UNDERSTANDING PERFORMANCE LIMITATIONS IN AUTOMATIC DRUM TRANSCRIPTION

Philipp Weyers[1] **Christian Uhle**[1,2] **Meinard Müller**[1,2]
Matthias Lang[1]
[1] Fraunhofer Institute for Integrated Circuits (IIS), Erlangen, Germany
[2] International Audio Laboratories Erlangen, Germany
philipp.weyers@iis.fraunhofer.de

ABSTRACT

Recent advancements in Automatic Drum Transcription (ADT) have improved overall transcription performance. However, state-of-the-art (SOTA) models still struggle with certain drum classes, particularly toms and cymbals, and the specific factors limiting their performance remain unclear. This paper addresses this gap by leveraging the Separate-Tracks-Annotate-Resynthesize Drums (STAR Drums) dataset to create multiple dataset versions that systematically eliminate potential performance constraints. We conduct experiments using three common ADT deep neural network (DNN) architectures to identify and quantify these limitations. For drum transcription in the presence of melodic instruments (DTM), the primary limiting factor is interference from melodic instruments and singing. Aside from this, performance improves by approximately five percent when training and testing use the same single drum kit, only strong onsets are present, or notes are not played simultaneously. For drum transcription of drum-only recordings (DTD), nearly error-free transcription is achieved when simultaneous onsets are removed. This confirms that overlapping drum hits are the main performance constraint. By identifying key ADT challenges, we provide insights to enhance SOTA models and improve overall transcription accuracy.

1. INTRODUCTION

As a sub-field of Automatic Music Transcription (AMT) within the broader field of Music Information Retrieval (MIR), Automatic Drum Transcription (ADT) focuses on identifying and classifying drum sounds in audio signals. Drum transcription of drum-only recordings (DTD) is considered less challenging due to the absence of sounds originating from other instruments, whereas drum transcription in the presence of melodic instruments (DTM) presents the challenge of drum sounds potentially being masked by melodic instruments and singing, or non-drum sounds being misclassified as drums [1].

Applications of ADT include music education software that provides real-time feedback for students practicing on acoustic drum kits, and music production tools that use transcriptions to add or replace drum samples [1].

For these applications, high-quality drum transcription is essential. However, achieving this requires overcoming several key challenges:

- Interference from melodic instruments and vocals, which can mask drum sounds [1].
- Overlapping drum sounds from different classes, leading to mutual masking [2].
- Weak onsets, which are difficult to detect due to low loudness and energy [3].
- Limited generalization, which affects transcription performance across diverse datasets.

We systematically investigate and quantify the impact of these challenges by creating multiple versions of the Separate-Tracks-Annotate-Resynthesize Drums (STAR Drums) dataset [4] that simplify the ADT problem.

Our main contribution is to provide insights into performance improvements achievable by systematically eliminating limiting factors in training and test data. Experiments are conducted for DTM and DTD separately, supporting the development of more robust ADT algorithms and a deeper understanding of the problem.

The paper is structured as follows: Section 2 reviews related work, Section 3 describes the methodology, Section 4 details the experiments, Section 5 presents the results, and Section 6 concludes.

2. RELATED WORK

The emergence of deep neural networks (DNNs) in ADT improved transcription performance. In [5], various Convolutional Neural Network (CNN) and Convolutional Recurrent Neural Network (CRNN) architectures were compared, and later, these were trained on large amounts of synthetic data generated from MIDI files, combined with smaller manually labeled datasets [2].

Subsequent works [4, 6–9] have cited [2] or [5] as state of the art (SOTA), and explored improvements using different datasets or alternative DNN architectures.

Few shot learning (FSL) has been applied to ADT with promising results, though it required examples for each class at inference time [6]. Dynamic FSL addressed this by eliminating the need to provide examples for drum classes present in initial training, while still allowing adaptation to drum sounds at inference time [4].

The authors of [7] and [10] created the Automatic Drums Transcription On Fire (ADTOF) dataset using crowd-sourced annotations and trained models with CRNN and CNN architectures incorporating self-attention. Both frame-wise and tatum grid synchronized models achieved similar performance. Tatum-level attention-based networks were also effective in [11].

In [8], the A2MD dataset was created using semi-automatic labeling. The proposed models also evaluated beat information, resulting in a modest performance increase. The authors of [9] employed a language model to regularize training for suppressing musically unnatural onsets. While this approach improved performance, the transcription was limited to the three main drum instruments, bass drum, snare drum, and hi-hat.

While state-of-the-art (SOTA) algorithms for DTM achieve good overall performance (global F-measures above 0.8 on hand-annotated datasets [3, 10]), classes such as toms and cymbals still show mediocre results.

In [3], the performance of models trained on the ADTOF dataset is analyzed in detail. Several hypotheses are proposed to explain transcription errors. The issue of soft onsets being masked is partly investigated by introducing a tempo octave F-measure, which disregards errors when transcriptions occur at half or double the tempo. The underlying assumption is that especially cymbals are often played with alternating weak and strong onsets, resulting in every second weak onset being missed. Additionally, confusion matrices are used to analyze class confusions, revealing similar problems identified in [2]: Misclassification frequently occurs among similar sounding instruments, such as hi-hat and cymbals. Other common errors are linked to weak onsets, characterized by low loudness, or masking effects.

In this paper, we go beyond describing current SOTA performance by attributing the remaining performance gap to to a perfect transcription to specific performance-limiting factors. This allows us to target these factors in future work and estimate the potential maximum performance gains.

3. METHODOLOGY

We take advantage of the STAR Drums dataset, first utilized in [4], to create training and test data where drum stems are modified to eliminate potential error sources, thereby progressively reducing the complexity of the ADT task. STAR Drums is created from audio recordings including melodic instruments, singing, and drums. We either utilize audio data provided as separate drum and non-drum stems or apply a Music Source Separation (MSS) algorithm to separate mixture recordings into drum and non-drum stems.

Subsequently, we annotate the drum stem using an ADT algorithm published alongside [2] and regard this information as estimated reference annotation. We then re-synthesize the drums by rendering the estimated reference annotations using several virtual drum kits and normalize the loudness of the re-synthesized drum stem according to Recommendation ITU-R BS1770-4 (2015) to match the loudness of the original drum stem. Finally, the re-synthesized drum stem is mixed with the original non-drum stem to create the audio signal for training an testing ADT algorithms.

The input data for STAR Drums originates from MUSDB18 [12], ISMIR04 [13], and MTG-Jamendo [14]. With data from ISMIR04 (originally for genre classification), STAR Drums covers a wide range of genres, while Rock and Pop are emphasized due to MUSDB18 and MTG-Jamendo. The data from MUSDB18 is used for validation and testing because it is already available as stems, thus avoiding biases in the results caused by artifacts introduced by MSS. By using $60\,\mathrm{s}$ excerpts from ISMIR04 and MTG-Jamendo data and full items from MUSDB18, we achieve reasonable ratios between the lengths of training, validation, and test splits, corresponding to $114.7\,\mathrm{h}$, $8.3\,\mathrm{h}$, and $1.6\,\mathrm{h}$, respectively.

STAR Drums contains recordings of instruments played by musicians and vocals, unlike fully synthetic datasets. In [4], STAR Drums outperformed training with Slakh [15], which uses only synthetic data. STAR Drums also allows full control over the re-synthesized drum stem, unlike other datasets [7, 8, 10] where separate drum stems are unavailable and drum sounds cannot be modified. Additionally, the results are not affected by labeling errors, as only the re-synthesized drum stem, which matches the estimated annotation exactly, is included. In contrast, the extent of labeling errors in ADTOF is unknown, and even human annotators often not fully agree as demonstrated in [3].

Leveraging STAR Drums's flexibility, we create simplified versions of the re-synthesized drum stem to systematically reduce ADT task complexity, allowing precise quantification of key limiting factors.

Table 1 presents the five variants of STAR Drums and the specific research questions they are designed to address. The `20Kits` version, serving as the baseline, utilizes 20 virtual drum kits to generate the re-synthesized drum stem. It includes simultaneous onsets and captures a full dynamic range, with MIDI velocity values from 40 to 127. The performance of models trained on different STAR Drums variants will be evaluated relative to this baseline.

For `10Kits`, we divide the 20 virtual drum kits into two distinct splits and train a model on each. To evaluate the impact of identical drum sounds in training and testing, we first test each model on the split it was trained on. Additionally, we perform 2-way cross-validation by testing each model on the split it was not trained on, which allows us to assess performance when the drum sounds in training and testing differ.

The `1Kit` version is created using a single drum kit. We

STAR Drums Variant	Identifier	Research Question
20 drum kits	`20Kits`	Baseline
10 drum kits (two splits)	`10Kits`	How does training and testing with a reduced number of different drum sounds impact performance? How does testing with drum sounds not included in training impact performance?
1 drum kit (four splits)	`1Kit`	How does training and testing with a reduced number of different drum sounds impact performance? How does testing with drum sounds not included in training impact performance?
20 drum kits - No weak onsets	`20KitsNoWeak`	How does the absence of weak onsets in training and testing impact performance?
20 drum kits - No simultaneous onsets	`20KitsNoSim`	How does the absence of simultaneous onsets in training and testing impact performance?

Table 1. Variants of STAR Drums with identifier and corresponding research questions.

repeat the dataset creation, training, and testing with four different drum kits to enhance generalizability and perform the same evaluation as for `10Kits`, assessing performance with both identical and different drum sounds in training and testing. For clarity, we report only the averaged performance across all splits for `10Kits` and `1Kit`.

The goal of `10Kits` and `1Kit` is to examine how the presence of identical versus different drum sounds in training and testing impacts performance. Additionally, we assess how reducing the number of drum kits affects transcription accuracy in both scenarios.

`20KitsNoWeak` uses all 20 virtual drum kits and contains only strong onsets with high loudness by limiting the MIDI velocity range to 100 to 127 during dataset creation, providing insight into the influence of weak onsets, characterized by low loudness, on transcription performance.

Lastly, we create the `20KitsNoSim` version by using all 20 virtual drum kits and ensuring a minimum inter-onset interval of $50\,\mathrm{ms}$ during the re-synthesis process to investigate the effect of simultaneous onsets. In cases where multiple onsets occur within a $50\,\mathrm{ms}$ window, we randomly choose one onset and discard the others.

For each of the five variants, we generate both a drum-only version and a full-mix version to compare transcription performance in the more challenging DTM scenario against the less demanding DTD scenario.

4. EXPERIMENTS

We train models using each of the five variants of STAR Drums listed in Table 1 across three different architectures. An overview of the DNN architectures used is provided in Table 2. The models process monaural mel spectra with 96 bands and an upper cut-off frequency of $16\,\mathrm{kHz}$. The spectra are computed using a 1024-point short-time Fourier transform (STFT) with a hop length of 512 samples, derived from audio signals sampled at $48\,\mathrm{kHz}$, resulting in a frame length of $10.7\,\mathrm{ms}$.

While all models utilize CNN layers, the `CRNN` model additionally incorporates Recurrent Neural Network (RNN) layers, and the `CNNSA` model employs self-attention

Model	# Frames	# Parms.	Architecture
`CNN`	25	2.4M	5 CNN layers 3 Dense layers
`CRNN`	400	2.9M	4 CNN layers 3 RNN layers 3 Dense layers
`CNNSA`	400	6.9M	4 CNN layers 2 Self-att. layers 3 Dense layers

Table 2. Overview of models used, including input frame length, parameter count, and architectural details.

layers. Each model incorporates three dense layers to map the output to the number of classes, followed by a sigmoid activation function to generate probability estimates.

The `CNN` model operates on blocks of 25 STFT frames. Networks utilizing RNN or self-attention layers can capture temporal dependencies in the input data, allowing them to process larger blocks of 400 STFT frames for improved context modeling. Similar block lengths were proposed in [5, 10].

We categorize onsets into eight classes, following the mapping proposed in [2]: bass drum, snare drum, hi-hat, toms, bell, cymbals, ride cymbals, and clave. To consider a wide range of drum sounds, we avoid using the three-class mapping, only including bass drum, snare drum, and hi-hat, and the five-class mapping utilized in [3, 7, 10]. At the same time, we refrain from using the 18-class mapping, as used in [2], since it remains unclear to which extent classification ambiguities arising from the fine-grained mapping impact performance. As noted in [2], no clear frequency ranges exist for low, mid, and high toms. Additionally, distinguishing between sounds like closed hi-hat and pedal hi-hat can be challenging, even for humans.

Onset times are extracted from the detection probabilities using a peak picking algorithm with a fixed threshold of 0.55 across all classes. We identify true positives, false positives, and false negatives using a tolerance window of $50\,\mathrm{ms}$, following [3, 10], with the Python package

Test Dataset	Same drums train + test?	Model	Train Dataset									
			20Kits		10Kits		1Kit		20KitsNoWeak		20KitsNoSim	
			DTM	DTD	DTM	DTD	DTM	DTD	DTM	DTD	DTM	DTD
20Kits	✓	CNN	0.73	0.89	0.71	0.85	0.56	0.64				
		CRNN	0.78	0.91	0.76	0.87	0.58	0.65				
		CNNSA	0.78	0.91	0.75	0.86	0.59	0.64				
10Kits	✓	CNN			0.73	0.89						
		CRNN			0.79	0.91						
		CNNSA			0.78	0.91						
10Kits	✗	CNN			0.67	0.78						
		CRNN			0.73	0.81						
		CNNSA			0.71	0.81						
1Kit	✓	CNN					0.78	0.91				
		CRNN					0.82	0.92				
		CNNSA					0.80	0.92				
1Kit	✗	CNN					0.61	0.70				
		CRNN					0.64	0.71				
		CNNSA					0.64	0.71				
20KitsNoWeak	✓	CNN	0.74	0.89					0.77	0.90		
		CRNN	0.80	0.91					0.82	0.92		
		CNNSA	0.79	0.91					0.83	0.92		
20KitsNoSim	✓	CNN	0.71	0.92							0.77	0.97
		CRNN	0.78	0.94							0.83	0.98
		CNNSA	0.75	0.94							0.81	0.98
ENST Drums	✗	CNN	0.70	0.74	0.69	0.72	0.62	0.63	0.69	0.73	0.62	0.65
		CRNN	0.73	0.76	0.73	0.73	0.65	0.64	0.72	0.75	0.62	0.65
		CNNSA	0.73	0.76	0.71	0.73	0.64	0.66	0.71	0.73	0.62	0.63
MDB Drums	✗	CNN	0.71	0.79	0.71	0.76	0.64	0.65	0.71	0.79	0.66	0.74
		CRNN	0.75	0.74	0.73	0.74	0.63	0.64	0.72	0.75	0.66	0.73
		CNNSA	0.71	0.76	0.69	0.73	0.61	0.63	0.71	0.78	0.67	0.71

Table 3. Results in terms of global F-measure when training and testing on different STAR Drums versions using the CNN, the CRNN, and the CNNSA model for DTM and DTD. The last two rows show results for MDB Drums and ENST Drums.

mir_eval [16]. The global F-measure is used for performance comparison and is computed using micro averaging [17], which involves summing all true positives, false positives, and false negatives across all classes and tracks before calculating the F-measure. This approach assigns equal weight to every onset.

In addition to testing on the STAR Drums test split, we use MDB Drums [18] and ENST Drums [19], publicly available ADT datasets commonly used for testing. MDB Drums and ENST Drums include 0.4 and 1.0 h of hand-annotated audio, respectively, provided as complete mixtures containing recordings of drum sounds alongside melodic instruments. Additionally, drum-only stems are provided by the authors.

5. RESULTS

Table 3 presents the results of all experiments. For clarity, we evaluate only the combinations of training and test datasets that address the research questions in Table 1. Each cell in Table 3 shows two global F-measures for the three model architectures from Table 2, where the first row corresponds to the CNN model, the second row to the CRNN, and the third row to the CNNSA model. The first value is the global F-measure for DTM, and the second value is for DTD. The results for the single splits of 10Kits and 1Kit are averaged.

Overall, the CRNN and CNNSA models outperform the CNN model, aligning with the findings of [2]. The CNNSA model achieves similar performance to the CRNN model, as observed in [10]. The best performance on MDB Drums and ENST Drums with an F-measure of 0.75 and 0.73, respectively, is slightly lower than reported in [10], likely due to their use of a less challenging five-class mapping. In all experiments with STAR Drums, DTD performance surpasses DTM performance.

In the following subsections, we provide a detailed analysis of the results in relation to the research questions outlined in Table 1.

5.1 Reducing the Number of Drum Kits

Performance remains similar for DTM when the number of drum kits in training and testing is reduced from 20 to 10 (10Kits), provided the drum sounds are identical. In contrast, performance decreases when drum kits in training and testing differ. For instance, comparing 20Kits and 10Kits, the results for the CNN model decrease from 0.73 to 0.67 and for the CRNN model from 0.78 to 0.73.

Reducing from 20 (20Kits) to 1 drum kit (1Kit) that is identical in training and testing, increases DTM performance notably for the CNN and CRNN models from 0.73 to 0.78 and 0.78 to 0.82, respectively. This performance improvement suggests that, in DTM, it can be beneficial for a model not to have to generalize across many different drum kits. Conversely, using 1Kit with different drum kits in training and testing leads to a significant performance drop.

The first and last two rows of Table 3 show slight DTM

performance decreases for models trained on `10Kits` compared to `20Kits` when testing on `20Kits` and MDB Drums (`CRNN`: 0.78 to 0.76 and 0.75 to 0.73), with performance remaining constant on ENST Drums. In contrast, training with `1Kit` leads to lower performance for all models, with the `CRNN` model dropping to 0.63 on MDB Drums.

MDB Drums and ENST Drums include different drum sounds than STAR Drums, featuring recorded drums rather than synthesized audio. The small performance decrease observed for training with `10Kits` suggests that even a relatively low number of virtual drum kits allows for the models to perform well on unseen real-world data.

For DTD, a greater performance decrease compared to DTM is observed with `10Kits` when the drum kits differ between training and testing (`CRNN`: 0.91 to 0.81). For `1Kit`, we see a performance decrease of up to 0.2 in F-measure (`CRNN`: 0.91 to 0.71). This decline may be attributed to DTM benefiting from the presence of melodic instruments and singing, which act as implicit data augmentation by introducing background noise to the drum sounds [20]. Consequently, in DTM, the model does not rely as heavily on training with a diverse range of drum sounds as it does for DTD. In contrast, DTD is more prone to overfitting due to the lack of this additional variability. When drum kits are identical in both training and testing, performance remains consistent across 20 and 10 drum kits. For one drum kit, the performance increase is less pronounced than for DTM (`CNNSA`: 0.91 to 0.92).

In summary, DTM generally performs worse than DTD but benefits more when drum sounds in training and test are consistent and origin from a single drum kit. In contrast, DTD relies more on diverse drum sounds for robust generalization.

5.2 No Weak Onsets

For DTM, using `20KitsNoWeak` in training and testing, where the velocity of weak notes is increased to include only strong onsets, results in a performance improvement across all models, with an increase of 0.04 to 0.05 in F-measure. For example, the `CNN` model's global F-measure improves from 0.73 to 0.77.

For DTD, there is a consistent but small performance increase across all models, indicating the transcription errors related to weak onsets are not a strong performance-limiting factor for DTD.

When only testing on `20KitsNoWeak` and training on `20Kits`, the performance for DTM increases slightly and remains constant for DTD.

5.3 No Simultaneous Onsets

In DTM, using a dataset which does not include simultaneous onsets (`20KitsNoSim`) leads to a performance increase similar to that achieved by avoiding weak onsets, with improvements ranging from 0.03 to 0.05 in F-measure (`CRNN`: 0.78 to 0.83).

For DTD, simultaneous onsets are the main performance-limiting factor when training and testing with identical drum sounds, resulting in an F-measure increase of 0.07 to 0.08 when eliminated. This results in a nearly perfect transcription performance with an F-measure 0.98 for the `CRNN` and `CNNSA` models, compared to 0.91 on `20Kits`.

When testing on `20KitsNoSim` after training on `20Kits`, performance remains constant for the `CRNN` model and decreases for the other two architectures for DTM. This decrease is mainly due to a reduction in precision, especially for bass drum and snare drum. A possible explanation is that the models, having learned conventions about classes frequently occurring simultaneously, generate more false positives when such simultaneity is absent in the test data. For DTD, we observe a smaller performance gain compared to when both training and testing are conducted on `20KitsNoSim`.

5.4 Comparison of DTM and DTD Performance

The presence of melodic instruments and singing significantly limits DTM performance, leading to worse results in all evaluations carried out with STAR Drums. For DTD, simultaneous onsets are the only significant performance-limiting factor when drum sounds in training and testing are identical, while DTM performance is also affected by weak onsets and can increase when drum sounds originate from one single and identical drum kit in training and testing.

In DTM, models trained with the `20Kits` dataset generalized well to the drum sounds of MDB Drums, resulting in a small performance gap of 0.02 to 0.07 in F-measure between the results for the `20Kits` test split and MDB Drums (`CNN`: 0.73 and 0.71). Conversely, models trained for DTD on `20Kits` exhibited a larger performance gap, ranging from 0.1 to 0.17 in F-measure (`CNNSA`: 0.91 and 0.76). The reasons may be again attributed to melodic instruments and singing serving as data augmentation as outlined in Section 5.1.

5.5 Relative Performance Changes

Table 4 summarizes findings relative to the research questions outlined in Table 1. To provide a clearer understanding of the performance changes detailed in Table 3, we first calculated the average global F-measure across all three model architectures for each STAR Drums variant. Subsequently, we computed the relative performance changes by dividing these averages from each variant by the average global F-measure of the `20Kits` variant. Values are provided for both DTM and DTD.

For DTM, transcription performance improves by $5.0\,\%$ when all drum sounds during training and testing origin from the same drum kit. For DTD, the same experiment results in an $1.6\,\%$ improvement.

When drum kits in training and testing differ, reducing the number of drum kits from 20 to 10 leads to an $8.1\,\%$ performance drop for DTM and $11.3\,\%$ decrease for DTD. Further reducing to a single drum kit results in a $17.4\,\%$ decrease for DTM and a $21.5\,\%$ decrease for DTD. These findings suggest that diverse drum sounds in training are more critical for DTD than for DTM.

How does transcription performance change when ...	Change in global F-measure	
	DTM	DTD
... reducing the number of kits in training from 20 to 1, with same drum sounds in training and testing.	$+5.0\,\%$	$+1.6\,\%$
... reducing the number of kits in training from 20 to 10, with different drum sounds in training and testing.	$-8.1\,\%$	$-11.3\,\%$
... reducing the number of kits in training from 20 to 1, with different drum sounds in training and testing.	$-17.4\,\%$	$-21.5\,\%$
... no weak onsets are present.	$+5.5\,\%$	$+1.1\,\%$
... no simultaneous onsets are present.	$+5.1\,\%$	$+8.5\,\%$
... no sounds of melodic instruments and singing are present.	$+14.5\,\%$	

Table 4. Relative changes in global F-measure averaged across used DNN architectures when comparing the results of different versions of STAR Drums and when comparing all DTM results to all DTD results across all STAR Drums versions.

The absence of weak and simultaneous onsets in DTM leads to similar performance increases of $5.5\,\%$ and $5.1\,\%$, respectively. In contrast, weak onsets have a minimal impact on DTD, with only an $1.1\,\%$ increase. However, preventing simultaneous onsets in DTD yields a more substantial performance increase of $8.5\,\%$.

Finally, performing DTD compared to DTM results in an average performance increase of $14.5\,\%$ across all experiments conducted using STAR Drums variants.

5.6 Qualitative Analysis of Transcription Errors

After presenting the quantitative results, we manually inspected transcriptions from the best-performing `CRNN` model (trained on `20Kits`) for MDB Drums to identify systematic errors related to previous findings and the challenges outlined in Section 1.

Jazz excerpts with many soft onsets are challenging: weak snare, bass drum, and cymbal onsets are often missed in DTM. Additionally, percussive events from melodic instruments can cause false positives. For example, accentuated bass guitar notes are sometimes labeled as bass drum hits.

When a cymbal's decay overlaps with the attack of another drum sound, the system may falsely detect a cymbal in DTM, while the transcription is correct for DTD. Simultaneous hi-hat and snare sounds are sometimes missed, presumably due to masking and similar spectral features. In contrast, false positive hi-hat detections can occur when only snare drum is active. Similar confusions arise between bass drum and low toms, and between low-pitched snares and toms. Heavily distorted drum sounds in some items are not classified reliably, presumably because they differ too much from the sounds included in training.

These observations support our quantitative findings: simultaneity of drum sounds, weak onsets, limited generalization, and interference from melodic instruments remain key challenges for modern ADT systems.

6. CONCLUSION

In this study, we utilized the STAR Drums dataset to quantify several performance-limiting factors in ADT. We created five increasingly simplified versions of STAR Drums and conducted experiments using three different DNN architectures.

Our findings highlight three factors in DTM that increase performance by approximately $5\,\%$ each:

- Training and testing use identical drum sounds originating from a single kit.
- No weak onsets are present.
- No simultaneous onsets are present.

For DTD, simultaneous onsets are the central performance-limiting factor, with nearly error-free transcriptions achieved in their absence. Moreover, DTD benefits more from a diverse set of training drum kits than DTM.

Increasing the diversity of training data of STAR Drums can be achieved by employing various data augmentation techniques, such as pitch shifting, dynamic range compression, and reverberation.

Our findings suggest that the transcription quality of music education apps, which analyze student recordings to provide feedback, will decrease if a student plays advanced drum patterns with more simultaneous onsets, or uses a kit whose timbre differs strongly from the training data. Employing FSL could help mimic the effects of having identical drum sounds for training and testing. Additionally, creating versions of STAR Drums with a higher number of simultaneous onsets with varying class combinations could facilitate efficient learning of these complex scenarios.

In music production tools that use transcriptions to add or replace drum samples, genre-specific playing styles matter: for example, a jazz track with frequent soft ride and ghost notes will be transcribed less accurately than a pop track with more uniform dynamics, especially for DTM. Exploring the optimal ratio of weak to strong onsets in STAR Drums could further contribute to performance improvements.

By sharing these insights, we aim to support further advancements in ADT by providing clear indications of how addressing key challenges can enhance transcription performance.

7. ACKNOWLEDGMENTS

The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits IIS.

8. REFERENCES

[1] C. Wu, C. Dittmar, C. Southall, R. Vogl, G. Widmer, J. Hockman, M. Müller, and A. Lerch, "A review of automatic drum transcription," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 9, pp. 1457–1483, 2018.

[2] R. Vogl, G. Widmer, and P. Knees, "Towards multi-instrument drum transcription," in *Proc. of 21st DAFx'18*, 2018.

[3] M. Zehren, M. Alunno, and P. Bientinesi, "In-depth performance analysis of the adtof-based algorithm for automatic drum transcription," in *Proc. of 25th ISMIR*, 2024, pp. 1060–1067.

[4] P. Weber, C. Uhle, M. Müller, and M. Lang, "Real-time automatic drum transcription using dynamic few-shot learning," in *Proc. of 5th International Symposium on the Internet of Sounds (IS2)*. IEEE, 2024.

[5] R. Vogl, M. Dorfer, G. Widmer, and P. Knees, "Drum transcription via joint beat and drum modeling using convolutional recurrent neural networks," in *Proc. of 18th ISMIR*, 2017, pp. 150–157.

[6] Y. Wang, J. Salamon, M. Cartwright, N. Bryan, and J. Bello, "Few-shot drum transcription in polyphonic music," in *Proc. of 21st ISMIR*, 2020.

[7] M. Zehren, M. Alunno, and P. Bientinesi, "ADTOF: A large dataset of non-synthetic music for automatic drum transcription," in *Proc. of 22nd ISMIR*, 2021, pp. 818–824.

[8] I. Wei, C. Wu, and L. Su, "Improving automatic drum transcription using large-scale audio-to-midi aligned data," in *Proc. of ICASSP*. IEEE, 2021, pp. 246–250.

[9] R. Ishizuka, R. Nishikimi, E. Nakamura, and K. Yoshii, "Tatum-level drum transcription based on a convolutional recurrent neural network with language model-based regularized training," in *Proc. of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, 2020, pp. 359–364.

[10] 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, 2023.

[11] 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.

[12] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "The MUSDB18 corpus for music separation," Dec. 2017. [Online]. Available: https://doi.org/10.5281/zenodo.1117372

[13] P. Cano, E. Gómez, F. Gouyon, P. Herrera, M. Koppenberger, B. Ong, X. Serra, S. Streich, and N. Wack, "ISMIR 2004 audio description contest," *Music Technology Group of the Universitat Pompeu Fabra, Tech. Rep*, 2006.

[14] D. Bogdanov, M. Won, P. Tovstogan, A. Porter, and X. Serra, "The MTG-Jamendo dataset for automatic music tagging," in *Machine Learning for Music Discovery Workshop, International Conference on Machine Learning*, Long Beach, CA, United States, 2019.

[15] 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 *Proc. of WASPAA*. IEEE, 2019, pp. 45–49.

[16] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, and D. P. W. Ellis, "MIR_EVAL: A transparent implementation of common MIR metrics." in *Proc. of 15th ISMIR*, 2014, pp. 367–372.

[17] K. Takahashi, K. Yamamoto, A. Kuchiba, and T. Koyama, "Confidence interval for micro-averaged f_1 and macro-averaged f_1 scores," *Appl. Intell.*, vol. 52, no. 5, pp. 4961–4972, 2022.

[18] C. Southall, C.-W. Wu, A. Lerch, and J. Hockman, "MDB drums: An annotated subset of MedleyDB for automatic drum transcription," in *Proc. of 18th ISMIR*, 2017. [Online]. Available: https://github.com/CarlSouthall/MDBDrums

[19] O. Gillet and G. Richard, "ENST-drums: an extensive audio-visual database for drum signals processing," in *Proc. of 7th ISMIR*, 2006, pp. 156–159.

[20] B. McFee, E. J. Humphrey, and J. P. Bello, "A software framework for musical data augmentation," in *Proc. of 16th ISMIR*, 2015, pp. 248–254.