MULTIMODAL EVALUATION METHOD FOR SOUND EVENT DETECTION

Seyed M.R. Modaresi[1,2] *Aomar Osmani*[1] *Mohammadreza Razzazi*[2,4] *Abdelghani Chibani*[3]

[1] LIPN-UMR CNRS 7030, Sorbonne Paris Nord University, France
[2] Computer Engineering Department, Amirkabir University of Technology, Tehran, Iran
[3] Images, Signals and Intelligent Systems Laboratory, University Paris-Est Creteil, France
[4] School of Computer Science, Institute for Research in Fundamental Sciences, Tehran, Iran

ABSTRACT

Time is an important dimension in sound event detection (SED) systems. However, evaluating the performance of SED systems is directly taken from the classical machine learning domain, and they are not well adapted to the needs of these systems such as recognizing the time, duration, detection, and uniformity of sound events. Despite its importance, it is not well-developed yet. Current methods are highly biased by their assumptions and may misleadingly present convincible results. This paper presents a novel multimodal method to evaluate SED systems from multiple perspectives such as detection, total duration, relative duration, and uniformity. Furthermore, the proposed method is simple, time-efficient, visualizable, extensible, open-source, and overcomes the limitations of existing methods. The benefits of the proposed approach are demonstrated by re-evaluating the best systems presented in a known challenge on sound event detection.

Index Terms— Evaluation, Metric, Sound Event Detection, Audio Signal Processing

1. INTRODUCTION & RELATED WORK

A sound event detection (SED) system recognizes sound events in audio tracks. It is a developing research field from both academia and industry due to its potential applications in healthcare, medical telemonitoring, surveillance, smart home, monitoring, security, audio content-based searching, etc. [1, 2, 3].

The time interval included in sound events is one of the important dimension in evaluating SED systems. These systems should determine the occurrence time interval of sound events in addition to their sound classes. Moreover, sound events can happen simultaneously (e.g., opening door sound events during a speech event) [2]. Evaluating the performance of SED systems is often done by comparing their predictions with the references [2]. One of the first evlauation methods was defined in CLEAR[1] 2006 challenge named acoustic event error rate [4]. It marks a reference event as correctly identified when the temporal center of the predicted event is inside it[4]. It also defines insertion, deletion and substitutions errors. The metric was ambiguous in some cases, e.g., whenever a part of a reference is well detected, and the other part has a substitution error [5]. In CLEAR 2007 challenge, recall, precision, and f-score (considering the above definition for correct prediction) were used; however, they redefine the acoustic event error rate by using frame-based methods [6, 2, 7]. Frame-based (segment) methods take fixed-duration intervals (e.g., 10 ms) as the basic atomic unit. It facilitates comparing different algorithms since each frame is independent of both the references and predictions [8]. In DCASE[2] 2013 challenge [7], the frame-based error rate, precision, recall, f-score and collar-based method were used. In collar-based methods, a reference is considered as correctly detected if the beginning (onset) or the ending (offset) or both are within a specific tolerance (e.g., 200 ms). This tolerance is necessary because of the inexact labeling of the data [2]. PSDS[3] is a recent method for SED evaluation which is proposed for more robust defining of true positive (TP), false positive (FP), and false negative (FN) by considering the intersection rate based on references and predictions [1]. It overcomes the dependency of the evaluation on sound event's duration and provides robustness to labeling subjectivity [9]. Researchers in [10] explore inequality of missing events in different scenarios. They break down FP into related and unrelated by considering the scene and sound event relation; then, they give double penalty to unrelated FPs in calculating f-score. The IEEE AASP[4] challenge on detection and classification of acoustic scenes and events [7] highlights the need for an appropriate metric. Still, no universally accepted metric is defined [11]. The previously mentioned methods consider few situations of errors and have some certain deficiencies. e.g., they are highly biased by their assumptions [9] and may misleadingly present convincible results. Our recent study highlights the benefits of measuring the quality of an activity recognition algorithm through various perspectives [12].

In this paper, we propose a novel multimodal metric to study different properties of a SED system including detection, uniformity, total duration, and relative duration rather than using common machine learning metrics. Furthermore, our proposed method resolves the dependency on pre-defined strict parameters such as ρ_{DTC} and ρ_{GTC} in PSDS, length in collar, time resolution in frame-based (segment) and event-based method that are widely used in SED evaluation [7, 2, 13]. For the sake of transparency and uniformity in implementation, our method is published as an open source library[5].

In the remaining sections, initially, in section 2, we formally present our proposed method. In section 3, we evaluate our method on state-of-the-art systems submitted in the DCASE 2020 challenge. Lastly, we conclude with a brief discussion for future directions.

2. PROPOSED METHOD

In classical problems, an instance is either correctly detected (TP) or not (FP or FN). However, instances in SED are durative (events start and end at a specific time). Therefore, predictions may identify parts of references (figure 1). As a result, the TP, FP, and FN should have

[1] CLassification of Events, Activities and Relationships

[2] Detection and Classification of Acoustic Scenes and Events

[3] Polyphonic Sound Detection Score

[4] Audio and Acoustic Signal Processing

[5] Accessible from github.com/modaresimr/SED-MME-eval

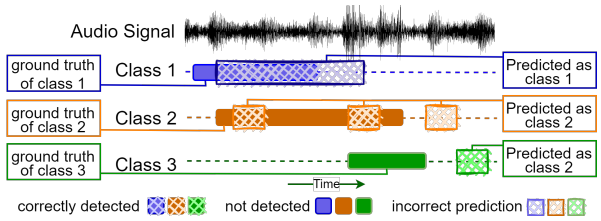

Fig. 1: SED instances are partially correct and partially incorrect while the classical ones are either correct or not.

x=[a:b]	x (a lower case letter) is an interval from a to b
x^s, x^e	start (x^s) and end (x^e) of interval x
X	uppercase letters indicate an interval set
$\|X\|$	number of elements in set X
x_i	i^{th} element of set X ($1 \leq i \leq \|X\|$)
R, P	positive references (R) and predictions (P) interval set
R^-	$\{[r^e_i : r^s_{i+1}] \| 1 \leq i < \|R\|\} \cup \{[0 : r^s_1], [r^e_{\|R\|} : MAX]\}$
P^-	$\{[p^e_i : p^s_{i+1}] \| 1 \leq i < \|P\|\} \cup \{[0 : p^s_1], [p^e_{\|P\|} : MAX]\}$
$x \cap y$	$\{[a:b] \| a = \max(x^s, y^s) \wedge b = \min(x^e, y^e) \wedge a < b\}$
$X \cap Y$	$\{z \| x \in X \wedge y \in Y \wedge z \in x \cap y\}$
$\mathcal{C}(x, Y)$	$\{y \in Y \| x \cap y \neq \emptyset\}$ //correlated to x in set Y
$\mathcal{T}$	interval: $\mathcal{T}(x) = x^e - x^s$ set: $\mathcal{T}(X) = \Sigma_{x \in X} \mathcal{T}(x)$
[ ]	represents the Iverson bracket. It is one when the enclosed condition is satisfied; otherwise, it is zero.

Table 1: Notation

a partial value between zero and one. Additionally, the situations of predicted events (e.g., predicting a reference event by multiple fragmented predictions) should be considered in the evaluation method. In our proposed method, we define a set of properties that are essential in SED. For each property, a set of measurements (in terms of TP, FP and FN) are defined. Therefore, well-known recall, precision and f1-score can be computed for each property. These measurements together constitute our proposed metric. Using a weighted combination of them can provide a single value or they can be considered collectively as a multi-objective metric. We find that researchers are interested in capturing some properties of an algorithm such as:
Detection (D): the capability of detecting reference events regardless of their duration.
Uniformity (U): the ability of detecting each reference event by a single prediction instead of multiple fragmented predictions.
Total Duration (T): the capability of recognizing duration of correctly detected parts of reference events.
Relative Duration (R): the ability of recognizing the duration of each reference event individually (normalized duration of correct parts).

In the following subsections, we formally define each property. Moreover, we assume that acceptable time shift of detected predictions is within the reference range. i.e., predictions and references are relevant when the intersection of their interval is not empty. Furthermore, we consider events have binary classes (positive and negative). To consider multi-classes events, each class are evaluated individually as a positive class and the rest as a negative class. Some basic notations used in this paper are defined in the table 1. In this table, R^- and P^- denote intervals of references and predictions that are negative. Zero and MAX term in the table 1 refere to the begining and ending of audio track. In the following, we describe the mentioned properties and present our formulas for measuring TP^P, FP^P and FN^P, ($P \in \{D, U, T, R\}$) of each property separately.

2.1. Detection Property (D):

In some applications (e.g., alarm systems) or in some sound classes (e.g., gun shouting class), detecting the occurrence of a sound event is more important than its duration [14, 15]. Detection property calculates the identifying a reference event even by a short prediction. A positive reference is considered either correctly identified (TP) when a part of it is detected by at least one prediction, or falsely not identified (FN) if it is not detected at all. Predictions that do not correspond to any reference events are considered as FP. It is formulated in equation (1) using notations described in table 1.

$$TP^D = \Sigma_{r:R}[\mathcal{C}(r, P) \neq \emptyset], \qquad FP^D = \Sigma_{p:P}[\mathcal{C}(p, R) = \emptyset], \quad (1)$$
$$FN^D = \Sigma_{r:R}[\mathcal{C}(r, P) = \emptyset]$$

Therefore, in equation (1), TP is the number of positive references that are detected by positive predictions without considering their duration. FN counts the positive references that are not detected by any positive predictions. Incorrect positive predictions are considered as FP (they are falsely predicted as positive while all the related references are completely negative).

2.2. Uniformity Property (U):

Some reference events may be detected by multiple predictions, or conversely, some predictions may cover multiple references. While detection property take into account identifying a reference, the uniformity property considers the detection of a reference event without fragmentation. It is essential when the number of occurrences of events is important. Researchers in [15] and [16] consider this as a primary property for anomaly detection and activity recognition; however, they do not provide a clear definition of uniformity. We define a reference event as correctly detected (TP) if it is identified by a single prediction; otherwise, we consider it as partial TP, FN and FP. They are formulated in equation (2):

$$\mathcal{Z}(y, X, Y) = \cup_{x:\mathcal{C}(y,X)} \mathcal{C}(x, Y)$$
$$TP^U = \sum_{r:R} \frac{1}{|\mathcal{Z}(r, P, R)|} \text{ if } \mathcal{C}(r, P) \neq \emptyset,$$
$$FN^U = \sum_{r:R} 1 - \frac{1}{|\mathcal{Z}(r, P, R)|} \text{ if } \mathcal{C}(r, P) \neq \emptyset, \quad (2)$$
$$FP^U = \sum_{p:P} 1 - \frac{1}{|\mathcal{Z}(p, R, P)|} \text{ if } \mathcal{C}(p, R) \neq \emptyset$$

Consequently, the function $\mathcal{Z}(y, X, Y)$, first selects elements in X that have intersection with y; then, it finds elements in Y that are correlated with selected elements of X. For example, $\mathcal{Z}(r, P, R)$ first looks for predictions that detect r; it then returns the references identified by those predictions. e.g., in figure 2, $|\mathcal{Z}(r_8, P, R)|$ is one and it is three for $|\mathcal{Z}(r_{12}, P, R)|$. Therefore, the TP value for r_8 is one and the FN value is zero; however, for r_{12}, the TP value is $1/3$, and the FN value is $1 - 1/3$. Similarly, in calculating FP, $\mathcal{Z}(p, R, P)$ returns predictions that identify the references predicted by p.

2.3. Total Duration Property (T):

The previously mentioned properties do not consider the duration of references. Total duration property is helpful in duration-sensitive sound classes. It is similar to time-frame and segment-based calculation [2, 8]. It divides the whole series by the boundaries of predictions and references; then, it classifies each part as either TP, FP, or FN. Equation (3) defines the total duration property.

$$\mathrm{TP}^{\mathrm{T}} = \mathcal{T}(\mathrm{R} \cap \mathrm{P}),\ \ \mathrm{FP}^{\mathrm{T}} = \mathcal{T}(\mathrm{P} \cap \mathrm{R}^{-}),\ \ \mathrm{FN}^{\mathrm{T}} = \mathcal{T}(\mathrm{P}^{-} \cap \mathrm{R}) \quad (3)$$

Therefore, TP is the duration of correctly detected parts of positive references, and FP (FN) is the duration of incorrect parts of positive (negative) predictions that are negative (positive) in the reference.

2.4. Relative Duration Property (R):

In the previously mentioned property (total duration), long events can affect the overall performance. Relative duration property normalizes each event's duration individually to lessen the effect of varying durations in events. It is useful where the correctly detected parts of each individual event is important. The normalization procedure is done based on references and is shown in equation (4).

$$\begin{aligned}
\mathrm{TP}^{\mathrm{R}} &= \sum_{\mathrm{e}:\mathrm{R}\cap\mathrm{P}} \frac{\mathcal{T}(\mathrm{e})}{\mathcal{T}(\mathcal{C}(\mathrm{e},\mathrm{R}))}, \\
\mathrm{FN}^{\mathrm{R}} &= \sum_{\mathrm{e}:\mathrm{R}\cap\mathrm{P}^{-}} \frac{\mathcal{T}(\mathrm{e})}{\mathcal{T}(\mathcal{C}(\mathrm{e},\mathrm{R}))} \text{ if } \mathcal{C}(\mathrm{e},\mathrm{P}) \neq \emptyset, \\
\mathrm{FP}^{\mathrm{R}} &= \sum_{\mathrm{e}:\mathrm{R}^{-}\cap\mathrm{P}} \frac{\mathcal{T}(\mathrm{e})}{\mathcal{T}(\mathcal{C}(\mathrm{e},\mathrm{R}^{-}))} \text{ if } \mathcal{C}(\mathrm{e},\mathrm{R}) \neq \emptyset
\end{aligned} \quad (4)$$

Accordingly, TP is the sum of the normalized duration of correctly detected positive predictions. i.e., it sums the percentage of correctly detected parts of positive references. FP (FN) sums the normalized duration of wrongly predicted parts of positive (negative) predictions regarding their negative (positive) references. The denominator of these formula will never be zero because the zero duration interval are filtered in the intersection which exist below the sigma. The if conditions in the calculation of FP and FN select only detected references and correct predictions.

3. EXPERIMENT

Earlier, several TPs, FPs, and FNs were defined that measure the properties of a SED system independently. They can be used to calculate recall, precision, f-score, accuracy, error rate, etc. for each property. These properties can be visualized and be extended easily. A total value can be obtained using a weighted combination or any other multi-objective method. In this experiment, equal weights linear combination is used.

In order to analyze our method, we first demonstrate its soundness by considering all major possible situations between references and predictions. Then, we compare the sound classes of two SED systems in detail. Lastly, we re-evaluate the best ten systems presented in DCASE 2020 challenge. For the sake of reproducibility, the data, source code and the details of the experiments are available in our repository at https://github.com/modaresimr/SED-MME-eval

3.1. Computation Complexity Analysis:

The presented formulations iterate on both sets of P and R; thus, it has a complexity of $O(|R| \times |P|)$. Interval tree helps us to optimize it to $O(|R|log|R| + |P|log|P|)$. When P and R are sorted by time, the complexity of $O(|R|+|P|)$ can be achieved by considering their time relationship since R and P are traversed once.

3.2. Detailed analysis on artificially generated data:

In this experiment, we consider all major possible situations between references and predictions using artificially generated data. This data contains four parts and is visualized in figure 2. The first part is about simple situations (one reference is related to only one prediction). It includes all 13 relations described in Allen's interval algebra [17]. The second part shows fragmented prediction. The third part considers a single prediction that cover multiple references. Lastly, fragmented and merged predictions are considered simultaneously. The evaluation outputs on each part are available on table 2.

Verifying detection property is straightforward. We consider all reference events that have at least one common part with predicted events as TP ($r_{2\ldots17}$), other reference events as FN ($r_{0,1}$) and false positive predictions are FP ($p_{0\ldots2}$) in this property.

The uniformity property captures detection of reference events by multiple predictions in a fragmented manner (e.g., r_{11} is recognized by three predictions ($p_{12\ldots14}$); thus, each one is partial FP ($^2/_3$)), or one prediction cover multiple references (e.g., $\mathrm{r}_{12\ldots14}$ are recognized by p_{15}; thus each one is partial FN ($\mathrm{FN}_{\mathrm{r}_{12\ldots14}} = {}^2/_3$) and partial TP ($\mathrm{TP}_{\mathrm{r}_{12\ldots14}} = {}^1/_3$)); otherwise, the predictions are complete TP (e.g., $\mathrm{TP}_{\mathrm{r}_{2\ldots7}} = 1$). In the fourth part, similar to the second one, each reference is identified by multiple predictions and also, similar to the third part, one prediction (p_{18}) covers three references ($\mathrm{FP}_{\mathrm{p}_{16,17}} = {}^2/_3, \mathrm{FP}_{\mathrm{p}_{18}} = {}^3/_4, \mathrm{FP}_{\mathrm{p}_{19}} = {}^1/_2$).

Total duration property divides the predictions to independent parts and mark them as TP, FP, and FN (similar to figure 1); then it sums their time intervals (e.g., $\mathrm{TP}_{\mathrm{r}_{2\ldots10}} = \mathrm{FP}_{\mathrm{p}_{2,7}} = \mathrm{FN}_{\mathrm{r}_3} = {}^1/_2$). The segment-based method is similar, but since it reduces the time resolution to one second [18]); they produce different results.

Evaluation of long events is the dominant output in the total duration property. Therefore, the relative duration property is used to calculate the normalized correctly recognized part of each event. Therefore, each partial TP, FP, and FN is normalized depend on its correlated reference events (e.g., $\mathrm{TP}_{\mathrm{r}_2} = 1, \mathrm{TP}_{\mathrm{r}_3} = \mathrm{FN}_{\mathrm{r}_3} = \mathrm{FP}_{\mathrm{p}_7} = {}^1/_2, \mathrm{TP}_{\mathrm{r}_4} = {}^1/_3$).

The state-of-the-art methods also exist in table 2. In their definition, each of TP, FP and FN is either zero or one; while they can have a partial values in our definitions. The collar method provides only one TP (r_2) because the collar range is 200 ms [9] and the timing errors are 500ms in this data. The $psd\ d/gtc$=0.8 has a similar situation because its acceptable time shift is 200ms for each second of events. The $psd\ d/gtc$=0.1 and our detection property provide similar results because its parameter is small enough in this artificial data. However, an inconsistency exists in the third part. The FP calculated by $psd\ d/gtc$=0.8 is 1 while the TP calculated by $psd\ d/gtc$=0.5 is 3. The result produced by $psd\ d/gtc$=0.8 for third part is similar to the p_2, which means it ignores existence of two references in third part. This shows that PSD method in state-of-the-art needs some improvements. However, our method does not show this inconsistency.

3.3. Detailed comparison of two SED systems:

The second analysis is made over the best systems in DCASE challenge[6] 2020 Task 4 (Miyazaki and Hao-CQU) on the public dataset [11]. We provide detailed information for sound classes by showing the TP, FN, and FP provided by different evaluation methods in figure 3. By decreasing the parameter of PSD, it will be closer to our detection property, and by increasing that parameter, it will be closer to the collar method. When the duration of all references and predictions is one second, 100ms collar and $psd\ d/gtc = 0.8$ provide close results. The segment method's objective is to allow some misalignment between the reference and prediction [2]; however, when a prediction is completely incorrect, this method provides more FP

[6]website: `https://dcase.community/challenge2020/`

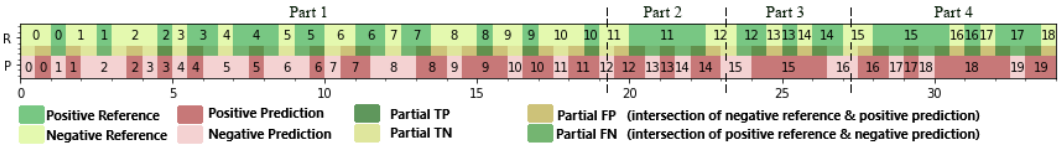

Fig. 2: An example scenario that contains all major possible situations between references and predictions. The numbers correspond to the i-th reference (prediction) and are indicated by r_i (p_i).

	Part 1			Part 2			Part 3			Part 4		
	TP	FN	FP	TP	FN	FP	TP	FN	FP	TP	FN	FP
*detection	9	2	3	1	-	-	3	-	-	3	-	-
*uniformity	9	-	-	1	-	2	1	2	-	1	2	2.6
*total durati.	4.5	4	4.5	1.5	1	1	1.5	1	1	3	1.5	2
*relative dur.	6.3	2.7	3.5	0.6	0.4	0.7	2	1	-	2.3	0.7	0.7
collar	1	10	11	-	1	3	-	3	1	-	3	4
segment	10	4	5	3	-	1	3	1	-	6	-	1
psd d/gtc=0.1	9	2	3	1	-	-	3	-	-	3	-	-
psd d/gtc=0.5	7	4	4	1	-	-	3	-	-	3	-	-
psd d/gtc=0.8	1	10	8	-	-	2	-	-	1	-	-	3

Table 2: Different methods for defining TP, FN, and FP on sample data. Our methods are identified by *.

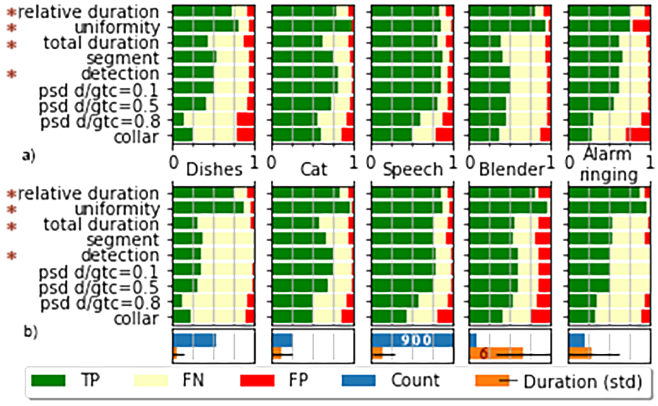

Fig. 3: Detailed information on different classes in two systems. a) S1:Miyazaki-NU-1, b) S3:Hao-CQU-4. * specifies our methods.

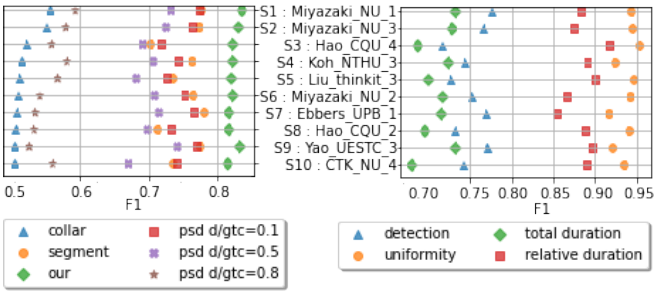

Fig. 4: (left) F1-score of different systems using event, segment, psd and our metric, (right) the details of our metric.

than usual (e.g., Blender class in figure 3.b). This is opposite to its goal. The hypothesis in segment method decreases the time resolution; thus, it may have a wrong impact on the final result. Therefore, we choose exact timing in calculating the total duration property.

Figure 3 shows that system (b) recognizes fewer events (detection) than system (a); however, its detections are less fragmented (better uniformity) and more precise in detecting the event's time interval (relative duration), while neither collar method nor PSD method can capture uniformity and relative duration.

3.4. Global Comparison of Top SED Systems:

To demonstrate the advantages of the proposed metric more globally, the top ten SED systems (based on collar evaluation) in DCASE challenge 2020 Task 4 are evaluated by F_1-score over public dataset [11] under collar, segment, PSD (with three different configurations d/gtc=0.8, 0.5, and 0.1) and ours in figure 4.

Re-interpretation of TP, FN, and FP causes changes in the rankings of those systems. The collar and PSD methods contain strict hypothesis that a prediction is acceptable when the timing difference between reference and prediction is within a fixed values. Hence, all the events that satisfy these conditions are similar; besides, they provide significant differences between two events that are close to the boundaries of their preset conditions. For these reasons, the system ranked ninth by collar method takes third place by using our method.

Unlike other approaches with strict predefined parameters, our metric is independent to parameters in calculating properties (detection, uniformity, total duration, and relative duration). The only optional parameter (weights) in our metric, is used to prioritize the properties, which can be easily changed without recalculating the properties. Therefore, an appropriate algorithm can be selected easier for a new application with different constraints by prioritizing those properties differently. e.g., an algorithm that performs better in uniformity property; is expected to be more suitable for an application that uniformity is essential.

4. CONCLUSIONS

In this paper, we presented a new multimodal metric. Our proposed method redefines the well-known TP, FP, and FN for SED systems by allowing partial value for each one versus current methods that each TP, FP and FN is either zero or one. We also presented an analysis of our methods versus state-of-the-art on the best SED systems in the DCASE 2020 challenge. Regarding to the experimental results, standard metrics fail to reveal the actual performance of SED systems. It shows that our method is expressive (by capturing several properties of SED systems such as detection, total duration, relative duration and uniformity), generalizable (prioritizing properties can support a wide range of applications) and extensible (adding a new property is straightforward and independent of others). Our proposed method can also facilitate the selection of a SED system for a new application by prioritizing the properties differently which will be explored in future studies. Future studies will also include extending this metric to directly deal with multi-class cases to build a new confusion matrix and to be formalized based on generalized interval [19].

5. REFERENCES

[1] Cagdas Bilen, Giacomo Ferroni, Francesco Tuveri, Juan Azcarreta, and Sacha Krstulovic, "A Framework for the Robust Evaluation of Sound Event Detection," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 61–65.

[2] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences (Switzerland)*, vol. 6, no. 6, 2016.

[3] T K Chan and Cheng Siong Chin, "A Comprehensive Review of Polyphonic Sound Event Detection," *IEEE Access*, vol. 8, pp. 103339–103373, 2020.

[4] Rainer Stiefelhagen, Keni Bernardin, Rachel Bowers, John Garofolo, Djamel Mostefa, and Padmanabhan Soundararajan, "The CLEAR 2006 Evaluation," in *Multimodal Technologies for Perception of Humans*, vol. 4122 LNCS, pp. 1–44. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[5] Andrey Temko, Climent Nadeu, Dušan Macho, Robert Malkin, Christian Zieger, and Maurizio Omologo, "Acoustic Event Detection and Classification," *Computers in the Human Interaction Loop*, pp. 61–73, 2009.

[6] Rainer Stiefelhagen, Keni Bernardin, Rachel Bowers, R. Travis Rose, Martial Michel, and John Garofolo, "The CLEAR 2007 evaluation," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 4625 LNCS, pp. 3–34, 2008.

[7] Dan Stowell, Dimitrios Giannoulis, Emmanouil Benetos, Mathieu Lagrange, and Mark D. Plumbley, "Detection and Classification of Acoustic Scenes and Events," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, 2015.

[8] Tim Van Kasteren, Hande Alemdar, and Cem Ersoy, "Effective performance metrics for evaluating activity recognition methods," in *ARCS*, 2011.

[9] Giacomo Ferroni, Nicolas Turpault, Juan Azcarreta, Francesco Tuveri, Romain Serizel, Çagdaş Bilen, and Sacha Krstulović, "Improving Sound Event Detection Metrics: Insights from DCASE 2020," 2020.

[10] Noriyuki Tonami, Keisuke Imoto, Takahiro Fukumori, and Yoichi Yamashita, "Evaluation Metric of Sound Event Detection Considering Severe Misdetections By Scenes," , no. November, pp. 1–5, 2020.

[11] Romain Serizel, Nicolas Turpault, Ankit Shah, and Justin Salamon, "Sound event detection in synthetic domestic environments," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2020-May, pp. 86–90, 2020.

[12] Seyed Modaresi, Aomar Osmani, Mohammadreza Razzazi, and Abdelghani Chibani, "Uniform Evaluation of Properties in Activity Recognition," in *Advances in Knowledge Discovery and Data Mining*. 2022, Springer International Publishing.

[13] Nicolas Turpault, Romain Serizel, Justin Salamon, and Ankit Parag Shah, "Sound Event Detection in Domestic Environments with Weakly Labeled Data and Soundscape Synthesis," pp. 253–257, 2019.

[14] Emmanuel Munguia Tapia, Stephen S. Intille, and Kent Larson, "Activity Recognition in the Home Using Simple and Ubiquitous Sensors," in *Pervasive Computing*, Alois Ferscha and Friedemann Mattern, Eds., pp. 158–175. Springer, Heidelberg, 2004.

[15] Nesime Tatbul, Tae Jun Lee, Stan Zdonik, Mejbah Alam, and Justin Gottschlich, "Precision and recall for time series," *Neural Information Processing Systems (NIPS)*, 2018.

[16] Jamie A. Ward, Paul Lukowicz, Hans W. Gellersen, and Ward, "Performance metrics for activity recognition," *ACM Transactions on Intelligent Systems and Technology*, 2011.

[17] Aomar Osmani, "STCSP : A Representation Model for Sequential Patterns," *Foundations and Applications of Spatio-Temporal Reasoning (FASTR)*, 2003.

[18] Toni Heittola, Annamaria Mesaros, Antti Eronen, and Tuomas Virtanen, "Context-dependent sound event detection," *Eurasip Journal on Audio, Speech, and Music Processing*, vol. 2013, no. 1, pp. 1–13, 2013.

[19] Aomar Osmani, "Learning Patterns in Multidimensional Space Using Interval Algebra," in *International Conference on Artificial Intelligence: Methodology, Systems, and Applications*. Springer, 2002, pp. 31–40.