On the Importance of Temporally Precise Onset Annotations for Real-Time Music Information Retrieval: Findings from the AG-PT-set Dataset

Domenico Stefani*
Gregorio Andrea Giudici*
Luca Turchet
domenico.stefani@unitn.it
gregorio.giudici@unitn.it
luca.turchet@unitn.it
Department of Information Engineering and Computer Science, University Of Trento
Trento, Italy

ABSTRACT

In real-time Music Information Retrieval (MIR), small analysis windows are essential for achieving low retrieval latency. In turn, event-based real-time MIR methods require precise onset detectors to correctly align with the beginning of events such as musical notes. Detectors are typically trained using ground-truth annotations from datasets of interest. Yet, most MIR datasets do not prioritize the accurate timing of onset labels, and the evaluation of detectors often relies on generous tolerance windows (even ±50ms). In this paper we present AG-PT-set, a new dataset of acoustic guitar techniques with precise onset annotations. The dataset features 32,592 individual notes and over 10 hours of audio, covering eight techniques. Moreover, we assess the importance of exact onset labels across multiple real-time MIR tasks. Our results show how accurate timing of onset labels and precise detectors are crucial for real-time MIR tasks, as the performance of most algorithms degrades with imprecise onsets. In few occasions, imprecise onset timing slightly improved results, hinting at a possible similarity to data augmentation methods. Taken together, our findings indicate that temporally precise labels and detectors are always preferable, as robustness can always be obtained via artificial augmentation, while precision cannot be obtained as easily

CCS CONCEPTS

• **Information systems** → **Music retrieval**; **Evaluation of retrieval results**; **Retrieval models and ranking**.

KEYWORDS

Music Information Retrieval, Real-time, Audio Processing

*Both authors contributed equally to this research.

AM '24, September 18–20, 2024, Milan, Italy

ACM ISBN 979-8-4007-0968-5/24/09
https://doi.org/10.1145/3678299.3678325

ACM Reference Format:
Domenico Stefani, Gregorio Andrea Giudici, and Luca Turchet. 2024. On the Importance of Temporally Precise Onset Annotations for Real-Time Music Information Retrieval: Findings from the AG-PT-set Dataset. In *Audio Mostly 2024 - Explorations in Sonic Cultures (AM '24), September 18–20, 2024, Milan, Italy*. ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3678299.3678325

1 INTRODUCTION

Real-time Music Information Retrieval (rtMIR) tasks often have to rely on very small signal analysis windows to extract meaningful features or signal properties. In the case of event-based real-time MIR (e.g., real-time pitch detection of a note), feature extraction windows are often aligned to the beginning of the event (e.g., a note event) with the aid of an onset detector. The onset of a note event is defined as the time at which the sound starts, which also marks the beginning of the attack phase that leads to the amplitude peak (see Figure 1).

Onset detectors can be divided into function-based detectors [3, 8, 13, 14, 16, 19, 26] or trainable detectors (i.e., machine learning-based) [5, 15]. The former type of detector exposes several parameters that are often tuned to perform best with the specific target sounds [30], while the latter comprises detectors that are directly trained on audio datasets, often in a supervised manner with ground-truth onset time annotation. In both cases, the best-performing onset detectors tend to be quite specialized for certain types of sounds and music [15] and must be tuned/trained for the specifics of the sounds at hand. As a result, the detection performance is as good as the alignment of ground-truth annotations.

However, to date many onset-annotated audio datasets that are freely available do not prioritize time-precision of the labels and, as a result, onset labels can be off by tens or even hundreds of milliseconds. As anticipated, this can have detrimental effects on both the training and evaluation of onset detectors, which tend to conform to the degree of precision of ground-truth labels. The performance of many real-time MIR systems can be affected by these detectors, originating from the poor precision of onset annotations. In particular, MIR systems that are meant to run in real-time (i.e., on performance systems, digital musical instruments) often use small signal analysis windows (few milliseconds) that are aligned to detected onsets. This is so that contained retrieval latency is

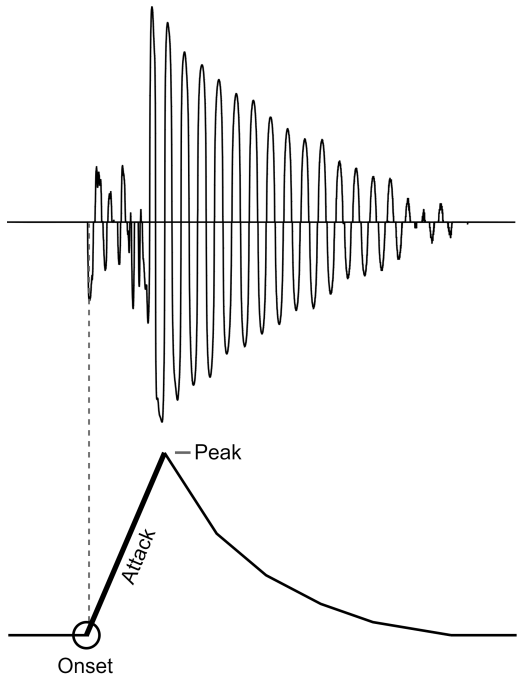

Figure 1: Onset of an individual note, identified as the beginning of the attack phase in the amplitude envelope. Adapted from [1].

maintained. In these cases, poor time-precision of onset detectors can lead to analysis windows having varying alignment with the beginning of notes, and potentially missing crucial parts of information in the attack phase of notes.

The scarcity of such precisely annotated datasets was particularly felt by the authors in relation to the guitar and the automatic playing technique recognition task. Such task is relevant to the design of intelligent musical instruments such as smart guitars capable of understanding in real-time what and how the guitarist is playing [35].

In this paper, we present *AG-PT-set*, a new dataset of individual, monophonic acoustic guitar sounds played with 8 different expressive playing techniques by 7 different players with 6 different instruments. The dataset was accurately annotated at the onset level by five musicians. The labeled part of the dataset comprises 32,592 notes totaling 10 hours and 4 minutes of 24-bit WAV recordings. Furthermore, the dataset contains an additional 4 playing techniques whose onset times have not been annotated yet (5 hours and 1 minute). The audio files and relative metadata (i.e., onset and pitch annotations) are available on Zenodo[1].

Furthermore, with the present study we set out to prove the relevance of precise annotations and detectors for real-time MIR tasks through a series of experiments. In particular, we approached the tasks of pitch detection, binary and multiclass playing technique classification, and playing dynamics classification.

[1]https://zenodo.org/doi/10.5281/zenodo.10159491

Each real-time task was addressed with two methods: a Deep Neural Network (DNN) and either a computational or Machine Learning (ML) method. Onset labels were then progressively perturbed, mimicking onset detectors trained/tuned on poorly aligned onset labels. Perturbation was performed up to a maximum of ± 50ms, which still would yield 100% correct detection score with the evaluation strategies of some studies [13, 22, 32] and challenges (e.g., MIREX, see Section 2). The code for all the experiments reported in this study is made freely available online[2].

The remainder of the paper is organized as follows. Section 2 presents related guitar datasets and onset detection studies. In Section 3 we present AG-PT-set in detail. Section 4 describes the experiments on the effect of varying onset label precision of various MIR tasks. Then we present the results in Section 5 and discuss them in Section 6. Finally, we draw our conclusions in Section 7.

2 RELATED WORKS

2.1 Guitar Datasets

With the increase of data-driven MIR methods and the importance of the guitar as a popular and widespread musical instrument, quality data with annotations have become highly valuable resources. However, a lack of comprehensive audio datasets specific to guitar has been widely acknowledged [36], especially those with annotated onset times[3].

The more relevant and freely available guitar audio datasets we identified are the following:

- EGDB (2022, Chen *et al.* [10]): this dataset contains different tone renditions and transcriptions of 240 guitar tablatures, executed on the electric guitar.
- GuitarSet (2018, Xi *et al.* [36]): this dataset was envisioned for automated guitar transcription, and contains 360 acoustic guitar excerpts recorded with an hexaphonic pickup.
- IDMT-SMT-Guitar (2014, [22]): This dataset contains about 5100 note events in its two original partitions and an additional 5 short and 64 longer pieces which were added in the following versions. Each of the four partitions of the dataset has different annotations, recording quality, and characteristics (e.g., monophonic and polyphonic). The first three parts of the datasets are recorded with electric guitars only, while the last includes acoustic guitars.
- Guitar Playing Techniques dataset (2014, Su *et al.* [33]): this dataset reportedly contained 6580 clips of single notes along with playing technique annotations (seven techniques). Despite acknowledging its existence, we were not able to include the dataset in our comparisons as the hyperlink to the dataset website has been broken for a number of years now, and attempts to contact the authors to obtain the data have been unsuccessful.
- EG-solo (2023, Huang [21]): this is an electric guitar solo dataset consisting of 6833 note events annotated with onset times, pitches, and playing techniques (e.g., palm mute, pull-off, harmonics). The authors provide MIDI and technique annotations, while the audio was obtained from several YouTube videos, and therefore cannot be provided separately.

[2]https://github.com/CIMIL/AG-PT-set_AM24_accompanying-material
[3]Some datasets can be found at https://ismir.net/resources/datasets/

Table 1 reports some of the characteristics of the mentioned datasets. Dataset size is not reported as each entry reports different metrics (i.e., number of note events, duration) or even a different metric for separate parts of the dataset (e.g., IDMT-SMT-Guitar).

Despite these efforts, there is still a significant lack of large-scale, well-annotated guitar audio datasets, particularly those with precise onset time annotations. This scarcity of suitable datasets poses challenges for researchers and developers working on guitar-related audio applications, such as transcription, source separation, and effects processing. The absence of comprehensive freely available guitar datasets can be attributed to several factors, including the time-consuming nature of manual annotation.

2.2 Time Precision of Onset annotations and detection

Large part of freely available datasets for MIR research do not include millisecond-accurate onset annotations [34]. Moreover, the annotation process is often not documented, as for the guitar datasets described in the previous section. In particular, The authors of EGDB [10] produced onset annotations via an onset detector. The authors also indicated that a few labels required manual corrections but provided few details on the process. In [36] the authors of GuitarSet used a similar "semi-automatic" approach to produce onset annotations with manual validation but did not provide other details. The authors of IDMT-SMT-Guitar [22] describe the use of a manual annotation procedure but only provide information about the file format of the annotations. Furthermore, the authors of the Guitar Playing Techniques dataset [33] provide no detail on the annotation procedure. Finally, the authors of EG-Solo [21] report the manual annotation of onsets with the aid of guitar tablatures.

However, the accurate annotation and detection of onsets are rather crucial in MIR, as the identification of the beginning of note events is often used to align signal analysis windows for algorithms that can extract various types of information and features. These can, in turn, be fed to machine learning algorithms to extract higher-level characteristics such as gestures, playing techniques, or note information for automatic transcription [6]. Alternatively, following more modern approaches, the recognition step can be integrated with feature extraction, and even onset detection, in a single *end-to-end* deep-learning neural network [29].

In this context, the aforementioned variance in the time precision of onset annotations in freely available datasets has two main negative implications:

(1) **Misleading Evaluation:** The evaluation of MIR methods is naturally affected by inaccurate dataset annotations in a negative manner. This holds for onset annotation in onset-based MIR methods. Furthermore, attempts to evaluate onset detection methods on loosely annotated data led many to use severely large tolerance windows around ground-truth annotations (e.g., ±50ms for the Music Information Retrieval Evaluation eXchange (MIREX) onset detection challenge, see the next section). One of the objectives of this paper is to show how even a detector that obtained a 100% accuracy score according to such tolerance windows, can be widely unsuitable for real-time MIR tasks;

(2) **Training with bad data:** The training of neural onset detectors and end-to-end MIR networks for onset-based tasks is affected by the quality of onset annotations in a garbage-in-garbage-out manner [17]. Additionally, data-based parameter tuning of parametric onset detectors will follow a similar behavior [30].

2.3 Large tolerance windows for onset evaluation

With regard to the issue of "misleading evaluation", we report the most relevant works on onset detection that employed what can be considered a rather large tolerance windows for real-time MIR.

Firstly the MIREX Onset Detection challenge states that the evaluation of the performance of challenge entries is done with a ±50ms tolerance window to consider detection correct. The reasoning reported in the challenge rules is the following: *"Time precision (tolerance from ±50ms to less). For certain file, we cannot be much more accurate than 50ms because of the weak annotation precision"* [4]. Other works that evaluated onset detectors with a ±50ms window include those by Dixon *et al.* [13], Stowell *et al.* [32], and Kehling *et al.* [22]. Bello *et al.* [1] describe using a *"relatively large"* 50ms window, without further specifying whether it referred to the entire length of a window centered on the ground-truth onset, or a ±50ms window as Bock *et al.* later argued [4].

A smaller tolerance window of ±25ms was used by others such as Böck *et al.* [4], who stated that this "rather strict" evaluation method yields more meaningful results for online detection than ±50ms. Other studies using 25ms tolerance windows include [15], [28], [27], and [7]. A different approach is used in [2], where the authors compare different detection methods for increasing tolerance window sizes (from 10 to 100ms). However, results showed dramatically low percentages of good detections for smaller windows.

2.4 Time Precision of dataset onset annotation

With respect to the mentioned guitar audio datasets, we set to sample the time precision of their onset annotations where possible. To do so, we extracted 100 random note events for each and asked one annotator (a musician) to manually label those that were the most clearly distinguished from other notes in the audio signal. This resulted in 50 to 80 labels for each dataset. Labeling was performed similarly to the dataset presented here, using Audacity with track multiview and zooming at millisecond level (see Section 3). An example of one of the corrections is presented in Figure 2.

Then, an error measure was obtained by subtracting the relative dataset labels from the new "corrections". A random sample of 50 notes was taken for each dataset. The Guitar Playing Techniques dataset was excluded since it was not available online. EG-solo was excluded since it required downloading the audio data from YouTube (i.e., highly compressed recordings) with backing tracks, and precisely annotating onsets was not possible. The respective absolute error distributions can be seen in Figure 3 and Table 2.

[4]From the MIREX 2021 Onset Detection challenge website https://www.music-ir.org/mirex/wiki/2021:Audio_Onset_Detection.

Table 1: Overview of other guitar audio datasets with onset annotations.

Dataset	Guitar Type	Transducer	Annotated Playing Techniques	Onset annotations
EGDB [10]	Electric	Internal	✗	✓(MIDI)
Guitar Set [36]	Acoustic	Internal	✗	✓
IDMT-SMT-Guitar [22]	Acoustic, Electric	Internal and External	3 excitation styles and 6 expression styles	✓
EG-Solo [21]	Electric (solo)	Internal	9 techniques	✓
AG-PT-set (Ours)	**Acoustic**	**Internal**	**12 playing techniques**	**✓(for 8 techniques)**

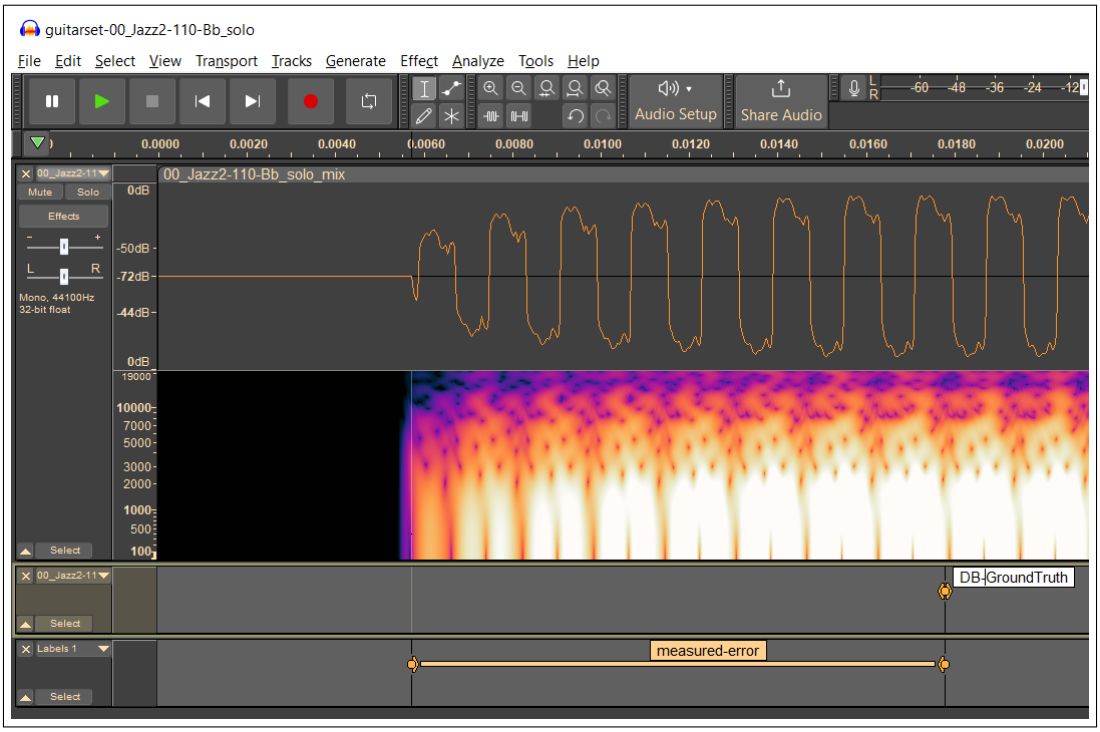

Figure 2: Example of one of the measurements of time precision of an onset annotation in GuitarSet which was misaligned with the real onset. The single label just below the spectrogram represents the misaligned dataset label, while the label bracket below represents the measured error.

Table 2: Mean and standard deviation of absolute onset time error measured on a small random sample of notes for each dataset (50 notes each). These measures were obtained by relabeling a random subset of onsets from each dataset and subtracting the dataset labels from the corrected labels.

Dataset	Mean Onset Error [ms]	Std Dev. [ms]
EGDB	34.6	31.4
GuitarSet	12.4	3.8
IDMT-SMT-Guitar	8.5	9.2

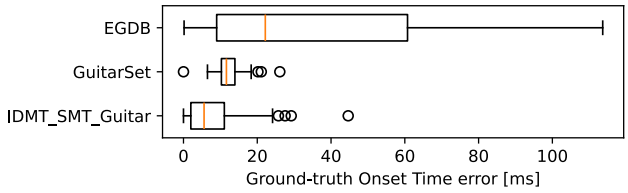

Figure 3: Absolute onset time error measured for three guitar audio datasets.

2.5 Datasets of Guitar Playing Techniques

Beyond the scarcity of guitar datasets in general, those that contain playing technique annotations are even fewer. However, these are required for the MIR task of automated *playing technique recognition*. This dataset originated from the authors' need for a dataset for acoustic guitar technique recognition. The dataset was briefly mentioned - while in development - in previous studies (see [30, 31]) and is now made public and described in detail.

3 THE DATASET

The Acoustic Guitar Playing Technique dataset (AG-PT-set) contains 10 hours and 4 minutes of accurately labeled recordings of individual acoustic guitar sounds (no polyphony) played with 8 playing techniques. Of these, four are common pitched techniques, (i.e., *Pick Over the Soundhole*, *Pick Near the Bridge*, *Palm Mute*, and *Natural Harmonics*), while the remaining four are percussive techniques. The latter type comprises techniques that are part of percussive fingerstyle.

The choice of percussive techniques drew inspiration from Martelloni *et al.* [25]. For the four selected percussive techniques, we adopted names from drum-kit pieces that most resembled their use in percussive fingerstyle, i.e., *Kick*, *Snare-A*, *Snare-B*, *Tom*. The onset-labeled portion of the dataset contains 32,592 notes. The labeling process is described in Section 3.3.

Moreover, the dataset contains five additional hours of recordings with four other playing techniques whose onset times have not been annotated yet (i.e., *Half-tone Bending*, *Hammer-on*, *Staccato*, and *Vibrato*). Despite missing onset labels, recordings for the four additional techniques are labeled with player ID, guitar ID, playing technique, and playing dynamics information. Moreover, each recording contains individual notes played starting from fret 0 (open string), on a single guitar string, with 3 repetitions per fret, making it possible to use them for other MIR tasks such as automatic transcription or player identification.

3.1 Playing Techniques

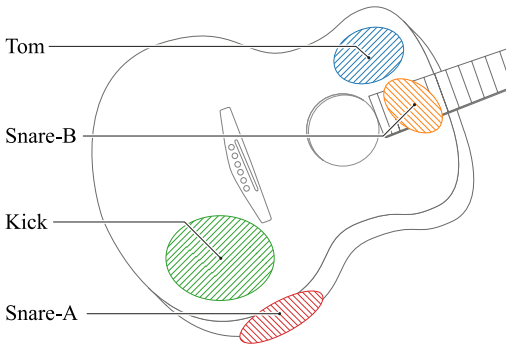

Figure 4: Diagram displaying the contact positions for the percussive selected for the dataset (from [31]).

The playing techniques recorded for the dataset are the following:

(1) *"Kick" technique* (percussive): producing a sound that resembles a kick drum by hitting the lower right part of the top of the guitar body (see 4);
(2) *"Snare-A" technique* (percussive): producing a sound by hitting the lower right side of the guitar body;
(3) *"Tom" technique* (percussive): producing a sound by hitting the area of the guitar body near the top of the end of the fretboard, using the thumb;
(4) *"Snare-B" technique* (percussive): producing a sound by hitting the muted strings over the end of the fretboard;
(5) *Bending technique*[5] (pitched): pulling the strings, raising the pitch (half-tone interval);
(6) *Hammer-on technique*[5] (pitched): sharply bringing a finger down onto the fingerboard, creating a legato sound (half-tone interval);
(7) *Natural Harmonics* (pitched): plucking the strings while lightly touching the string with the fretting finger (i.e., not pressing the string fully), therefore letting only some harmonic overtones ring;
(8) *Palm Mute*: partially muting the strings with the palm of the picking hand, resulting in a muffled sound.
(9) *"Pick Near Bridge"* (pitched): plucking the string near to the guitar bridge, producing sounds with great high-frequency content;
(10) *"Pick Over the Soundhole"* (pitched): plucking the string over the soundhole, producing sounds with lower treble content and greater intensity;
(11) *Staccato* (pitched): playing short notes;
(12) *Vibrato* (pitched): Moving the fretting finger to warp the pitch and tone of the sound.

Percussive techniques were recorded by each guitarist with three dynamics (i.e., *piano*, *mezzoforte*, *forte*). For the first four musicians, 10 repetitions of each percussive technique were performed for each dynamic level. For the remaining musicians, the number of repetitions was first increased to 100 and then 300 for a more fair balance with pitched techniques. The pitched technique classes encompassed more repetitions simply due to the number of strings and frets played. For these, individual notes were played within specified fret ranges[6].

Each fret on each string was played three times for each of the dynamics/intensity levels mentioned earlier. Therefore, as anticipated, most pitched techniques total a higher number of recorded notes with respect to percussive sounds. Table 3 contains a summary of all the techniques with details about their type (percussive or pitched), total recorded time, and time of labeled recordings. The distribution of dynamics between the recordings in the dataset is presented in Table 4.

3.2 Recordings

The present study is part of a larger project targeting the creation of self-contained instruments such as smart guitars [35]. As such, the dataset was envisioned for real-time MIR tasks to be performed on small embedded computers placed inside guitars, and the audio was captured with the internal transducers of each guitar. All the guitars used featured a piezoelectric transducer under the bridge of the instrument, and 6 out of 7 featured an additional internal condenser microphone. Where present, the condenser microphone was used in conjunction with the piezoelectric pickup (50% blend). The guitars used in the dataset are presented in Table 5. Audio was

[5] As for other techniques, note onsets for the bending and hammer-on techniques are considered to correspond to the very beginning of the note event.
[6] Pitched techniques cover a range from open strings up to a fret between the 15th and 20th, depending on the physical attributes of each guitar (e.g., cutaway, string gauge, guitar scale). Natural harmonics were recorded only for frets 5, 7, and 12 as these are the most likely to resonate with sufficient loudness across different guitars.

Table 3: Playing techniques in the dataset with the respective type (i.e., pitched or percussive technique), the total recorded time, and the recording time of the relative labeled portion of the dataset. The percentages next to the total and labeled portion time refer to the fractions of the entire and labeled portion of the dataset represented by the technique recordings.

ID	name	Type	Tot Rec. Time	Labeled Rec. Time	# Labeled Notes
1	Kick	percussive	20' (2.2%)	20' (3.3%)	1,950
2	Snare-A	percussive	19' (2.0%)	19' (3.2%)	1,968
3	Tom	percussive	19' (2.1%)	19' (3.3%)	2,167
4	Snare-B	percussive	19' (2.1%)	19' (3.3%)	1,943
5	Natural Harmonics	pitched	52' (5.6%)	52' (8.7%)	2,092
6	Palm Mute	pitched	1h 49' (11.8%)	1h 49' (18.2%)	7,588
7	Pick Near Bridge	pitched	2h 40' (17.1%)	2h 40' (26.5%)	7,227
8	Pick Over the Soundhole	pitched	3h 22' (21.7%)	3h 22' (33.6%)	7,657
9	Bending	pitched	1h 17' (8.3%)	0	0
10	Hammer-on	pitched	1h 02' (6.7%)	0	0
11	Staccato	pitched	1h 11' (7.7%)	0	0
12	Vibrato	pitched	1h 58' (12.7%)	0	0

Table 4: Distribution of playing dynamics in the dataset.

Dynamics	Tot Rec. Time	Labeled Rec. Time
forte	4h 55' (31.0%)	3h 05' (30.8%)
mezzoforte	5h 53' (37.0%)	3h 48' (37.8%)
piano	5h 00' (31.5%)	3h 09' (31.4%)

recorded at 48kHz with 24bit depth as WAV files with an Elk Pi Audio Board.

3.3 Onset Labeling Process

The onsets for techniques 1 to 8 were labeled by five musician. Recordings were split between the five annotators depending on their respective availability, with no overlap. Annotation was carried out with the Audacity[7] software. Annotators were provided with an audacity configuration[8] file with specific settings to better label onsets: in particular, the track visualization was set to a split between the audio signal (in dB) and a Mel Spectrogram with very high temporal resolution, to quickly identify onsets (See Figure 5). Additionally, the configuration included a key combination to toggle the zoom between the default and a zoomed-in visualization where each tick in the time ruler identified a single millisecond. The lower bound of the signal visualization range was also set to -84dB instead of the default -60 dB, to help visualize quieter sounds. Annotation projects were prepared with candidate onset labels placed with an onset detector (*aubioonset* [9]) to aid the labeling process. Annotators were prompted to go through each file by first marking each missed onset from the wide zoom configuration, then removing potential false positives, and finally toggling to the zoomed-in view and precisely aligning each label to the onset with the visual aids. Since the number, pitch, and sequence of notes in each file were known, these were used to identify annotation mistakes (with the aid of a pitch detector), which were promptly fixed afterwards. The number of onsets labeled by each annotator is presented in Table 6.

[7]https://www.audacityteam.org/

[8]The configuration file and annotation instructions are available in the project's repository: https://github.com/CIMIL/AG-PT-set_AM24_accompanying-material

3.4 Data and Metadata

The dataset is composed of audio data and annotations (metadata). The dataset directory contains `data` folder which includes the folder `audio` with the WAV recordings. Furthermore, `data` contains the folders `onset_labels` and `pitch_labels` with label files formatted for Audacity, one for each audio file. The `metadata` folder contains the following CSV files:

(1) `expressive_techniques.csv`: This contains a description of the different expressive techniques in the dataset. The most relevant columns are:
   (a) *"id"*: numeric id,
   (b) *"name"*: technique name,
   (c) *"description"*;
(2) `files.csv`: This contains labels for all the audio files in the dataset. The most relevant columns are:
   (a) *"filename"*: filename (table key),
   (b) *"duration"*: duration in seconds,
   (c) *"bits_per_sample"*: bit-depth,
   (d) *"guitar_id"*: id of the guitar used,
   (e) *"player_id"*: id of the guitarist,
   (f) *"playing_dynamics_or_intensity"*: *piano*, *mezzoforte*, or *forte* dynamics,
   (g) *"sha256"*: SHA256 checksum to verify file integrity,
   (h) *"labeler_id"*: id of the labeler;
(3) `instruments.csv`: This contains a description of the different instruments in the dataset similarly to Table 5.
(4) `note_labels.csv`: This contains labels for all the notes in the dataset. The most relevant columns are:
   (a) *"onset_label_seconds"*: onset label in seconds (Composite key 1),
   (b) *"audio_file_path"*: filename of the WAV containing the note (Composite key 2),
   (c) *"onset_label_samples""*: onset label in samples,

Table 5: Acoustic Guitars used to record the dataset.

ID	Brand	Model	Pickup	CM?*	Tot rec. time	Labeled rec. time
0	Eko	unknown	Factory Option	✓	2h 54' (18.3%)	1h 42' (16.9%)
1	Eko	WOW 018 KOA	Fishman Flex Blend	✓	2h 44' (17.2%)	1h 20' (13.3%)
2	Taylor	114 CE	LR Baggs Anthem	✓	2h 53' (18.1%)	1h 27' (14.4%)
3	Maton	EBG-808	AP mic	✓	2h 30' (15.8%)	1h 21' (13.5%)
4	Yamaha	APX-8A	Factory Option	✗	1h 50' (11.6%)	1h 18' (13.1%)
5	Crafter	TB-G-1000	Factory Option	✓	1h 31' (9.6%)	1h 23' (13.9%)
6	Crafter	GLXE-4000/RS	Factory Option	✓	1h 30' (9.5%)	1h 30' (15.0%)

CM?: Has condenser microphone. If true, the condenser microphone was set to a 50% blend with the piezoelectric transducer for recording. If not, only the piezoelectric pickup was used.

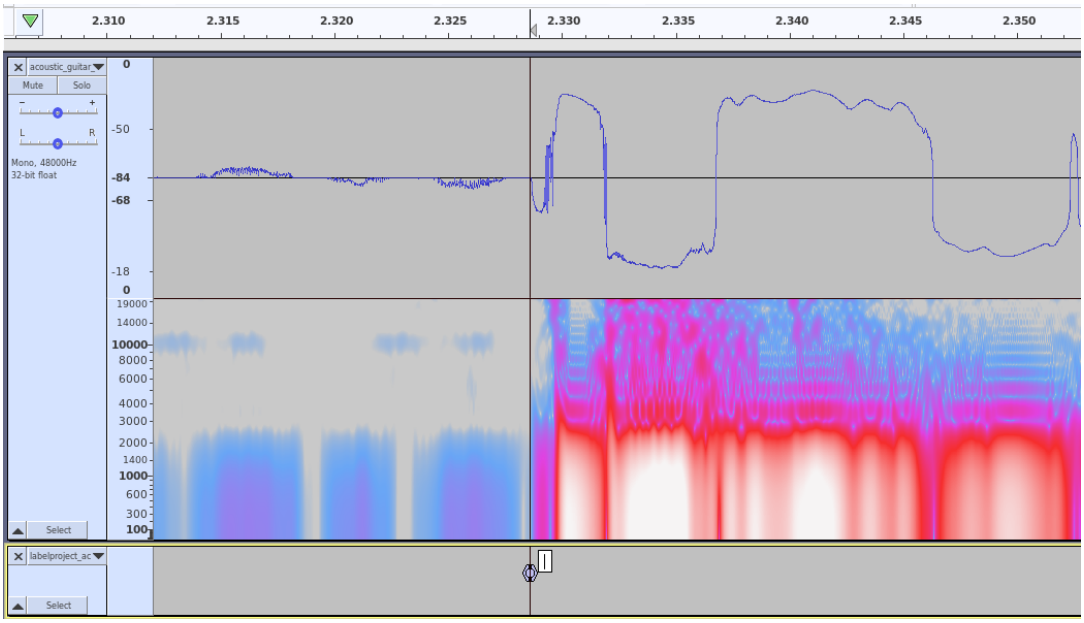

Figure 5: Annotated onset in Audacity with the time zoom set so that the time ruler ticks are one millisecond apart.

Table 6: Number of onset times labeled by each annotator and the duration of the labeled recordings.

Annotator	# Labeled Onsets	Tot Rec. Time
A1	6,981 (21.4%)	2h 13'
A2	11,041 (33.9%)	2h 56'
A3	4,988 (15.3%)	1h 39'
A4	4,534 (13.9%)	1h 36'
A5	5,048 (15.5%)	1h 38'

(d) *"expressive_technique_id"*: Foreign key pointing to expressive_techniques.csv,
(e) *"pitch_midi"*: Ground-truth pitch as MIDI number,
(f) *"string_number"*: from 1 (low E) to 6 (high E),
(g) *"playing_intensity"*: dynamics;

Finally, the dataset includes the `AG_PT_set.py` Python script which can be used to load the four CSV files as Pandas DataFrames[9] and perform integrity tests on load.

4 EXPERIMENTS

We selected four different tasks to demonstrate the importance of precise onset labels for real-time MIR. The selected tasks fall into the category of real-time MIR algorithms that are triggered by the detection of onsets and are meant to extract relevant information (e.g., pitch, playing technique) as soon as possible. Moreover, approaches that address each task can be evaluated with ground-truth annotations other than the onset times themselves. Each task was

[9]https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.html

therefore addressed with different analysis window sizes between 6.25 and 100ms, resulting in different retrieval latencies. The tasks are introduced in Section 4.1. The analysis windows were first aligned with the ground-truth onsets, and then with progressively perturbed onset labels, up to a maximum of ±50 ms. The onset perturbation is a proxy for either a loosely labeled dataset or a rather temporally imprecise onset detector, and a decrease in performance would indicate a correlation between the fine temporal alignment of onset labels and task results. The perturbation of the onset labels is described in Section 4.3.

Furthermore, each task was addressed with two methods: a deep learning method and either a computational or conventional ML solution[10]. These tasks represent a few examples of possible use cases of the proposed dataset. The code for the experiments is available in the project's repository[11].

4.1 Tasks

The four selected MIR tasks are the following:

(1) **Pitch Detection Task (PD):** the process of identifying the fundamental frequency of a musical sound, i.e., the lowest frequency component of a complex sound wave.
(2) **Percussive/Pitched Binary Classification Task (BC):** a classification task set to distinguish between the two main categories of playing techniques in the dataset: pitched and percussive. Pitched techniques are characterized by producing a discernible musical pitch. On the other hand, percussive sounds lack a distinct pitch and are characterized by their transient nature and emphasis on attack and decay rather than sustained tones.
(3) **Dynamics Classification Task (DC):** The categorization of musical notes based on their playing dynamics. Each category represents a different level of loudness and timbral properties, with *forte* representing a strong intensity, *mezzo-forte* indicating a medium intensity, and *piano* denoting a soft or quiet intensity. This dynamic classification approach is valuable in audio processing applications where controlling and categorizing the dynamic levels of audio signals are essential for achieving desired sound aesthetics and perceptual effects.
(4) **Multiclass Playing Technique Classification (TC):** categorization of musical notes into one of multiple classes describing the playing technique employed by the performer. This type of classification system can provide valuable insights into the performance styles and expressive nuances used by musicians, and complement tasks such as music transcription, instrument recognition in audio recordings, and the development of interactive music learning tools.

4.2 Experimental Setup

For each task, we run two respective retrieval methods with five different analysis windows of 6.25ms, 12.5ms, 25ms, 50ms, and 100ms, respectively. Each analysis window was aligned to either the ground-truth labels (without perturbation) or to a perturbed version of the labels. Perturbation was performed by applying a Gaussian-type distribution to the ground-truth labels with 3 different ranges of perturbation, i.e., ±5ms, ±25ms, or ±50ms (See Section 4.3).

For the PD task, two types of pitch detectors, Yin [11] and CREPE [23], were used to analyze the influence of onset both with computational methods and deep learning methods for monophonic pitch estimation. Yin is an algorithm that estimates the fundamental frequency, or pitch, of an audio signal by analyzing its autocorrelation function and identifying the lowest valley in the Cumulative Mean Normalized Difference Function (CMND). In contrast, CREPE is a deep learning-based pitch detection algorithm that uses Convolutional Neural Network (CNN) to directly estimate the fundamental frequency of audio signals from raw spectrograms. For this particular task, we used a pre-trained version of CREPE, which was obtained via pypi[12].

For the remaining classification tasks (BC, DC, and TC), both a ML classifier and a DNN classifier were trained. For the ML classifier, we chose a K-Nearest Neighbors (KNN), with K=3 neighbors, because of its simplicity and flexibility with no assumptions on data distribution. The features used as input for the KNN were Mel Frequency Cepstral Coefficient (MFCC), spectral centroid, spectral bandwidth, and Root Mean Square (RMS). The DNN classifier used was a ResNet18 model [20], pre-trained for image classification on ImageNet [12] and then fine-tuned for audio classification with our proposed dataset. The adaptation of pre-trained models on data from different domains is an already established method [18, 24], and ResNet has been a cornerstone in the development of DNN for image classification, often serving as a baseline for comparison in research and benchmarking competitions due to its consistently high performance. The inputs fed to the ResNet model were the logarithmic Mel spectrograms, delta features, and delta-delta features.

For each classification task, we applied Stratified Grouped k-fold cross-validation (k = 3), natural groups in the data (i.e., guitar player) are kept separate, providing an accurate estimate of model performance. Additionally, to address class imbalance within the dataset for some tasks, we employed Synthetic Minority Oversampling Technique (SMOTE), which artificially inflates the number of training samples of minority classes via synthetic generation. This approach helps mitigating the impact of class imbalance on model training and evaluation, thereby enhancing the robustness and reliability of our classification results.

4.3 Onset Perturbation

To test how onset label/detection time precision can affect various real-time MIR tasks, we perturbed the ground-truth onset values using a Gaussian-type distribution. This was found to closely resemble time-error distributions from many onset detectors. We implemented a Gaussian probability density function (PDF) with a mean equal to zero and a standard deviation equal to 1/3 of the maximum range value, so that 99.73% of the perturbation values generated by the distribution would fall within the range value.

Three different perturbation ranges were chosen. The first and largest perturbation range of ±50ms was chosen following the

[10] Despite deep learning being a subset of ML, we follow common practice and refer to non-deep learning approaches as conventional ML or just ML.
[11] https://github.com/CIMIL/AG-PT-set_AM24_accompanying-material
[12] https://pypi.org/project/crepe/

guidelines of the MIREX onset detection challenge[13]. The rationale behind this choice is that an onset detector yielding the same distribution of time-precision would have the same evaluation score as the much more precise onset labels, according to MIREX, while potentially negatively affecting the precision of triggered MIR methods. The second range of ±25ms was chosen following the range used in [27] for validating the CNN-based onset detector. This is considered by many to be the current state-of-the-art. The last perturbation range is ±5ms, representing an optimal onset detector. This was obtained by performing a series of experiments by tuning the parameters of various onset detectors of the Aubio library [8, 9], obtaining on our dataset an F1-Score of 98.4% with the onset detector "hfc". The resulting error distribution yielded a standard deviation value of about 2ms.

5 RESULTS

Separate results for each task are presented in this Section. For each task, we present two plots, respectively reporting the results for the computational/ML approach (i.e., Yin or KNN) and the DNN (i.e., CREPE or ResNet).

5.1 PD Task

Results for the PD task are presented in Figures 6 and 7. The absolute pitch error was measured as the distance between the pitch value and the output from the pitch trackers (in MIDI notes). As expected, the Yin pitch tracker obtained progressively lower error with larger analysis windows. Moreover, for all analysis window sizes, the first and third quartile (i.e., the extremes of the boxplots) tend to spread as the applied perturbation range increases (see Figure 6). Similarly, CREPE obtained a reduction in error with larger windows, but in contrast to the Yin tracker, the results with no perturbation and 5ms perturbation are almost identical (see Figure 7)

The performance of the Yin detector is acceptable for real-world applications starting from 50ms windows, as long as the variability of onset timing does not exceed ±25ms. CREPE instead shows far lower errors with 25ms analysis windows, as long as the start of windows is within ±5ms of the actual onset. Interestingly, most errors with larger onset variability correspond to one octave. However, computation time can differ depending on the approach, and the choice between the two can depend on both accuracy and total retrieval latency.

5.2 BC Task

Results for the BC task are presented in Figures 8 and 9. The vertical axis of each plot reports the macro average F1-Score between classes, averaged across 3-fold group cross-validation. KNN struggled to obtain acceptable results with varying windows sizes. With respect to onset perturbation, KNN yielded a consistent slight decrease of performance with the smallest analysis window (6.25ms) and every range of perturbation. The same decrease is observed with 12.5ms analysis windows and perturbation in the 5ms and 25ms ranges. With 25ms and 50ms analysis windows, KNN shows a slight but progressive performance increase with larger perturbation ranges (at most < 0.05). With the larger 100ms window, performances with varying perturbation ranges remain unchanged, except for the 50ms perturbation range, where a slight improvement is observed (< 0.03).

Differently, ResNet shows consistently better performance, reaching 0.98 F1-score points with the larger analysis window and no perturbation. Moreover, the performance decrease of ResNet was consistent with onset perturbation, with the exception of the 25ms analysis windows where slight improvements are observed. The largest performance gaps were found with the smallest analysis window, with performance changes of more than 0.3 F1-score points.

5.3 DC Task

Results for the DC task are presented in Figures 10 and 11. Similarly to the previous task, KNN obtained worse results than ResNet. With respect to onset perturbation, KNN obtained consistent and progressive performance reductions with most perturbation ranges with the exception of 5ms perturbations (on all but the smallest window) where the F1-score is virtually unchanged from the relative non-perturbed results.

ResNet presented consistently better results, but the reduction of performances follows a similar trend to KNN. The performance of ResNet is increased by using larger analysis windows, more evidently than the KNN. The best results are with the 100ms window and no perturbation, reaching 0.88 F1-Score points.

5.4 TC Task

Results for the TC task are presented in Figures 12 and 13. The multiclass technique problem proves to be more challenging than the previous PD, BC and DC tasks [31], only reaching 0.3 F1-score points with KNN. ResNet reaches instead a much higher 0.76 score with the largest window (100ms) and no perturbation.

With respect to onset perturbation KNN obtained a progressive decrease in performance with the two smallest windows (6.25ms and 12.5ms). The performance is, however, consistently higher with 12.5ms windows. In contrast, with larger windows (25ms, 50ms, and 100ms), perturbation yielded slight performance gains (at most < 0.02). The less pronounced effect of the perturbation on these results can be attributed to the already-low performance of the KNN in the non-perturbed labels.

ResNet obtained a progressive improvement in performance with larger analysis windows. In addition, for all window sizes performances always decreased when applying onset perturbation, always obtaining the best results with ground-truth onset alignment and no perturbation. The performance of ResNet without perturbation is similar between 50ms and 100ms windows, but with the 100ms window, perturbation caused less performance deterioration.

6 DISCUSSION

Our experiments consistently showed that performance decreased when we applied perturbations to the onset ground-truth under strict real-time constraints (e.g., analysis window of 6.25ms and 12.5ms). In addition, in many cases, a larger perturbation range corresponded to an even greater and more progressive deterioration of algorithm performance. This decrease in performance can be attributed to the sensitivity of real-time algorithms to the alignment of their small analysis windows to real onsets.

[13] https://www.music-ir.org/mirex/wiki/2021:Audio_Onset_Detection

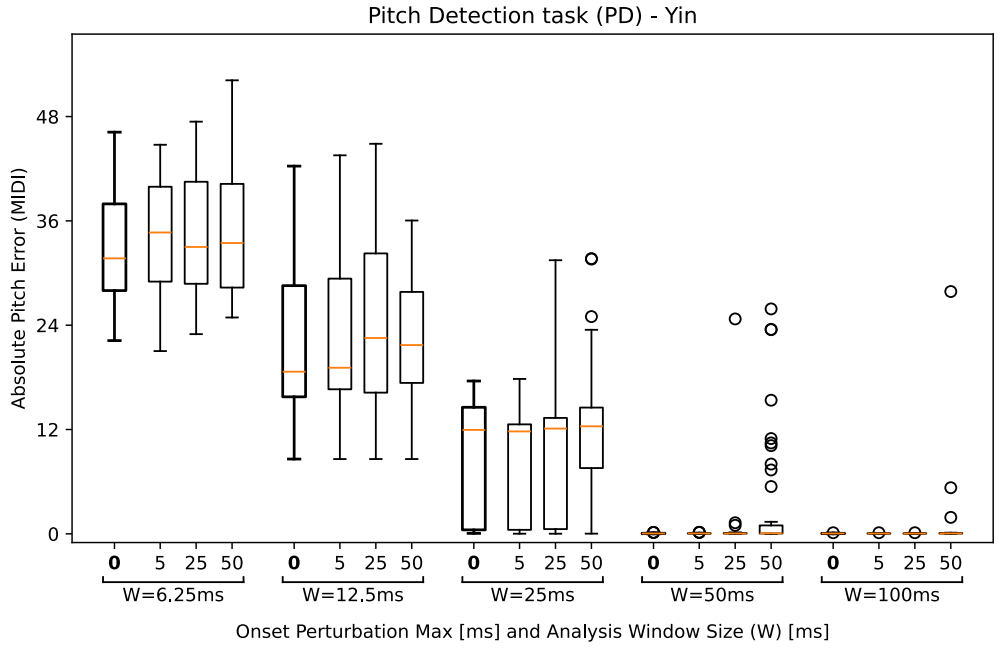

Figure 6: Results for the PD task (Pitch Detection) with the Yin pitch tracker. Absolute pitch error is measured as the distance between the ground-truth pitch and the output of the detector (in MIDI notes). Five groups of boxes are presented along the horizontal axis, where each represents the results of the detector with a specific analysis window size, ranging from 6.25 to 100ms. The first box of each group represents the result with the analysis window aligned with the dataset's ground-truth onsets, while subsequent boxes present results obtained with onset perturbation of increasing magnitude (±5, ±25, ±50 milliseconds).

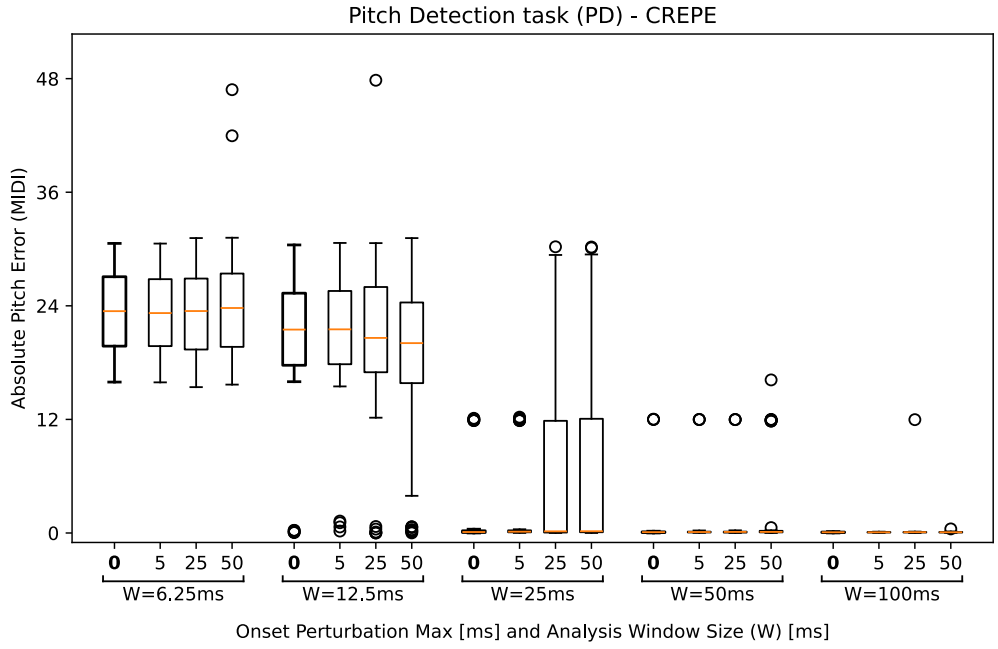

Figure 7: Results for the PD task (Pitch Detection) with the CREPE pitch tracker. Similar to Figure 6, the vertical axis represents the absolute pitch detection error as MIDI note distance, while results are horizontally grouped by analysis window size, each group containing ground-truth-aligned results and onset-perturbed detection results with increasing perturbation magnitude.

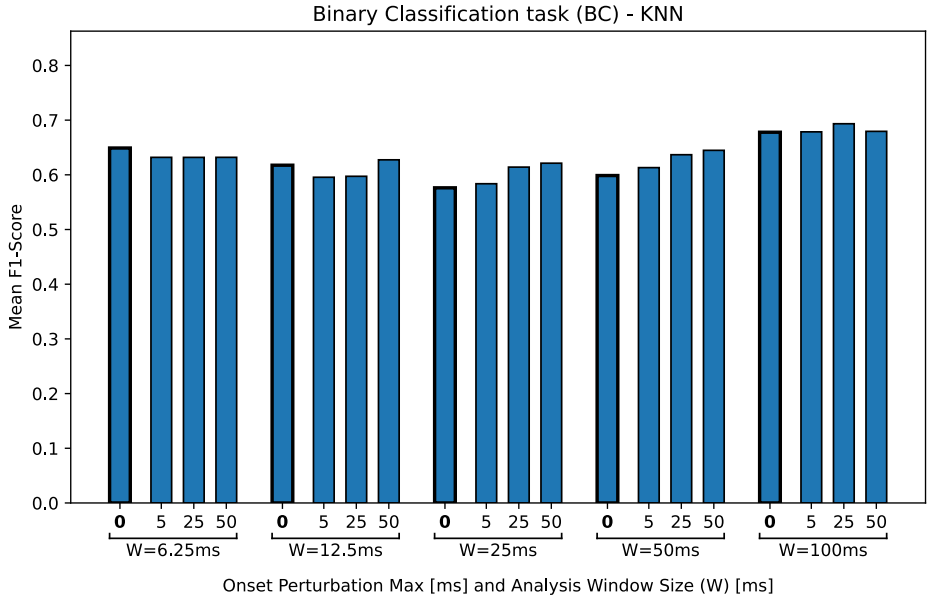

Figure 8: Results for the BC task (Binary percussive/pitched classification) with a KNN approach. The vertical axis represents the mean of the macro average F1-Score across 3-fold cross-validation, while results are horizontally grouped by analysis window size, each group containing ground-truth-aligned results (bold) and onset-perturbed detection results with increasing perturbation magnitude.

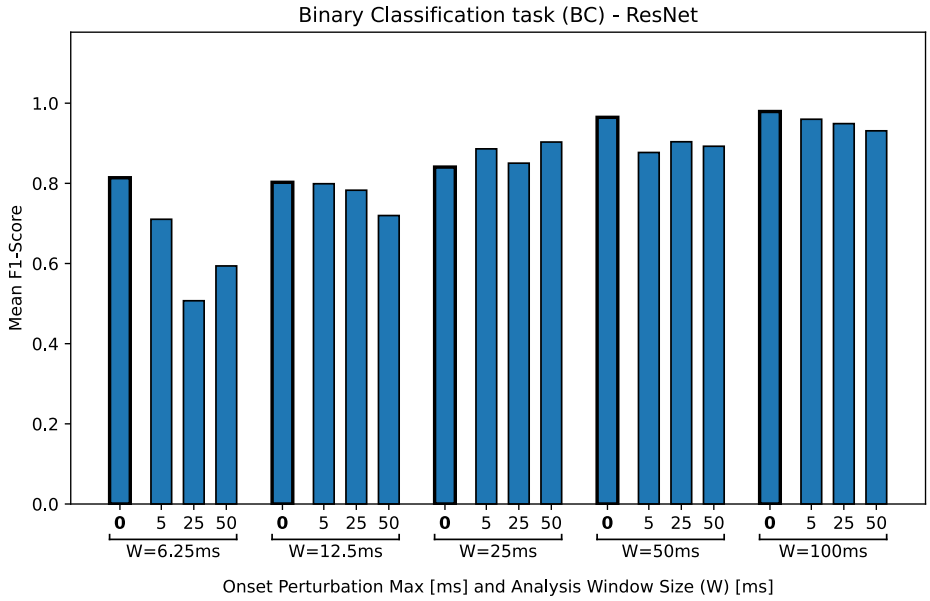

Figure 9: Results for the BC task (Binary percussive/pitched classification) with a ResNet classifier.

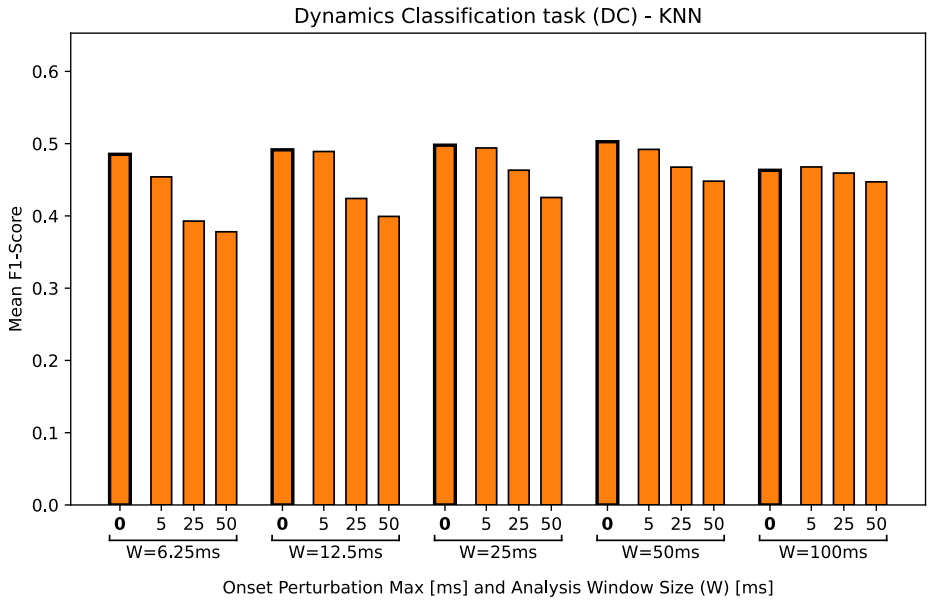

Figure 10: Results for the DC task (Dynamics classification) with a KNN classifier.

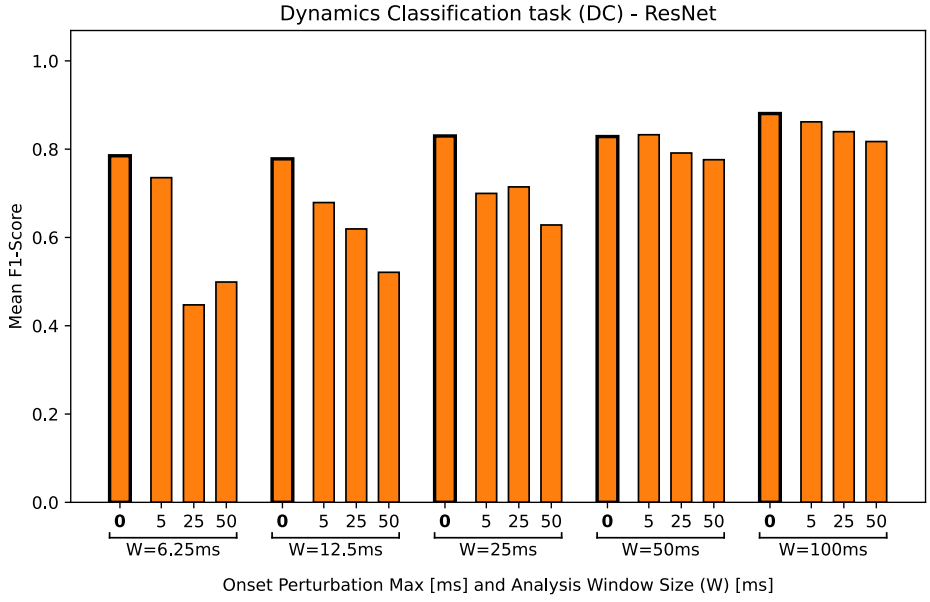

Figure 11: Results for the DC task (Dynamics classification) with a ResNet classifier.

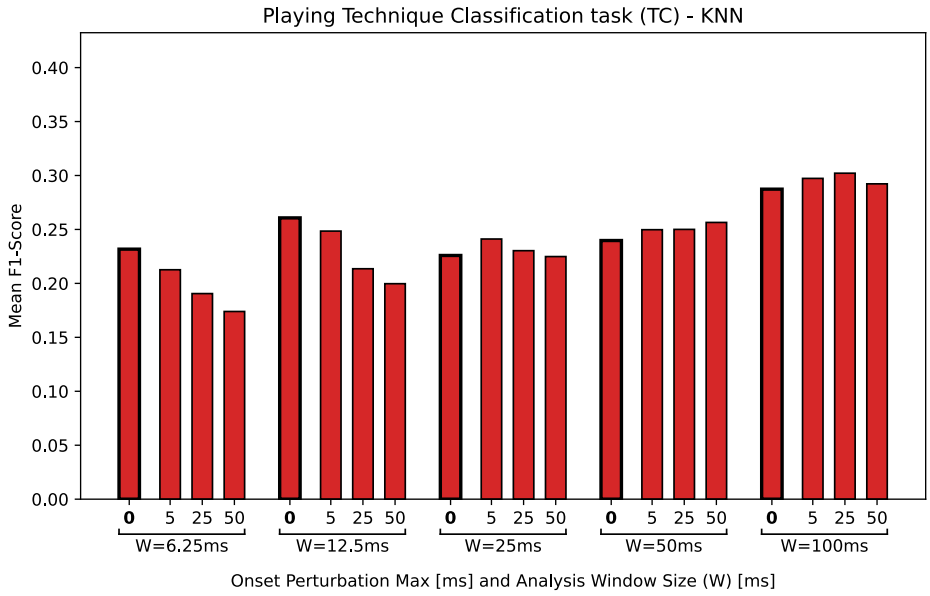

Figure 12: Results for the TC task (Multiclass technique classification) with a KNN classifier.

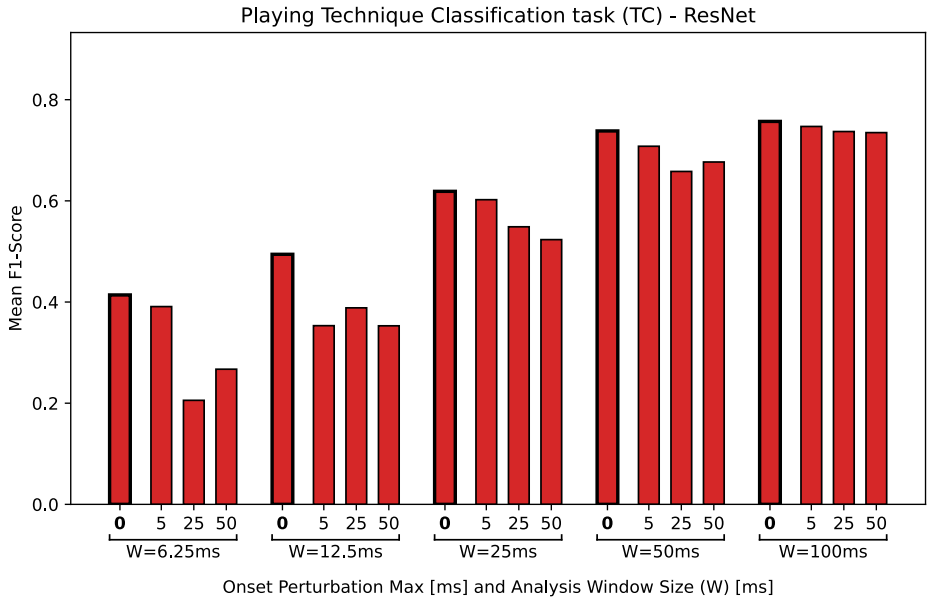

Figure 13: Results for the TC task (Multiclass technique classification) with a ResNet classifier.

In this context, even minor inaccuracies in onset timing can lead to significant errors across different tasks. These challenges were only exaggerated with the introduction of perturbations to simulate time-inaccuracies in the detection process or labeling.

Interestingly, results also reveal a few scenarios where perturbations improve performance. This seemingly paradoxical result can be explained by the similarity of our perturbation with some data augmentation approaches. By introducing controlled perturbations, we actually increased the diversity of the training dataset, exposing the system to a wider range of window alignments and inaccuracies that it might encounter in real-world conditions. This process may improve the robustness and generalization of some real-time MIR methods, enabling them to perform better under non-ideal conditions. It is important, however, that this behavior is never observed with smaller analysis windows, hinting at the fact that *severe* shifts in onset labels, and in turn window alignment, cannot serve as augmentations as they have a detrimental effect on performance. Moreover, augmentation by means of perturbation of the alignment of analysis windows can be applied artificially to precisely-annotated datasets, and the resulting performances can be evaluated, while precise alignment cannot be obtained from inaccurate annotations.

The results presented in this paper are relative to acoustic guitar sounds, and could potentially be extended to electric guitar and some plucked instruments. The applicability of our findings to different instruments, e.g., those that produce longer attack phases, would need to be confirmed by repeating the experimental tasks. The code required to repeat all the experiments is available online[14].

7 CONCLUSIONS

In this paper, we presented AG-PT-set, a novel dataset of individual acoustic guitar sounds with playing technique labels and precise onset annotations.

Furthermore, we set out to assess the importance of time-accurate onset labels for real-time Music Information Retrieval (rtMIR) tasks, which involve the alignment of small signal analysis windows with note onsets. Onset label perturbations mimicked the probability distribution of inaccurate onset detectors and served as a proxy for such. Our exploration of the effects of inaccurate onset times provided clear evidence of the need for high precision in data annotation, which in turn affects the time precision of trained onset detectors.

Future efforts will be devoted to annotating the additional recordings in AG-PT-set, to offer a wider range of playing techniques with onset labels to support automating technique recognition. Other future works could focus on applying explainability analysis methods to gain better insights into the effect of misaligned windows on machine and deep learning methods. Moreover, this study focused solely on acoustic guitar sounds, while future works could extend this analysis to more audio datasets involving different musical instruments, providing a solid benchmark for the quality and time precision of onset annotations.

We hope that our work can inspire a more in-depth analysis of the precision of onset annotations in existing and new musical audio datasets.

[14]https://github.com/CIMIL/AG-PT-set_AM24_accompanying-material

ACKNOWLEDGMENTS

Funded by the European Union under NextGenerationEU. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or The European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. We thank all the dataset annotators for volunteering their time.

REFERENCES

[1] Juan Pablo Bello, Laurent Daudet, Samer Abdallah, Chris Duxbury, Mike Davies, and Mark B Sandler. 2005. A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing* 13 (2005), 1035–1047.

[2] Juan Pablo Bello, Chris Duxbury, Mike Davies, and Mark Sandler. 2004. On the use of phase and energy for musical onset detection in the complex domain. *IEEE Signal Processing Letters* 11, 6 (2004), 553–556.

[3] Juan Pablo Bello and Mark Sandler. 2003. Phase-based note onset detection for music signals. *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)* 5 (2003), V–441.

[4] Sebastian Böck, Florian Krebs, and Markus Schedl. 2012. Evaluating the Online Capabilities of Onset Detection Methods. In *ISMIR*. 49–54.

[5] Sebastian Böck and Markus Schedl. 2011. Enhanced beat tracking with context-aware neural networks. In *Proceedings of the 14th International Conference on Digital Audio Effects (DAFx-11)*. 135–139.

[6] Sebastian Böck and Markus Schedl. 2012. Polyphonic piano note transcription with recurrent neural networks. In *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 121–124.

[7] Sebastian Böck and Gerhard Widmer. 2013. Maximum filter vibrato suppression for onset detection. In *Proc. of the 16th Int. Conf. on Digital Audio Effects (DAFx). Maynooth, Ireland (Sept 2013)*, Vol. 7. 4.

[8] Paul M Brossier. 2006. *Automatic annotation of musical audio for interactive applications*. Ph. D. Dissertation. Centre for Digital Music, Queen Mary University of London, London, UK.

[9] Paul M. Brossier. accessed July 23, 2024. Aubio, a library for audio labelling. (accessed July 23, 2024). http://aubio.piem.org.

[10] Yu-Hua Chen, Wen-Yi Hsiao, Tsu-Kuang Hsieh, Jyh-Shing Roger Jang, and Yi-Hsuan Yang. 2022. Towards Automatic Transcription of Polyphonic Electric Guitar Music: A New Dataset and a Multi-Loss Transformer Model. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 786–790. https://doi.org/10.1109/ICASSP43922.2022.9747697

[11] Alain de Cheveigné and Hideki Kawahara. 2002. YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America* 111 (2002), 1917–1930.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[13] Simon Dixon. 2006. Onset detection revisited. In *Proceedings of the 9th International Conference on Digital Audio Effects (DAFx-06)*, Vol. 120. 133–137.

[14] Chris Duxbury, Juan Pablo Bello, Mike Davies, and Mark Sandler. 2003. Complex domain onset detection for musical signals. In *Proceedings of the 6th International Conference on Digital Audio Effects (DAFx03)*, Vol. 1. 6–9.

[15] Florian Eyben, Sebastian Böck, Björn Schuller, and Alex Graves. 2010. Universal onset detection with bidirectional long-short term memory neural networks. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR), Utrecht, The Netherlands*. 589–594.

[16] Jonathan Foote and Shingo Uchihashi. 2001. The beat spectrum: a new approach to rhythm analysis. *IEEE International Conference on Multimedia and Expo (ICME)* (2001), 881–884.

[17] R. Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. 2020. Garbage in, garbage out? do machine learning application papers in social computing report where human-labeled training data comes from?. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 325–336. https://doi.org/10.1145/3351095.3372862

[18] Yuan Gong, Yu-An Chung, and James Glass. 2021. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778* (2021).

[19] Stephen Hainsworth and Malcolm D Macleod. 2003. Onset Detection in Musical Audio Signals. In *Proceedings of the International Computer Music Conference (ICMC)*.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[21] Tung-Sheng Huang, Ping-Chung Yu, and Li Su. 2023. Note and Playing Technique Transcription of Electric Guitar Solos in Real-World Music Performance. In

ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 1–5. https://doi.org/10.1109/ICASSP49357.2023.10095225

[22] Christian Kehling, Jakob Abeßer, Christian Dittmar, and Gerald Schuller. 2014. Automatic Tablature Transcription of Electric Guitar Recordings by Estimation of Score-and Instrument-Related Parameters. In *DAFx*. 219–226.

[23] Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello. 2018. Crepe: A Convolutional Representation for Pitch Estimation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 161–165. https://doi.org/10.1109/ICASSP.2018.8461329

[24] Khaled Koutini, Jan Schlüter, Hamid Eghbal-Zadeh, and Gerhard Widmer. 2021. Efficient training of audio transformers with patchout. *arXiv preprint arXiv:2110.05069* (2021).

[25] Andrea Martelloni, Andrew McPherson, and Mathieu Barthet. 2020. Percussive Fingerstyle Guitar through the Lens of NIME: an Interview Study. In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*. 440–445. https://doi.org/10.5281/zenodo.4813463

[26] Paul Masri. 1996. *Computer modeling of Sound for Transformation and Synthesis of Musical Signal*. Ph. D. Dissertation. University of Bristol, UK.

[27] Jan Schlüter and Sebastian Böck. 2013. Musical onset detection with convolutional neural networks. In *6th international workshop on machine learning and music (MML), Prague, Czech Republic*. sn.

[28] Jan Schlüter and Sebastian Böck. 2014. Improved musical onset detection with Convolutional Neural Networks. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 6979–6983. https://doi.org/10.1109/ICASSP.2014.6854953

[29] Siddharth Sigtia, Emmanouil Benetos, and Simon Dixon. 2016. An End-to-End Neural Network for Polyphonic Piano Music Transcription. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24, 5 (2016), 927–939. https://doi.org/10.1109/TASLP.2016.2533858

[30] Domenico Stefani and Luca Turchet. 2021. Bio-Inspired Optimization of Parametric Onset Detectors. In *Proceedings of the 24th International Conference on Digital Audio Effects (DAFx20in21)* (Vienna, Austria), Vol. 2. 268–275. https://doi.org/10.23919/DAFx51585.2021.9768293

[31] Domenico Stefani and Luca Turchet. 2022. On the Challenges of Embedded Real-Time Music Information Retrieval. In *Proceedings of the 25-th Int. Conf. on Digital Audio Effects (DAFx20in22)* (Vienna, Austria), Vol. 3. 177–184.

[32] Dan Stowell and Mark Plumbley. 2007. Adaptive whitening for improved real-time audio onset detection. In *Proceedings of the 2007 International Computer Music Conference, ICMC 2007*. 312–319.

[33] Li Su, Li-Fan Yu, and Yi-Hsuan Yang. 2014. Sparse Cepstral, Phase Codes for Guitar Playing Technique Classification. In *ISMIR*. 9–14.

[34] L. Turchet. 2018. Hard real time onset detection for percussive sounds. In *Proceedings of the Digital Audio Effects Conference*. 349–356.

[35] L. Turchet. 2019. Smart Musical Instruments: vision, design principles, and future directions. *IEEE Access* 7 (2019), 8944–8963.

[36] Qingyang Xi, Rachel M Bittner, Johan Pauwels, Xuzhou Ye, and Juan Pablo Bello. 2018. GuitarSet: A Dataset for Guitar Transcription. In *ISMIR*. 453–460.