Improving Sound Event Detection Metrics: Insights from DCASE 2020

Giacomo Ferroni, Nicolas Turpault, Juan Azcarreta, Francesco Tuveri, Romain Serizel, Çağdaş Bilen, Sacha Krstulović

To cite this version:

HAL Id: hal-02978422
https://inria.hal.science/hal-02978422
Submitted on 26 Oct 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
ABSTRACT

The ranking of sound event detection (SED) systems may be biased by assumptions inherent to evaluation criteria and to the choice of an operating point. This paper compares conventional event-based and segment-based criteria against the Polyphonic Sound Detection Score (PSDS)’s intersection-based criterion, over a selection of systems from DCASE 2020 Challenge Task 4. It shows that, by relying on collars, the conventional event-based criterion introduces different strictness levels depending on the length of the sound events, and that the segment-based criterion may lack precision and be application dependent. Alternatively, PSDS’s intersection-based criterion overcomes the dependency of the evaluation on sound event duration and provides robustness to labelling subjectivity, by allowing valid detections of interrupted events. Furthermore, PSDS enhances the comparison of SED systems by measuring sound event modelling performance independently from the systems’ operating points.

Index Terms— sound event detection, evaluation metrics, segment vs event criteria, polyphonic sound detection score.

1. INTRODUCTION

Given an audio recording, sound event detection consists not only of detecting which sound events have occurred during the recording but also when it occurred [1]. SED and the broader domain of ambient sounds analysis have received a lot of attention in the past few years because of their numerous potential applications including smart home consumer electronics, smartphones, earbuds, assisted living, smart cities or security [2–8].

A common way to solve the SED challenge is to use supervised machine learning on a training dataset composed of strongly labeled recordings, i.e., with timestamps for the sound events’ onset and offset. However, obtaining these annotations is time and resource consuming, as well as prone to errors and disagreement between annotators in their perception of where the onsets and offsets should be, thus making strong annotations difficult to obtain from crowd sourcing. Since 2018, Task 4 of the international challenge on Detection and Classification of Acoustic Scenes and Events (DCASE) has been addressing this issue by using weakly annotated training data [9] and synthetic soundscapes with strong annotations [10], which are much cheaper to obtain. However, the evaluation set remains strongly annotated by humans, with potential inconsistencies in annotations due to the aforementioned disagreements and ambiguities, whereas the systems are evaluated on a segmentation task, i.e., in terms of their capacity to localise a sound event in time in addition to predicting the correct sound event classes. Therefore, evaluation metrics should be able to measure the time-localisation abilities of the compared systems at different time scales while remaining independent of the labeling strategy. With that in mind, and until recently, SED systems were evaluated on two different classes of metrics summarised into F1-scores [11, 12]: event-based metrics, which measure how successful the models are in correctly predicting each annotated sound event, or segment-based metrics, which measure how successful the models are in correctly identifying the most dominant sound event within fixed sized segments of audio.

In 2020, DCASE Challenge Task 4 introduced the Polyphonic Sound Detection Score (PSDS) [13] as an alternative evaluation metric [14, 15]. PSDS relies on an intersection criterion to validate the predicted sound events against the reference annotations. It can be adjusted to enforce similar level of constraints to segment or event based evaluations and it can provide the advantages of both approaches (Table 1). PSDS also provides an alternative metric to the F1-score insofar as it is a performance figure that is global across a set of operating points, rather than depending on a single operating point chosen a priori for each submitted system as implied by

UL/INRIA’s work for this article was partly supported by the French National Research Agency (project LEAUDS “Learning to understand audio scenes” ANR-18-CE23-0020) and by the French region Grand-Est.
the F1-score. Therefore PSDS prevents the erroneous conclusion that a system performs sound event modelling better than another one where, instead, the difference relies only on a different choice of system sensitivity tuning.

This paper therefore presents the experimental comparison, over PSDS compliant submissions to DCASE 2020 Challenge Task 4, of the intersection-based metric versus the conventional event or segment-based metrics, with particular regards to flexibility, robustness to annotation variations and independence from the operating point.

Section 2 describes the compared SED evaluation metrics and the analysis methodology used in this paper. Section 3 and Section 4 report the experimental results which outline the relevance of the choice of evaluation criterion and the independence of the PSDS score from the operating point, respectively. Finally, the conclusions of the paper are summarised in Section 5.

2. THE COMPARED SED EVALUATION CRITERIA

Several metrics have been proposed for the evaluation of polyphonic SED systems [12]. This section presents the metrics that have been used in DCASE Challenge Task 4 during the past 4 years as well as the metric recently introduced and presented in [13]. In particular, this paper focuses on two aspects of importance for the evaluation of SED systems: different possible ways to validate the predicted time localisation of a sound event, either at a segment level or at an event level, and different approaches available to compute an overall performance metric from the validated event-wise or segment-wise decisions.

<table>
<thead>
<tr>
<th>Approaches</th>
<th>SB</th>
<th>CB</th>
<th>IB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allows interrupted detections</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Allows merged detections</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Each event has an equal weight in the metric</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Metric parameters can be chosen independently of event durations</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
</tbody>
</table>

Table 1. Compared characteristics of the segment-based (SB), collar-based (CB) and intersection-based (IB) approaches.

2.1. Event validation criteria

Time localisation can be determined either by comparing each predicted sound event with the reference annotations in the case of event-based metrics, or by determining the presence of a sound event within a segment of fixed length in the case of segmented-based metrics. Although intersections and collars are both event-based criteria, in this article the two approaches are distinguished as intersection-based and collar-based for the sake of comparison.

**Collar-based criterion**: Traditional event-based metrics match predicted event occurrences with reference events in the transcription up to a tolerance collar around the onset and the offset. These collars generally have a fixed size for the onset, e.g., 200ms in DCASE Challenge Task 4. The size of the collar on the offset can depend on the duration of the sound event, e.g., the maximum between 200ms and 20% of the duration of the sound event in DCASE Challenge Task 4. As a consequence, onset detection can have a large impact on event-based metrics. For short sound events, typically below 1 second in length, it could seem fair to impose a constraint of 200ms on onset estimation. However, for longer sound events the impact of this collar might become disproportionate.

**Segment-based criterion**: Segment-based metrics are computed by comparing the system outputs against the reference across segments of fixed duration. This approach is more tolerant than event-based metrics to the occurrence of short pauses between consecutive sound events as well as to errors in the detection of sound event boundaries [16]. The segment-based criterion requires to define a fixed segment size, e.g., 1 second for DCASE Challenge Task 4. This value is application-dependent and has to be chosen carefully in order to avoid performance measure artifacts. Indeed, multiple short sound events (dog barks, dishes, etc.) present in the same segment may be artificially collapsed into a single target event, thus leading to an overestimation of performance. Conversely, long sound events (vacuum cleaner, blender, etc.) may be artificially subdivided into sequences of sound events, one per segment, thus possibly underestimating performance and overemphasising the total contribution of long event detections to the final performance figure.

**Intersection-based criterion**: The intersection criterion is a novel approach, formally presented in [13]. It redefines true positives (TP), false positives (FP) and cross-triggers (CT) in the context of polyphonic SED by exploiting the amount of overlap between the annotations and the detections. Thus, it allows the evaluation to support merged detections, i.e., single detections across multiple annotations, and interrupted ones, i.e., multiple detections across single annotations. Furthermore, the evaluation is invariant to the time scale of the event durations. These properties provide the flexibility needed to handle subjective labelling and to avoid event duration biases, i.e., events with certain durations have a higher weight in the evaluation. The minimum amount of overlap required to validate the detections against the annotations is parameterised by three intersection ratios: the Detection Tolerance Criterion (DTC), the Ground Truth intersection Criterion (GTC) and the Cross-Trigger Tolerance Criterion (CTTC) [13].

A summary of the main differences between the three event validation criteria is shown in Table 1.
2.2. Overall performance metrics

Once a criterion has been chosen to determine the validity of the time localisation for each predicted sound event, a metric needs to be computed from the predictions of the system at the corresponding time scale.

**F1-score**: SED systems are conventionally evaluated with an event error rate (ER) or an F1-score [12]. The computation of these metrics relies on the evaluation of TP and FP rates computed from system predictions made at a single operating point. The issue about computing metrics on a single operating point will be addressed in Section 4.

**ROC-curves/PSDS**: Receiver Operating Characteristic (ROC) curves are computed over ranges of operating points. As such, the Area Under ROC Curves (AUC) [17] summarises global performance information without depending on the choice of a single operation point. However, the AUC is traditionally used for binary classification. PSDS defines a similar metric, but for polyphonic SED computed from a ROC-like curve [13].

### 2.3. Analysis methodology

In this analysis, we used the PSDS compliant systems submitted to the DCASE 2020 Challenge Task 4. In total, eight teams were selected for the analysis\(^1\), and only the best submission for each team in terms of PSDS score computed with \(\alpha _{CT} = \alpha _{ST} = 0\), \(\text{DTC} = \text{GTC} = 0.5\), \(\text{CTTC} = 0.3\) and \(\epsilon _{\text{max}} = 100\) events-per-hour [13]. The dataset used throughout the comparison is the domestic environment sound event detection (DESED) public evaluation dataset [10, 18]. It is entirely composed of strongly annotated, real soundscapes and it comprises the 10 different sound event classes shown in Table 2.

The predictions from the selected systems are evaluated using the metrics presented in Section 2 with 200ms and \(\max (200\text{ms}, 0.2 \times \text{annotation\_length})\) as onset and offset collars, respectively. The segment size for segment-based evaluation is set to 1 second [19]. The intersection-based evaluation uses \(\text{DTC} = \text{GTC}\) values ranging from 0.1 to 0.9 with a step of 0.2 while the CTTC threshold is not considered since \(\alpha _{CT}\) is set to 0. The analysis presented in the next two sections aims to study how the different evaluation criteria compare to each other, and how evaluating systems independently from the operating point keeps the evaluation focused on sound event modelling.

### 3. COMPARING EVENT VALIDATION CRITERIA

Throughout this section the metric is set to macro F1-score, computed as the average of the per-class F1-scores, and is used to compare the different event validation criteria.

\(^1\)The evaluation data presented in this paper have been computed by the Task 4 organisers (UL/INRIA) with code provided by Audio Analytic, in order to preserve the confidentiality of the challenge’s prediction data.

---

![Fig. 1. Macro F1-score across systems for the compared evaluation criteria: segment-based (SB), intersection-based (IB-DTC=GTC), collar-based (CB) and collar onset only (CB-O).](image)

<table>
<thead>
<tr>
<th>Class</th>
<th>(\mu \pm \sigma), m</th>
<th>Class</th>
<th>(\mu \pm \sigma), m</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dishes</td>
<td>0.6±0.7, 0.33</td>
<td>El. shaver</td>
<td>5.4±2.1, 5.76</td>
</tr>
<tr>
<td>Dog</td>
<td>0.8±3.9, 5.96</td>
<td>Blender</td>
<td>0.33</td>
</tr>
<tr>
<td>Cat</td>
<td>1.1±1.1, 0.74</td>
<td>Run. water</td>
<td>6.4±1.1, 7.85</td>
</tr>
<tr>
<td>Speech</td>
<td>1.4±2.2, 1.04</td>
<td>Frying</td>
<td>8.7±3.6, 10.00</td>
</tr>
<tr>
<td>Alarm</td>
<td>3.2±9.2, 0.83</td>
<td>Vacuum</td>
<td>8.3±0.6, 9.96</td>
</tr>
</tbody>
</table>

Table 2. Class names and label duration statistics: mean ± standard deviation and median in seconds.
gle detections over multiple ground truths, a trade-off that the conventional evaluation criteria are not able to achieve.

Furthermore, collars introduce a dependency on the annotation lengths: event durations comparable to the collar size imply more tolerance on onset detection, e.g., in the case of 500ms-long events with a 200ms collar. Comparatively, the same tolerance applied to 8-seconds-long events is stricter because it ignores a lot of the matching information, in particular the possibility for valid detections to occur outside of the collar boundaries but within the whole event duration.

Three observations confirm this from Figure 2, which shows the F1-scores per-class averaged across the eight systems for the compared evaluation criteria, in growing order of class duration from Table 2: (1) IB-0.9 always yields lower F1-scores than at IB-0.1 (blue vs red), i.e., the strictness of the validation of detections is proportional to the required amount of overlap. However, IB-0.9 always yields higher F1-scores on long event classes than CB, and vice versa for short event classes, (blue vs orange) thus confirming that the collars are imposing a stricter validity constraint, perhaps unfairly, on long classes than on short classes. (2) CB always yields lower F1-scores than CB-O (orange vs green). This is expected, since the former case is a more relaxed constraint because offset detection is not required. However, the performance delta is lower for long events than for short events: offset validation has less influence on the perceived performance for long events than for short ones. (3) IB-0.9 versus CB-O (blue vs green) yields very similar F1-scores for long events only, meaning that validating the onset for long events is a difficult task, i.e., as difficult as generating detections that almost fully overlap with the corresponding annotations.

4. INDEPENDENCE FROM OPERATING POINT

The F1-score evaluation requires participants to tune thresholds that optimise the compared systems for a specific cost of false positives and missed detections [20]. This process can have a considerable impact on system ranking, as illustrated in Figure 3. For example, when selecting a system for a low FP use-case (leftmost ranking), system T4 would be preferred to system T3. However, in a scenario where FPs have a lower cost (rightmost ranking), T3 would be a better candidate model than T4. Scientific evaluation tasks such as DCASE 2020 Challenge Task 4 are interested in finding the best modelling method in general, regardless of user-related specifications of error costs and trade-offs.

5. CONCLUSIONS

This paper analyses the influence of evaluation metrics in comparing the performance of SED systems in the context of DCASE 2020 Challenge Task 4, and considers two factors leading to comparison bias: the event validation criterion and the dependency on the operating point. Sound event validation based on the collar-based criterion treats events of different duration with different levels of strictness. Alternatively, the segment-based criterion appears more tolerant to localisation errors but it lacks precision and may still be application dependent because the chosen time resolution may overemphasise the contribution of particular event lengths to the final performance figure, while de-emphasising others. A solution is shown to be the intersection-based criterion, which adapts to the audio event duration by design, does not focus on annotation boundaries nor on defining a fixed time resolution for the evaluation, and provides robustness to labelling subjectivity by allowing valid detections across interruptions. Once an event validation criterion is chosen,
system rankings based on the F1-score are still locally biased by multiple choices of operating point: PSDS overcomes this dependency. Overall, this paper suggests that PSDS, powered by the intersection-based criterion, captures a more direct relationship between performance improvements and sound event modeling methods than the conventional evaluation metrics.

6. REFERENCES


