i.e., bass drum notes played with both feet) by omitting the notes played by the left foot. Despite such simplifications, the model still manages to achieve a high F-measure when testing on ADTOF-YT, which does not contain simplified annotations.

Although data is not enough to determine accurately an upper limit to the performance of the models, we believe that the agreement we measured among annotators is a reasonably good guess. Because it is not possible to know which of the annotators made a mistake (possibly both), the discrepancies between them do not always impact the measure of the model's performance, making this estimation pessimistic. [5] However, considering that the best performance we achieve on ADTOF-YT is close to the agreement among humans, it is intuitive that any improvement of the model beyond this point will not be easily measurable. In other words, this model is not far from a perfect classifier on this dataset.

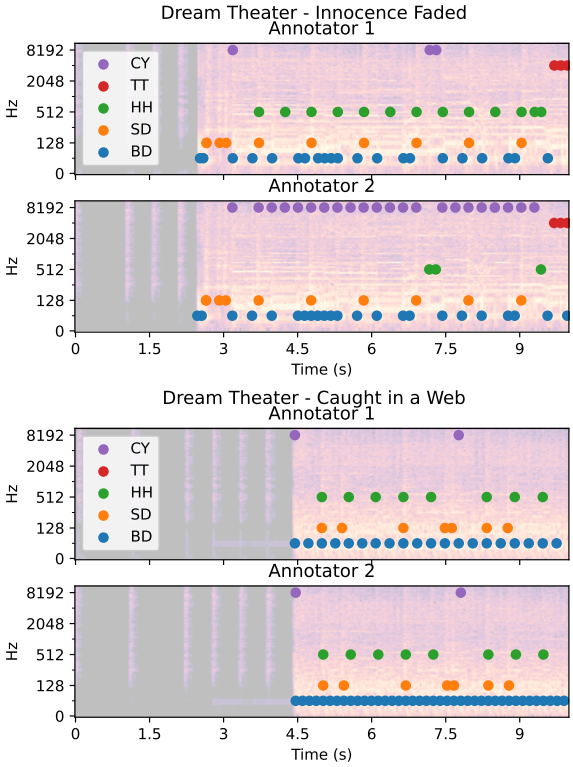

Figure 4: Spectrograms and annotations from ADTOF-RGW and ADTOF-YT for the first 10s of two tracks. Notice the confusion and the use of different subdivisions between CY and HH (top), as well as the simplification of the BD for fast rhythms (bottom).

6. CONCLUSIONS

In this work, we analyzed the performance of a state-of-the-art model for automatic drum transcription [1]. First, through the F-measure for the individual classes, we identified that ADTOF-YT is the only dataset able to train a model to such a high level of accuracy on its data distribution. In this context, when training and testing on ADTOF-YT, the transcription is: almost perfect for the bass drum (BD), better than previous methods for the sparse class of tom-toms (TT), but less reliable for cymbals. Second, to understand why cymbals are more difficult to transcribe, we used a new metric we named Octave F-measure as well as a pseudo confusion matrix. We then concluded that what hinders the cymbals' transcription is their typical accentuated-note/quiet-note pattern and their similar timbre to each other. Last, because the test data has been annotated by many people with different levels of expertise, we aimed to quantify the errors due to discrepancies in the ground truth rather than to mistakes made by the model. By estimating the accuracy of the annotations through the agreement between multiple annotators of the same tracks, we identified that the human-level accuracy is on par with the performance of the model. Thus, it is not clear whether the differences between the estimations and annotations originate from the model or the annotators, even though their causes are the same.

With this study, we quantified the main difficulties faced by the model or the annotators. The errors caused by the cymbals could be the focus of future research in ADT, which we believe could be tackled in one of two ways: Either existing annotations could be verified, possibly via a semi-automatic method relying on the estimation of a pre-trained model to detect likely errors, or complementary training data could be generated, possibly in a synthetic way, to ensure a perfect ground truth.

[5] In the hypothetical scenario where each disagreement is caused by only one of the annotator making a mistake, the discrepancies will affect the evaluation only 50% of the cases.

7. ACKNOWLEDGMENTS

The computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

8. REFERENCES

[1] M. Zehren, M. Alunno, and P. Bientinesi, "High-Quality and Reproducible Automatic Drum Transcription from Crowdsourced Data," *Signals*, vol. 4, no. 4, pp. 768–787, Nov. 2023. [Online]. Available: https://www.mdpi.com/2624-6120/4/4/42

[2] R. Vogl, G. Widmer, and P. Knees, "Towards multi-instrument drum transcription," in *21th International Conference on Digital Audio Effects (DAFx-18)*, Jun. 2018. [Online]. Available: http://arxiv.org/abs/1806.06676

[3] C.-W. Wu, C. Dittmar, C. Southall, R. Vogl, G. Widmer, J. Hockman, M. Muller, and A. Lerch, "A Review of Automatic Drum Transcription," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1457–1483, Sep. 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8350302/

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," *arXiv:1706.03762 [cs]*, Dec. 2017, arXiv: 1706.03762. [Online]. Available: http://arxiv.org/abs/1706.03762

[5] R. Ishizuka, R. Nishikimi, and K. Yoshii, "Global Structure-Aware Drum Transcription Based on Self-Attention Mechanisms," *Signals*, vol. 2, no. 3, pp. 508–526, 2021. [Online]. Available: https://www.mdpi.com/2624-6120/2/3/31

[6] L. Callender, C. Hawthorne, and J. Engel, "Improving Perceptual Quality of Drum Transcription with the Expanded Groove MIDI Dataset," *arXiv:2004.00188 [cs]*, May 2020, arXiv: 2004.00188. [Online]. Available: http://arxiv.org/abs/2004.00188

[7] M. Cartwright and J. P. Bello, "Increasing Drum Transcription Vocabulary Using Data Synthesis," in *21th International Conference on Digital Audio Effects (DAFx-18)*, 2018, pp. 72–79.

[8] E. Manilow, G. Wichern, P. Seetharaman, and J. Le Roux, "Cutting music source separation some Slakh: A dataset to study the impact of training data quality and quantity," in *IEEE workshop on applications of signal processing to audio and acoustics (WASPAA)*. IEEE, 2019.

[9] I.-C. Wei, C.-W. Wu, and L. Su, "Improving Automatic Drum Transcription Using Large-Scale Audio-to-Midi Aligned Data," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Toronto, ON, Canada: IEEE, Jun. 2021, pp. 246–250. [Online]. Available: https://ieeexplore.ieee.org/document/9414409/

[10] "MIREX 2021: Drum Transcription," 2021. [Online]. Available: https://www.music-ir.org/mirex/wiki/2021:Drum_Transcription#Evaluation

[11] J. Gardner, I. Simon, E. Manilow, C. Hawthorne, and J. Engel, "MT3: Multi-Task Multitrack Music Transcription," in *International Conference on Learning Representations*, 2021, p. 21, arXiv: 2111.03017. [Online]. Available: https://openreview.net/pdf?id=iMSjopcOn0p

[12] R. Vogl, "Deep Learning Methods for Drum Transcription and Drum Pattern Generation," Ph.D. dissertation, Institute of Computational Perception, Linz, Austria, Nov. 2018.

[13] M. Zehren, M. Alunno, and P. Bientinesi, "Analyzing and reducing the synthetic-to-real transfer gap in Music Information Retrieval: the task of automatic drum transcription," Jul. 2024, arXiv:2407.19823 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2407.19823

[14] R. Vogl, M. Dorfer, G. Widmer, and P. Knees, "Drum transcription via joint beat and drum modeling using convolutional recurrent neural networks," in *18th International Society for Music Information Retrieval Conference*, Suzhou, China, 2017, pp. 150–157.

[15] O. Gillet and G. Richard, "ENST-Drums: an extensive audio-visual database for drum signals processing," in *7th International Society for Music Information Retrieval Conference (ISMIR)*, Victoria, Canada, 2006, pp. 156–159.

[16] C. Southall, C.-W. Wu, A. Lerch, and J. Hockman, "MDB drums- An annotated subset of MedleyDB for Automatic Drum Transcription," Suzhou, China, 2017.

[17] M. Zehren, M. Alunno, and P. Bientinesi, "ADTOF: A large dataset of non-synthetic music for automatic drum transcription," in *22nd International Society for Music Information Retrieval Conference (ISMIR)*, Online, 2021, pp. 818–824.

[18] E. Manilow, P. Seetharaman, and B. Pardo, "Simultaneous Separation and Transcription of Mixtures with Multiple Polyphonic and Percussive Instruments," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Barcelona, Spain: IEEE, May 2020, pp. 771–775. [Online]. Available: https://ieeexplore.ieee.org/document/9054340/

[19] Y. Wang, J. Salamon, M. Cartwright, N. J. Bryan, and J. P. Bello, "Few-Shot Drum Transcription in Polyphonic Music," in *21st International Society for Music Information Retrieval Conference (ISMIR)*, Montréal, 2020, pp. 117–124.

[20] D. Hernandez, J. Kaplan, T. Henighan, and S. McCandlish, "Scaling Laws for Transfer," Feb. 2021, arXiv:2102.01293 [cs]. [Online]. Available: http://arxiv.org/abs/2102.01293

[21] H. Schreiber, J. Urbano, and M. Müller, "Music Tempo Estimation: Are We Done Yet?" *Transactions of the International Society for Music Information Retrieval*, vol. 3, no. 1, p. 111, Aug. 2020. [Online]. Available: https://transactions.ismir.net/article/10.5334/tismir.43/

[22] D. Rolnick, A. Veit, S. Belongie, and N. Shavit, "Deep Learning is Robust to Massive Label Noise," Feb. 2018, arXiv:1705.10694 [cs]. [Online]. Available: http://arxiv.org/abs/1705.10694

[23] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," Jan. 2015, arXiv:1409.0575 [cs]. [Online]. Available: http://arxiv.org/abs/1409.0575

[24] M. Zehren, M. Alunno, and P. Bientinesi, "M-DJCUE: A Manually Annotated Dataset of Cue Points," in *Late Breaking/Demo at the 20th International Society for Music Information Retrieval*, Delft, The Netherlands, 2019, p. 2.

[25] O. Nieto, G. J. Mysore, C.-i. Wang, J. B. L. Smith, J. Schlüter, T. Grill, and B. McFee, "Audio-Based Music Structure Analysis: Current Trends, Open Challenges, and Applications," *Transactions of the International Society for Music Information Retrieval*, vol. 3, no. 1, pp. 246–263, Dec. 2020. [Online]. Available: http://transactions.ismir.net/articles/10.5334/tismir.54/

[26] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, and A. Stolcke, "The Microsoft 2017 Conversational Speech Recognition System," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 5934–5938, arXiv:1708.06073 [cs]. [Online]. Available: http://arxiv.org/abs/1708.06073

[27] A. Flexer and T. Grill, "The Problem of Limited Inter-rater Agreement in Modelling Music Similarity," *Journal of New Music Research*, vol. 45, no. 3, pp. 239–251, Jul. 2016. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/09298215.2016.1200631