Sound Event Bounding Boxes

Ebbers, Janek; Germain, François G; Wichern, Gordon; Le Roux, Jonathan

TR2024-118 September 04, 2024

Abstract

Sound event detection is the task of recognizing sounds and determining their extent (onset/offset times) within an audio clip. Existing systems commonly predict sound presence posteriors in short time frames. Then, thresholding produces binary frame-level presence decisions, with the extent of individual events determined by merging presence in consecutive frames. In this paper, we show that frame-level thresholding deteriorates event extent prediction by coupling it with the system's sound presence confidence. We propose to decouple the prediction of event extent and confidence by introducing sound event bounding boxes (SEBBs), which format each sound event prediction as a combination of a class type, extent, and overall confidence. We also propose a change-detection-based algorithm to convert frame-level posteriors into SEBBs. We find the algorithm significantly improves the performance of DCASE 2023 Challenge systems, boosting the state of the art from .644 to .686 PSDS1.

Interspeech 2024

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

Sound Event Bounding Boxes

Janek Ebbers, François G. Germain, Gordon Wichern, Jonathan Le Roux

Mitsubishi Electric Research Laboratories, Cambridge, MA, USA
{ebbers,germain,wichern,leroux}@merl.com

Abstract

Sound event detection is the task of recognizing sounds and determining their extent (onset/offset times) within an audio clip. Existing systems commonly predict sound presence posteriors in short time frames. Then, thresholding produces binary frame-level presence decisions, with the extent of individual events determined by merging presence in consecutive frames. In this paper, we show that frame-level thresholding deteriorates event extent prediction by coupling it with the system's sound presence confidence. We propose to decouple the prediction of event extent and confidence by introducing sound event bounding boxes (SEBBs), which format each sound event prediction as a combination of a class type, extent, and overall confidence. We also propose a change-detection-based algorithm to convert frame-level posteriors into SEBBs. We find the algorithm significantly improves the performance of DCASE 2023 Challenge systems, boosting the state of the art from .644 to .686 PSDS1.

Index Terms: sound event detection, polyphonic sound detection, post processing, change detection

1. Introduction

Automatically recognizing and processing sounds in diverse environments is a highly desired technology for many applications, such as wildlife monitoring, autonomous driving, and surveillance. Different sound recognition tasks focus on parsing acoustic scenes at different levels of detail. In particular, audio tagging [1] and sound event detection (SED) [1, 2] tasks aim at exhaustively inventorying the sounds in a scene. SED differs from audio tagging by requiring identification of the temporal extent of sound events on top of their event class. In mathematical terms, it asks to detect the events e_j $(j = 1, \ldots, J)$ expressed as triplets $(c_j, t_{\text{on},j}, t_{\text{off},j})$ with c_j the event class label, and $t_{\text{on},j}$ (resp. $t_{\text{off},j}$) the event's onset (resp. offset) time.

While a few event-level models such as those proposed in [3, 4] directly output event triplet predictions $\hat{e}_j = (\hat{c}_j, \hat{t}_{\text{on},j}, \hat{t}_{\text{off},j})$, the large majority of SED systems rely on frame-level class presence detection models. As such, thresholding of the presence confidence cannot directly output event predictions, and post-processing is needed to consolidate frame-level class presence predictions into event predictions [2]. In that respect, current state-of-the-art models [5–8] overwhelmingly compute event predictions as blocks of consecutive frame-level class presence predictions (i.e., confidences falling above the aforementioned threshold). As traditionally understood in detection tasks, the threshold then controls the minimum presence confidence triggering an event detection in a binary fashion. As such, appropriate threshold value(s) can be chosen depending on application requirements, with, for example, some applications requiring high recall and others high precision. Crucially, the current approach means varying the threshold also affects the event predictions in non-trivial and, we argue, detrimental ways. For example, additional frame-level detections due to a lower threshold can change the detected onset/offset times of a predicted event, or even merge multiple predicted events into a single one. This, in turn, substantially diminishes the interpretability of current evaluation procedures.

Here, we show this behavior to be substantially suboptimal. As a remedy, we propose a new structure for SED systems to explicitly decouple the prediction mechanisms for onset/offset times and event presence, by introducing the sound event bounding box (SEBB) output format. Motivated by bounding box predictions in image object detection [9], the SEBB format corresponds to a series of event candidates where event class, onset time, and offset time predictions are complemented by a scalar presence confidence. Then, predicted events become a series of SEBBs whose presence confidences exceed a (now event-level) confidence threshold. Crucially, this threshold now intuitively controls only whether a SEBB is predicted as an event, without affecting its onset/offset times, and eliminates the undesirable behaviors observed with the current approach due to coupling of extent prediction with prediction confidence. Note that adding an event-level presence confidence to an event with fixed boundaries and class label to obtain a SEBB substantially differs from previous approaches separating boundary detection from class label prediction without considering prediction confidence at all [10].

SEBBs can be predicted in various ways, including in an end-to-end manner. However, we acknowledge that one reason for the enduring popularity of frame-level models is that multiple instance learning (MIL) techniques [11–13] allow for training without ground-truth onset and offset times (i.e., *weakly labeled training*), while end-to-end prediction usually requires strongly labeled training data [3]. In that context, we also propose a post-processing algorithm to convert the frame-level presence confidence scores into SEBBs for any frame-level system. In it, conversion relies primarily on a change-detection approach. Note that change-/slope-based algorithms can be found in prior post-processing signal chains for SED in [14]. However, these would perform SED solely based on change/slope without considering absolute confidence at all and did not ultimately improve performance. In contrast, we find that deploying our proposed post-processing, on top of unlocking the conceptual benefits of SEBBs, substantially improves performance. In particular, our post-processing boosts performance of all 13 considered systems from the recent DCASE 2023 Challenge Task 4a [15] and establishes a new state of the art. Source code is publicly available[1].

2. SED with Sound Event Bounding Boxes

2.1. Preliminaries

As stated earlier, SED systems commonly consist of a frame-level multi-label classifier followed by a post-processing to output predicted events. In mathematical terms, the classifier corresponds to the operation $\mathbf{Y} = f(\mathbf{X})$, with f denoting a prediction model, $\mathbf{X} = [\mathbf{x}_0, \ldots, \mathbf{x}_{N-1}]$ a sequence of input feature vectors $\mathbf{x}_n$ (e.g., log-mel spectrogram frames), and $\mathbf{Y} = [\mathbf{y}_0, \ldots, \mathbf{y}_{N-1}]$ a sequence of frame-level class probability vectors $\mathbf{y}_n$, with n the frame index. $y_{n,c} \in [0, 1]$ then rep-

[1] https://github.com/merlresearch/sebbs

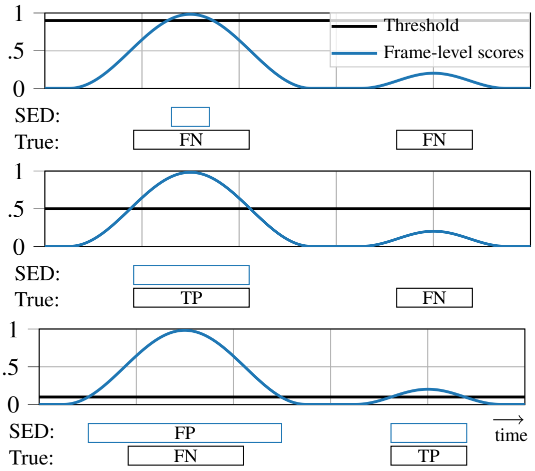

Figure 1: *Example of detection with different frame-level thresholds and comparison with ground-truth events.*

resents the predicted confidence of sound class c being present in frame n. Note that, in practice, input and output sequence lengths may differ, for example when f is a convolutional neural network with striding and/or pooling. For conciseness, we assume same sequence lengths with no loss of generality.

$\mathbf{Y}$ is then fed to post-processing. It may be first (optionally) altered, e.g., by median filtering $y_{n,c}$ in times. Ultimately, event predictions are obtained through a frame-level thresholding operation, turning $y_{n,c}$ into a binary $z_{n,c} = \mathbb{1}_{[y_{n,c}>\lambda_c]}$ (with λ_c a class-dependent threshold), followed by a merging operation where each block of consecutive $z_{n,c} = 1$ becomes a single detected event $\hat{e}_j$ of sound class c. The detected onset (resp. offset) time then corresponds to the beginning (resp. end) of the first (resp. last) frame of that block.

For applications seeking for meaningful connected event predictions, event-based evaluation is employed, which is recently favored by benchmarks and challenges. For given sets of predicted and ground truth events, counts of true positive (TP), false positive (FP), and false negative (FN) events are obtained, with two main approaches currently in use. Collar-based evaluation [16] makes determinations based on whether the onset and offset times of a predicted event match the onset and offset times of a ground-truth event of the same class up to a maximal allowed divergence. Intersection-based evaluation [17, 18] is based instead on the intersection of predicted events with ground-truth events.

Notably, as threshold selection criteria vary widely depending on the target application, the community currently relies on threshold-independent metrics which aggregate performance over various thresholds λ_c. For example, the recent DCASE Task 4 challenges used polyphonic sound detection score (PSDS) [17, 19], which is computed as the normalized area under the PSD-ROC curve, i.e., the average of class-level ROC curves from intersection-based TP/FP/FN results, plus a penalty on inter-class standard deviation.

Crucially, a known problem resulting from the aforementioned conversion of event predictions centered around frame-level confidence thresholding is that, typically, it leads to TP/FP/FN results for which the ROC curve is no longer monotonic, unlike what is expected in traditional classification. As a workaround monotonicity is restored by only taking the (oracle) best-case operating points into account [17], but it results in artificially inflated scores and limits the intuitive interpretation of the metric as a performance score.

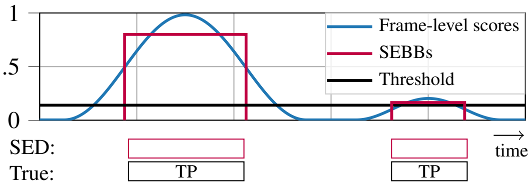

Figure 2: *Examples of SEBBs with event-level decision threshold and comparison with ground-truth events.*

2.2. Effects of Frame-level Thresholding

In this section, we show a practical illustration of the impact of frame-level thresholding on event boundary detection, and its detrimental effect on intersection-based evaluation. For this, we consider the example in Fig. 1 as representative of the frame-level presence confidence output for a single sound class found in existing SED systems. We then see how a lower frame-level threshold, while triggering more individual event detections, also has a non-trivial impact on event boundary predictions.

For example, we consider a typical intersection-based evaluation based on the ground-truth events. We use a required intersection rate of $\rho_{\text{DTC}} = \rho_{\text{GTC}} = 0.7$ for both the detection tolerance criterion (DTC) and ground-truth intersection criterion (GTC), i.e., predictions must intersect with a ground-truth event by at least 70 % to not be FP and ground-truth events must be covered by non-FP detections by at least 70 % to be TP [20].

Then, we can see that, when gradually lowering the threshold down from 1, we will first get a prediction corresponding to the first ground-truth event, but with an underestimated extent, leading to FN. When lowering the threshold further, that matching prediction remains, but its predicted extent grows longer to the point where it yields TP. However, when lowering the threshold even further, the predicted extent will ultimately grow overestimated yielding now both FN and FP, even as we might get a TP in predicting the second ground-truth event. TPs turning back to FNs (i.e., having the true positive rate decrease) when the threshold decreases is different from standard binary classification tasks and ultimately makes ROC curves decrease again after some point, breaking their monotonic properties. As we can see, this is ultimately because the threshold that detects the correct extent depends on the geometry of the frame-level scores (e.g., the overall peak heights in the case of Fig. 1). Crucially, we see that no threshold could get both ground-truth events right at the same time in our example.

This demonstrates how frame-level thresholding is sub-optimal for event detection due to the event-level entanglement of both boundary and confidence information in the frame-level scores. Therefore, we propose to decouple extent and confidence prediction as presented in the next section.

2.3. Sound Event Bounding Boxes

To solve above issues, we propose the concept of SEBBs as new SED system output format. In mathematical terms, we define SEBBs as quadruples $\hat{b}_j = (\hat{c}_j, \hat{t}_{\text{on},j}, \hat{t}_{\text{off},j}, \overline{y}_j)$ which intuitively represent sound event candidates defined by sound class $\hat{c}_j$, a fixed extent given by onset time $\hat{t}_{\text{on},j}$ and offset time $\hat{t}_{\text{off},j}$, plus an overall presence confidence score $\overline{y}_j$. Fig. 2 shows a graphical representation of SEBBs for our earlier example in Fig. 1. The key idea is that the temporal extent of sound event candidates should be determined independently from the event candidate confidence score. An event-level thresholding can then be employed to control a system's sensitivity without affecting the temporal extents of event predictions. In particular, even if the decision threshold is lowered far below a SEBB's confidence score, the temporal extent will not change. This ensures not to disturb high-confidence event detections when using low decision thresholds, such as in applications aiming for a high recall. With SEBBs, monotonically-increasing ROC curves are thus guaranteed again, and sound event candidates of high and low confidence, as in the above example, may be jointly detected correctly.

3. SEBBs from Frame-level Outputs

Now, as already stated, the vast majority of existing systems outputs frame-level multi-label presence confidence scores. As such, we now present a few post-processing approaches to enable conversion of their output to SEBBs.

Threshold-based SEBBs: A simple approach to generate SEBB predictions is akin to the threshold-based process illustrated in Fig. 1. Here too, we use median filtering plus frame-level thresholding followed by merging. But, instead of interpreting the results as event predictions, we use the resulting set of $\hat{c}_j, \hat{t}_{\text{on},j}, \hat{t}_{\text{off},j}$ together with $\overline{y}_j$, computed as the average over frame-level presence confidence y_{n,c_j} between $\hat{t}_{\text{on},j}$ and $\hat{t}_{\text{off},j}$, as predicted *tSEBB*. The purpose of the frame-level thresholds $\lambda_{c,\text{ext}}$ and median filter lengths hence is the detection of the events' extents and they are set through joint tuning on a validation set (as is commonly done for any frame-level post-processing of $\mathbf{Y}$ [21,22]). A second threshold (now event-level) would be used at inference to turn tSEBBs into predicted events. However, Fig. 1, where the two events cannot be detected with the same threshold, hints at how tSEBBs could still lead to poor detection performance in typical scenarios.

Change-detection-based SEBBs: Alternatively, we propose a change-detection-based algorithm. We first compute "delta" (i.e., change) scores by filtering $y_{n,c}$ with an ideal step filter. As different systems use different frame lengths, we perform the filtering in continuous time, interpolating $y_{n,c}$ as framewise constant. For filter length τ_c (in seconds), a delta score corresponds to the difference between the average of $y_{n,c}$ in the next $\tau_c/2$ and the previous $\tau_c/2$ seconds. Now, local maxima (resp. minima) of the delta scores become tentative onsets (resp. offsets). Tentative events (resp. gaps) are formed between each tentative onset (resp. offset) and the next tentative offset (resp. onset). Fig. 3 shows an example of frame-level scores in the upper plot, with corresponding deltas and tentative onsets (local maxima) and offsets (local minima) in the lower plot.

Then, as some tentative gaps may be due to only small spurious variations of $y_{n,c}$, we employ the following merging strategy. For every tentative gap, we compare its lowest $y_{n,c}$ with the highest $y_{n,c}$ in the tentative events immediately preceding and following it. If the comparisons fall under a predefined merge threshold γ_c, the tentative offset and onset around the gap are removed (i.e., the preceding and following tentative events and the gap between them are merged into the same event). We test γ_c both as threshold on the difference (i.e., absolute threshold) and ratio (i.e., relative threshold) between scores. Finally, we form a predicted *cSEBB* of class c from each remaining onset as $\hat{t}_{\text{on}}$, the following remaining offset as $\hat{t}_{\text{off}}$, and the average of $y_{n,c}$ between $\hat{t}_{\text{on}}$ and $\hat{t}_{\text{off}}$ as $\overline{y}$. The upper plot in Fig. 3 shows cSEBBs obtained from tentative onset/offsets (lower plot) using a relative $\gamma = 3$ (i.e., checking if neighboring tentative events' maxima fall below 3 times a gap's mimimum). The class-dependent filter length τ_c and threshold γ_c (either absolute or relative) are hyperparameters to tune on a validation set.

Hybrid SEBBs: We further propose a hybrid of the two previous approaches, where we predict a set of *hSEBB*s as follows. We first predict tSEBBs and select those above a certain confidence $\lambda_{c,\text{hyb}}$. These are then complemented by cSEBBs, discarding any cSEBB that overlaps with selected tSEBBs. Here, the predicted cSEBBs may find additional low-confidence SEBBs. The following (class-wise) hyperparameters are to be tuned on a validation set: median filter lengths and $\lambda_{c,\text{ext}}$ for tSEBB prediction, $\{\tau_c, \gamma_c\}$ for cSEBB prediction, and $\lambda_{c,\text{hyb}}$.

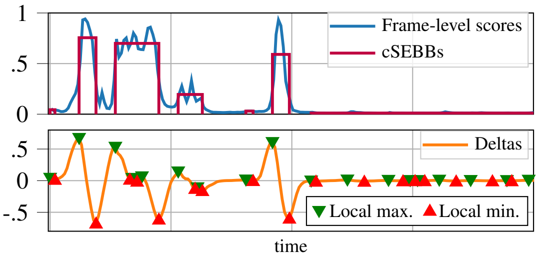

Figure 3: *Example of frame-level scores with deltas/change values from which proposed cSEBBs' on-/offset times are inferred.*

4. Experiments

We evaluate our proposed methods on DCASE 2023 Challenge Task 4a submissions [23] which target SED in domestic environments using the DESED dataset [24]. A system submission comprises three series of timestamped frame-level scores for the evaluation set generated corresponding to three (independent) system training runs. Additionally, if a proposed system included any frame-level post-processing, participants were asked to also provide "raw" frame-level scores before post-processing.

We run our evaluations on the public portion of the evaluation set for which ground-truth annotations are publicly available. To not apply our methods on top of other post-processing schemes, we only consider teams that provided raw scores, i.e., 13 teams (counting the baseline). For each team, we only report the system that performed best in terms of the challenge PSDS1 metric, i.e., Kim-2 [6], Chen-2 [25], Xiao-4 [26], Wenxin-6 [27], Li-6 [28], Cheimariotis-1 [29], Guan-3 [30], Liu-NSYSU-7 [31], Baseline-2 [32], Wang-1 [33], Lee-1 [34], Liu-SRCN-4 [35], and Barahona-2 [22].

As evaluation metrics, we use two metrics from the challenge, i.e., 1) PSDS1, i.e., PSDS with $\rho_{\text{DTC}} = \rho_{\text{GTC}} = 0.7$ and an inter-class standard-deviation penalty weight of $\alpha_{\text{ST}} = 1$, and 2) collar-based F_1-score [16], henceforth simply referred to as F_1, with a $200\,\text{ms}$ onset and max($200\,\text{ms}$, $20\,\%$ of ground truth event length) offset collar. We are not considering the PSDS2 metric, as it is more tuned as an audio tagging metric than an SED metric [36].

As previously mentioned, PSDS considers only best-case decision thresholds, i.e., thresholds leading to less TPs at more FPs than another threshold are discarded from the ROC curves. Whether an operating point is best-base or not, however, can only be determined by evaluation w.r.t. ground truth. This is then in square contradiction with any practical scenario where this oracle information would be, of course, inaccessible. Therefore, when using legacy event prediction (i.e., frame-level thresholding followed by merging), we also report *non-oracle PSDS (noPSDS)*, i.e., PSDS without best-case selection, that is, the normalized area under the (possibly non-monotonic) PSD-ROC. In order to simply limit the descent of the PSD-ROC, however, we pre-tune minimum thresholds $\lambda_{c,\text{noPSDS}}$, for each class c, which give the maximal number of TPs (#TP) on a validation dataset (i.e., below which #TP only decreases).

For each system, we report performance using legacy event prediction in conjunction with the original submission post-

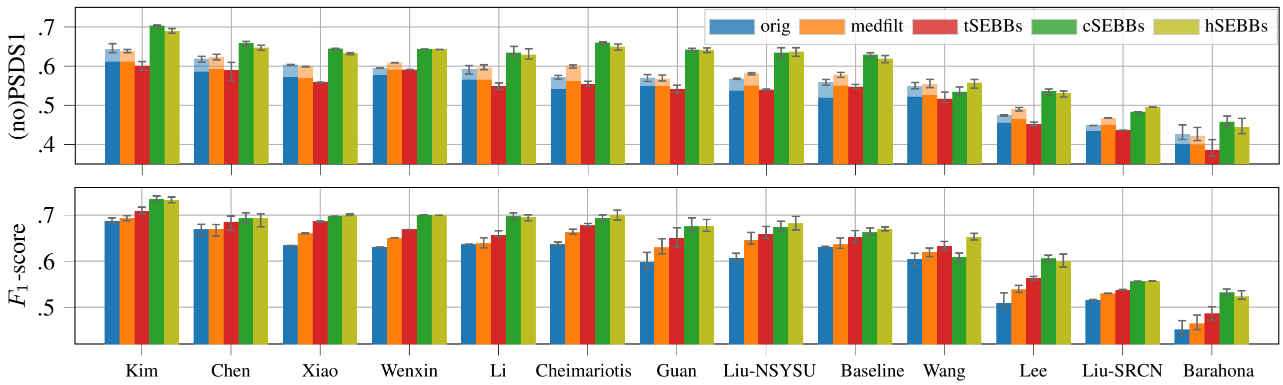

Figure 4: *Results on public evaluation set using 5-fold cross-validation for hyper-parameter tuning.*

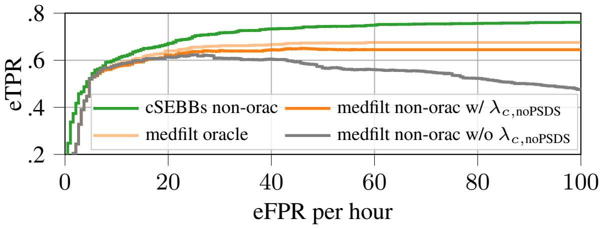

Figure 5: *PSD-ROCs for Kim with different post-processing.*

processing (orig) and common median filter post-processing (medfilt), and for SEBB-level thresholding in conjunction with our three proposed post-processing methods which output tSEBBs, cSEBBs, and hSEBBs, respectively. For each system and event class, the following hyper-parameters are to be tuned on a validation set:

- orig: $\lambda_{c,\text{noPSDS}}$,
- medfilt: median filter length, $\lambda_{c,\text{noPSDS}}$,
- tSEBBs, cSEBBs, hSEBBs: see lists in Sec. 3,

plus, for each method, a decision threshold $\lambda_{c,F}$ for F_1 evaluation. Different hyperparameter sets are tuned for PSDS and F_1 evaluation, respectively. Optimal thresholds can be efficiently tuned using sed_scores_eval [19]. Median filter lengths are chosen out of $\{0\,\text{s}\,(\text{no filter}), 0.2\,\text{s}, \dots, 2\,\text{s}\}$. (τ_c, γ_c) is chosen out of $\{0.32\,\text{s}, 0.48\,\text{s}, 0.64\,\text{s}\} \times \{.15\,\text{abs.}, .2\,\text{abs.}, .3\,\text{abs.}, 1.5\,\text{rel.}, 2\,\text{rel.}, 3\,\text{rel.}\}$. Hyperparameters are tuned to maximize the respective metric with the following exception. For hSEBBs, for simplicity, we do not tune all parameters. Instead, we adopt tSEBB-related parameters from stand-alone tSEBBs and cSEBB-related parameters from standalone cSEBBs and optimize only $\lambda_{c,\text{hyb}}$.

Without access to outputs on the challenge's validation set, we instead report noPSDS1 (using $\lambda_{c,\text{noPSDS}}$ for legacy event prediction) and F_1 for a 5-fold cross-validation on the evaluation set outputs, where predictions for each fold are generated using hyper-parameters tuned on the four other folds, in Fig. 4. For each condition, we show the mean, lowest and highest score over the three provided runs. For legacy event prediction, we also show the (higher) PSDS1 using a lighter color. Interestingly, we see that most systems would already perform better with legacy event prediction if they just traded their current post-processing for a straightforward class-specific median filter post-processing. PSDS1 performance deteriorates with tSEBBs, which suggests that indeed a fixed threshold ($\lambda_{c,\text{ext}}$) does not correctly predict the extents of events that have different detection confidences. For F_1-score, which evaluates a single operating point, it however is clearly beneficial to have different extent detection and event-level confidence thresholds. For the more sophisticated cSEBBs and hSEBBs, they overall substantially outperform the other methods. cSEBBs improves over median filtering for all systems but Wang [33], achieving an average gain of $4.1\,\%\text{pt}$ for PSDS1 and $3.4\,\%\text{pt}$ for F_1. They boost the winning system's PSDS1 from .644 to .703, and F_1-score from .688 to .734 setting the state of the art on this particular setup. At the same time, hSEBBs outscores median filtering for all systems, albeit only improving over cSEBBs for a select few. That hSEBBs deteriorate performance over cSEBBs for most systems can be explained by poorer generalization of the increased number of hyper-parameters.

In Fig. 5, we further see the expected benefits of cSEBBs on the Kim-2 system, when comparing the corresponding PSD-ROC curve with curves for the legacy event prediction with median filter post-processing. First, it can be seen how the oracle modification of PSDS distorts the intuition behind the AUC-like component of the PSDS, substantially diverging from the non-oracle curve, which is partially mitigated by the addition of $\lambda_{c,\text{noPSDS}}$. Further we can see that the inflated oracle modification still fails to close the gap with the cSEBBs' PSD-ROC curve, ending up lower in every operating range.

Finally, to evaluate our method in the proper challenge setting, i.e., tuning hyperparameters on validation set output scores, we contacted participants and asked whether they could share these[2]. We received raw validation scores from Xiao [26], Li [28], Barahona [22], and the baseline [32], and post-processed validation scores from Kim [6]. We then optimized hyperparameters on that data before scoring on the full evaluation set. Note that only having post-processed validation scores for Kim means a mismatch between validation (with original post-processing added) and evaluation (without). Our method again significantly improves performance for all systems and we achieve new challenge state-of-the-art performances of .686 PSDS (by Kim [6]) and .706 F_1 (by Xiao [26]), with Kim's performance likely hurt by the validation/evaluation mismatch.

5. Conclusions

In this work, we demonstrated how the commonly used frame-level thresholding for SED results in a harmful coupling of event extent and confidence prediction. As solution, we introduced sound event bounding boxes (SEBBs) as a new general SED output format, which also overcomes the ill-definition of recent event-based evaluation metrics. We further proposed a change-detection-based algorithm to infer SEBBs from frame-level model outputs. Our experiments showed that our proposed method allows for substantially improved performance for a large range of systems and sets a new state of the art on the DCASE 2023 Challenge Task 4a benchmark.

[2] We would like to thank all teams who responded to our request.

6. References

[1] T. Virtanen, M. D. Plumbley, and D. Ellis, *Computational analysis of sound scenes and events*. Springer, 2018.

[2] A. Mesaros, T. Heittola, T. Virtanen, and M. D. Plumbley, "Sound event detection: A tutorial," *IEEE Signal Process. Mag.*, vol. 38, no. 5, pp. 67–83, 2021.

[3] Z. Ye, X. Wang, H. Liu, Y. Qian, R. Tao, L. Yan, and K. Ouchi, "Sound event detection transformer: An event-based end-to-end model for sound event detection," *arXiv preprint arXiv:2110.02011*, 2021.

[4] S. Bhosale, S. Nag, D. Kanojia, J. Deng, and X. Zhu, "DiffSED: Sound event detection with denoising diffusion," *arXiv preprint arXiv:2308.07293*, 2023.

[5] H. Nam, S. Kim, and Y.-H. Park, "Frequency dynamic convolution: Frequency-adaptive pattern recognition for sound event detection," in *Proc. Interspeech*, 2022.

[6] J. W. Kim, S. W. Son, Y. Song, H. K. Kim, I. H. Song, and J. E. Lim, "Label filtering-based self-learning for sound event detection using frequency dynamic convolution with large kernel attention," in *Proc. DCASE*, 2023.

[7] Y. Xin, D. Yang, F. Cui, Y. Wang, and Y. Zou, "Improving weakly supervised sound event detection with causal intervention," in *Proc. ICASSP*, 2023.

[8] N. Shao, X. Li, and X. Li, "Fine-tune the pretrained atst model for sound event detection," *arXiv preprint arXiv:2309.08153*, 2023.

[9] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, real-time object detection," in *Proc. CVPR*, 2016.

[10] S. Kothinti, K. Imoto, D. Chakrabarty, G. Sell, S. Watanabe, and M. Elhilali, "Joint acoustic and class inference for weakly supervised sound event detection," in *Proc. ICASSP*, 2019.

[11] A. Kumar and B. Raj, "Audio event detection using weakly labeled data," in *Proc. ACM Multimedia*, 2016.

[12] A. Shah, A. Kumar, A. G. Hauptmann, and B. Raj, "A closer look at weak label learning for audio events," *arXiv preprint arXiv:1804.09288*, 2018.

[13] Y. Wang, J. Li, and F. Metze, "A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling," in *Proc. ICASSP*, 2019.

[14] L. Cances, P. Guyot, and T. Pellegrini, "Evaluation of post-processing algorithms for polyphonic sound event detection," in *Proc. WASPAA*, 2019.

[15] DCASE 2023 Challenge Task 4a description. [Online]. Available: https://dcase.community/challenge2023/task-sound-event-detection-with-weak-labels-and-synthetic-soundscapes

[16] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016.

[17] Ç. Bilen, G. Ferroni, F. Tuveri, J. Azcarreta, and S. Krstulović, "A framework for the robust evaluation of sound event detection," in *Proc. ICASSP*, 2020.

[18] G. Ferroni, N. Turpault, J. Azcarreta, F. Tuveri, R. Serizel, Ç. Bilen, and S. Krstulović, "Improving sound event detection metrics: insights from dcase 2020," in *Proc. ICASSP*, 2021.

[19] J. Ebbers, R. Haeb-Umbach, and R. Serizel, "Threshold independent evaluation of sound event detection scores," in *Proc. ICASSP*, 2022.

[20] F. Ronchini and R. Serizel, "A benchmark of state-of-the-art sound event detection systems evaluated on synthetic soundscapes," in *Proc. ICASSP*, 2022.

[21] J. Ebbers and R. Haeb-Umbach, "Self-trained audio tagging and sound event detection in domestic environments," in *Proc. DCASE*, 2021.

[22] S. Barahona, D. de Benito-Gorron, S. Segovia, D. Ramos, and D. T. Toledano, "Multi-resolution conformer for sound event detection: Analysis and optimization," in *Proc. DCASE*, Tampere, Finland, 2023.

[23] J. Ebbers and R. Serizel, "Submissions dcase 2023 task4a," 2023. [Online]. Available: https://doi.org/10.5281/zenodo.8248775

[24] N. Turpault, R. Serizel, A. P. Shah, and S. Justin, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Proc. ICASSP*, 2019.

[25] W.-Y. Chen, C.-L. Lu, H.-F. Chuang, Y.-H. C. Cheng, and B.-C. Chan, "Sound event detection system using pre-trained model for DCASE 2023 Task 4," DCASE 2023 Challenge, Tech. Rep., 2023.

[26] Y. Xiao, T. Khandelwal, and R. K. Das, "FMSG submission for DCASE 2023 challenge Task 4 on sound event detection with weak labels and synthetic soundscapes," DCASE 2023 Challenge, Tech. Rep., 2023.

[27] W. Duo, X. Fang, , and J. Li, "Semi-supervised sound event detection system for DCASE 2023 Task4a," DCASE 2023 Challenge, Tech. Rep., 2023.

[28] K. Li, P. Cai, and Y. Song, "Li USTC team's submission for DCASE 2023 challenge Task4a," DCASE 2023 Challenge, Tech. Rep., 2023.

[29] G.-A. Cheimariotis and N. Mitianoudis, "Sound event detection of domestic activities using frequency dynamic convolution and BEATS embeddings," DCASE 2023 Challenge, Tech. Rep., 2023.

[30] Y. Guan and Q. Shang, "Semi-supervised sound event detection system for DCASE 2023 Task 4," DCASE 2023 Challenge, Tech. Rep., 2023.

[31] C.-C. Liu, T.-H. Kuo, C.-P. Chen, C.-L. Lu, B.-C. Chan, Y.-H. Cheng, and H.-F. Chuang, "CHT+NSYSU sound event detection system with pretrained embeddings extracted from BEATS model for DCASE 2023 Task 4," DCASE 2023 Challenge, Tech. Rep., 2023.

[32] DCASE 2023 Challenge Task 4a baseline. [Online]. Available: https://github.com/DCASE-REPO/DESED_task/tree/master/recipes/DCASE2023_task4_baseline

[33] Y. Wang, H. Dinkel, Z. Yan, J. Zhang, and Y. Wang, "PEPE: Plain efficient pretrained embeddings for sound event detection," DCASE 2023 Challenge, Tech. Rep., 2023.

[34] S. Lee, N. Kim, J. Lee, C. Hwang, S. Jang, and I.-Y. Kwak, "Sound event detection using convolution attention module for DCASE 2023 challenge Task4a," DCASE 2023 Challenge, Tech. Rep., 2023.

[35] M. Chen, Y. Jin, J. Shao, Y. Liu, B. Peng, and J. Chen, "DCASE 2023 challenge Task4 technical report," DCASE 2023 Challenge, Tech. Rep., 2023.

[36] H. Nam, B.-Y. Ko, G.-T. Lee, S.-H. Kim, W.-H. Jung, S.-M. Choi, and Y.-H. Park, "Heavily augmented sound event detection utilizing weak predictions," *arXiv preprint arXiv:2107.03649*, 2021.