JOINT MEASUREMENT OF LOCALIZATION AND DETECTION OF SOUND EVENTS

Annamaria Mesaros, Sharath Adavanne, Archontis Politis, Toni Heittola, Tuomas Virtanen

Computing Sciences, Tampere University
Korkeakoulunkatu 1, 33720, Tampere, Finland
email: name.surname@tuni.fi

ABSTRACT

Sound event detection and sound localization or tracking have historically been two separate areas of research. Recent development of sound event detection methods approach also the localization side, but lack a consistent way of measuring the joint performance of the system; instead, they measure the separate abilities for detection and for localization. This paper proposes augmentation of the localization metrics with a condition related to the detection, and conversely, use of location information in calculating the true positives for detection. An extensive evaluation example is provided to illustrate the behavior of such joint metrics. The comparison to the detection only and localization only performance shows that the proposed joint metrics operate in a consistent and logical manner, and characterize adequately both aspects.

Index Terms— Sound event detection and localization, performance evaluation

1. INTRODUCTION

Sound event detection and sound localization or tracking are traditionally two different areas of research, as they deal with completely separate aspects of identifying sounds: one aiming to find the correct label and temporal position of sounds, the other aiming to find the correct spatial and temporal position. Recent research addresses localization and detection of sound events within the same system [1, 2, 3, 4]; however, for lack of a better measurement method, the localization and detection performance are evaluated separately [5, 6]. This is a paradoxical solution for evaluating performance of a system in which the two are modeled and predicted jointly.

Sound event detection metrics measure the ability of the system in finding the correct sound events or the amount of errors the system makes. The sound events are organized in classes defined by their labels, and evaluation measures if the detected events at a given time are assigned to the correct class. Different metrics serve different purpose, and generally the selection of a performance measure is dictated by the application [7, 8]. Most often used are error rate and F1-score, calculated in fixed length segments of typically 1 second [9], but metrics that consider sound event instances are also available [10]. Similar metrics are used for example in polyphonic music transcription [11], speech recognition and speaker diarization. These metrics have no means of representing location information for the detected sound; as a consequence, localization errors–sounds with incorrect position but correct label–would be considered correctly detected.

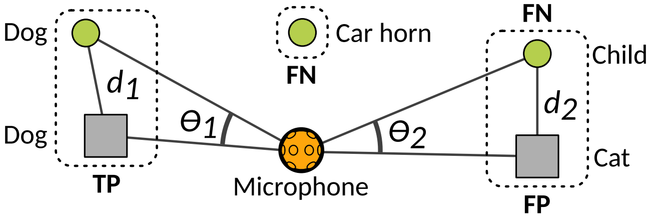

Figure 1: Example reference and predicted sound events and locations. Circles denote reference sounds, rectangles system output.

On the other hand, localization performance is commonly assessed through statistical analysis of the instantaneous angular or positional localization errors between the system output and the reference. Purely localization systems have generic source models with no class information. In the case of multi-source localization [12], evaluation can be based on an average instantaneous localization error after a minimum distance assignment has been made between the reference positions and the estimates, e.g. with the Hungarian algorithm [13]. Tracking systems do associate identities to sources, but only for the purpose of forming continuous location estimates, i.e. a track, for each source, therefore unrelated to to the signal content or sound class. For example if a sound event stops being active after being tracked with a certain identifier, and re-appears after a long time, a new identity can be assigned to its new track without any penalty on the performance of the system. If multiple tracks occur simultaneously and identities are swapped between them, the metrics should be able to strike a balance between localization accuracy and maintaining temporal association of localization estimates with the appropriate reference tracks [14, 15]. Evaluation is more complex in the case of multi-source tracking, and suitable metrics are still a topic of research [16].

We illustrate joint localization and detection for one time frame in Fig. 1: the reference annotation contains three sound events belonging to classes *dog*, *car horn* and *child*, while the system predicts two: *dog* and *cat*, each at their respective reference and predicted positions. Sound event detection aims to find the sound events in this frame and label them correctly, and its evaluation consists in comparing the labels of the reference and predicted sound events. This results in one true positive (reference “dog”, prediction “dog”), one false positive (prediction “cat”, no reference “cat”) and two false negatives (reference “car horn”, “child”, no prediction for these classes). Sound event localization aims to find the sound events at correct spatial positions, and the evaluation considers only the spatial errors between the closest sound sources. This results in two error measurements (“dog”-“dog” and “child”-“cat”). In contrast, sound event localization and detection requires both the labels and the locations to be correct, therefore it would only consider the “dog”-“dog” pair as correct localization and detection.

This work has received funding from the European Research Council under the ERC Grant Agreement 637422 EVERYSOUND

978-1-7281-1123-0/19/$31.00 ©2019 IEEE

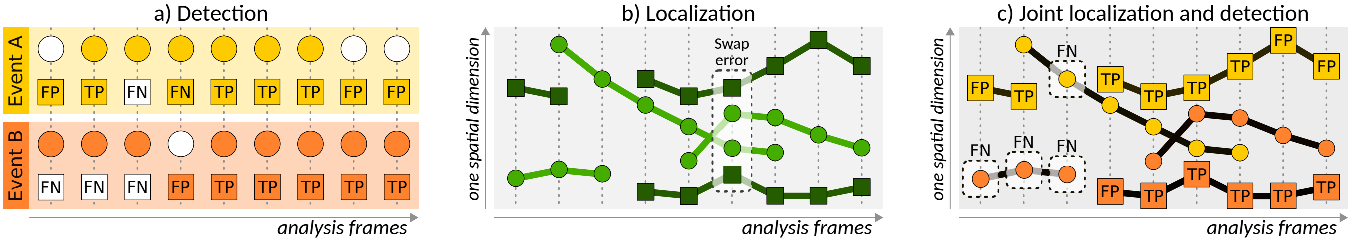

Figure 2: Evaluation of detection, localization, and their joint measurement, with two different sound events active at the same time. Circles denote reference sounds, rectangles system output. For simplification, only one spatial dimension is used in the illustration.

In this paper, we formulate a procedure for joint measurement of localization and detection performance. We start with the frame-based formulation, then consider the generalization to a segment-based version. Segment-based evaluation is used in sound event detection to alleviate the effect of onset/offset subjectivity in the reference annotations, therefore we present a similar way of measuring joint localization and detection. We approach the joint measurement from both detection and localization perspectives, and formulate *location-sensitive detection metrics* that count correct and erroneous detection cases within certain spatial error allowance, and *class-sensitive localization metrics* that measure the spatial error between sound events with same label. We show that the proposed metrics behave consistently and characterize adequately both aspects, in comparison with localization or detection only metrics.

The paper is organized as follows: Section 2 introduces the metric and presents the frame-based formulation, and Section 3 describes its generalization to segment-based measuring. Section 4 uses an example system for illustrating how the proposed metrics behave, and Section 5 presents a more general discussion of their characteristics. Finally Section 6 presents conclusions and future work.

2. PROPOSED METHOD

Localization and detection of sound events imply detection of a sound event with correct label at the correct temporal position and correct spatial location, as given by reference annotation. For measuring the localization and detection performance, we propose comparing both label and location at the same time when deciding if the system output is correct. This effectively means taking into account only the spatial errors between sounds that belong to the same class, and counting all other reference or predicted ones as errors.

Fig. 2 illustrates the different evaluation cases, with Fig. 2.c showing the labels and locations of two sound sources in consecutive frames. Sound event detection evaluation, illustrated in Fig. 2.a, checks only the event labels, for presence or absence of events belonging to the same class. Intermediate statistics are counted as true positives (*TP*, reference and predicted event active at the same time), false positives (*FP* or *insertions* *I*, reference inactive, predicted active), false negatives (*FN* or *deletions* *D*, reference active, predicted inactive); one true positive and one true negative appearing at the same time count as a single *substitution* error S, and N is the total number of reference events. Common sound event detection metrics include precision, recall, F-score and error rate, and they are calculated based on the total counts TP, FP, FN, and D, I, S, N respectively:

$$P = \frac{TP}{TP+FP} \quad R = \frac{TP}{TP+FN} \quad F = \frac{2PR}{P+R} \qquad (1)$$

$$ER = \frac{D+I+S}{N} \qquad (2)$$

For the same case, spatial errors will be measured by pairing the closest detected and reference sounds irrespective of their label, as illustrated in Fig.2.b. For the illustrated case, the method in [15] would count one swap error when the tracks intersect, and continue the optimal pairing afterwards. For a single pair, the spatial error can be expressed as an angular distance θ between directions-of-arrival (DoA) or Euclidean distance d between actual positional estimates, depending on the localization output of the system:

$$\theta = \arccos(\mathbf{u}_{\text{ref}} \cdot \mathbf{u}_{\text{so}}), \qquad d = ||\mathbf{x}_{\text{ref}} - \mathbf{x}_{\text{so}}|| \qquad (3)$$

where $\mathbf{u}$ is a unit vector pointing to a DoA, and $\mathbf{x}$ is a Cartesian position vector.

The proposed method is illustrated in Fig. 2.c, and includes both detection and localization criteria. Location-sensitive detection metrics can be calculated by counting true and false positives and negatives, and class-sensitive localization metrics can be calculated by measuring the spatial error between the detected and reference sound events with same label. In multi-instance cases, with potentially multiple prediction and reference events of the same class, one of the established tracking metrics could be used for the associations; however, this case is not yet within the scope of our work.

Location-sensitive detection: For detection evaluation, we consider allowing a small error in location, so that if a sound is localized approximately at its reference location, it is considered correctly detected. We therefore measure the spatial error between the predicted and reference events with same label, and count a true positive only when its label is correct and its location is within a threshold θ or d from its reference location, calculated according to (3). The detection criterion is included through the label use, and the localization criterion through the spatial error threshold.

The intermediate statistics are defined as in sound event detection [10], with the only modification for the true positive: in order to be considered a true positive, the spatial error for a detected event must be within the given threshold from the reference. If the predicted event is further than this threshold, the event is detected at a different location, therefore produces a false positive, while the undetected reference event produces a false negative. With multiple classes, these errors are likely to be counted towards substitution errors, therefore resulting in a single penalty for the pair. Once the intermediate statistics are obtained from the comparison of the system output with the reference, any metric defined for sound event detection can be calculated, e.g. F1-score, ER, etc.

The drawback of such a metric is that it discards information on the actual value of the spatial error, which is in fact very important in the sound localization and tracking research, so from the tracking point of view, expressing system performance based on the true and false positives and negatives is unsuitable.

978-1-7281-1123-0/19/$31.00 ©2019 IEEE

Class-sensitive localization: To evaluate more closely localization, we want to keep detailed information on the spatial errors. Inclusion of the detection criterion into the original localization metric is straightforward: we only calculate the spatial errors between sounds with the same label in each frame. Furthermore, as there is no threshold for the true positives, all intermediate statistics are counted as in detection. System performance can then be presented in terms of average spatial error with its corresponding detection performance–to indicate how much of the total number of test cases this spatial error characterizes.

3. GENERALIZATION TO DIFFERENT TIME SCALE

For practical reasons, sound event detection is often evaluated using segment-based metrics. Segment-based metrics allow evaluation with a temporal resolution independent of the resolution of the system, and alleviate the effect of onset/offset subjectivity in reference annotations. However, extending the formulation of this joint measurement for detection and evaluation to an arbitrary segment length is not trivial, because segment-based metrics require estimating one activity indicator (event active or inactive) and its corresponding spatial error within the segment.

Consider the case in Fig. 2.c. If all the illustrated consecutive frames form a *segment*, according to [10], both sound events are correctly detected within this segment, because both system output and reference have both A and B sound events active within this segment. The question is: how to apply the localization criterion?

One possibility is to calculate the average position of the sound within the evaluation segment, then calculate the spatial error between the average locations of the predicted and reference events. This can be accomplished by calculating the mean Cartesian DoA $\hat{\mathbf{u}}$ or position vector $\hat{\mathbf{x}}$ within the segment for the corresponding predicted or reference event:

$$\hat{\mathbf{u}}(n) = \sum_{i=1}^{L_n} \mathbf{u}_i / || \sum_{i=1}^{L_n} \mathbf{u}_i ||, \qquad \hat{\mathbf{x}}(n) = \sum_{i=1}^{L_n} \mathbf{x}_i / L_n, \tag{4}$$

where L_n are the number of reference points or system output estimates inside the segment. In consequence, the location information is transformed to a more coarse time resolution, as illustrated in Fig. 3. If errors exist in the prediction, the average location of the detected sound and that of the reference are calculated using different number of points, i.e. based on the amount of information available for each. The spatial error for each segment is then calculated between the two estimates.

A more tracking-oriented way is to calculate the average localization error within this segment for all pairs of reference and predicted events, frame-by-frame:

$$\hat{\theta}_{\text{segm}}(n) = \sum_{i=1}^{K_n} \theta_i / K_n, \qquad \hat{d}_{\text{segm}}(n) = \sum_{i=1}^{K_n} d_i / K_n, \tag{5}$$

where θ_i and d_i are the frame-based spatial errors according to Eqn. 3, and K_n is the number of frames with true positives in the n-th segment. This is a more accurate estimation of the spatial error than using the average location, as it includes only the true positives and the errors are calculated based on the reference and the system output in exactly the same time frames.

Once the spatial and detection information is transformed to segment-based resolution, all the previously described frame-based metrics can be applied: we can measure the spatial error between the predicted and reference events for the class-sensitive localization, or use a threshold for counting the true positives and evaluate it as location-sensitive detection.

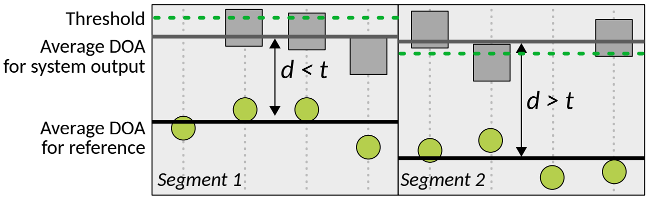

Figure 3: Location estimate as average DoA within the segment; spatial error for the segment is calculated based on the averages

4. EVALUATION

We consider the baseline system provided for DCASE 2019 Challenge Task 3: Sound Event Localization and Detection [17]. The system consists of a convolutional recurrent neural network that maps the input magnitude and phase components of a spectrogram into two outputs: temporal activity of sound classes in the dataset (sound event detection) as a multi-class multi-label classification task, and spatial trajectory when the respective sound class is active (DoA trajectory estimation) as a multi-output regression task. The system is trained on the TAU Spatial Sound Events 2019 - Microphone Array dataset provided for the same DCASE task [1]. An early stopping criterion of 25 epochs was used to halt training if the standalone detection and localization performance did not improve. The best results were obtained after about 75 epochs of training.

We present detailed results in Tables 1 and 2 using TAU Spatial Sound Events 2019 - Microphone Array dataset[18]. The dataset was synthesized using spatial room impulse responses (IRs) from five indoor locations, at 504 unique combinations of azimuth-elevation-distance, and stationary sound sources from 11 event classes [17]. For comparison with the official DCASE task, we measure spatial error as directional error DE, calculated as the average of frame-wise angular distances [6]. Additionally, we evaluate the same system using another dataset [19] that contains the same sound events, but uses simulated IRs [20], and sources are moving in varying velocities in complete azimuth and elevation angles[2].

Table 1 presents the class-sensitive localization performance. We include performance at different stages during training, to observe the metrics behavior within the system. For comparison, the table includes the DCASE baseline performance for localization only, DE$_L$, and the frame recall FR$_L$, which is the accuracy of the system in detecting the correct number of sources [6]. During training, all measured aspects improve as the system learns to predict better the location and labels for sound events. Frame-based directional error DE$_L$ is 30.8, increasing to a class-sensitive DE$_{CL}$ of 34.3 when both localization and detection criteria are included–which means that DE$_L$ includes some pairings that are wrongly labeled events. Evaluated in 1 s segments, DE$_{CL}$ calculated based on average location within the segment (AL_seg method in Table 1) is only slightly smaller than DE$_{CL}$ calculated using average spatial error within the segment (SE_seg method), showing that even though different, the two approaches reach very similar conclusion.

Table 2 presents the location-sensitive detection performance for different angular error thresholds θ, using a segment of 1 s;

[1] https://doi.org/10.5281/zenodo.2599196
[2] https://doi.org/10.5281/zenodo.2636586

	20ms frame				Class-sens. loc., 0.5s segment				Class-sens. loc., 1s segment			
	Loc. (baseline)		Class-sensitive loc.		AL_seg		SE_seg		AL_seg		SE_seg	
Epochs	DE$_L$	FR$_L$	DE$_{CL}$	F$_{CL}$	DE$_{CL}$	F$_{CL}$	DE$_{CL}$	F$_{CL}$	DE$_{CL}$	F$_{CL}$	DE$_{CL}$	F$_{CL}$
5	69.4	72.8	79.7	67.4	81.8	68.8	81.8	68.6	82.1	69.6	82.1	68.9
25	42.1	82.0	48.0	78.6	50.7	78.5	51.4	78.3	51.0	78.5	51.9	78.2
75	30.8	84.0	34.3	80.9	36.3	80.0	37.5	79.9	**36.7**	**79.7**	**38.1**	**79.3**

Table 1: Class-sensitive localization performance: average directional error (degrees) and F-score (%) for different segment sizes

	Det. (baseline)			AL_seg		SE_seg	
Epochs	ER$_D$	F$_D$	θ	ER$_{LD}$	F$_{LD}$	ER$_{LD}$	F$_{LD}$
5	0.48	69.6	10	1.02	1.2	1.03	0.3
			20	0.99	4.5	1.01	2.5
			30	0.95	9.0	0.97	7.1
			40	0.90	14.7	0.92	13.3
75	0.35	80.0	10	0.92	11.9	0.98	5.6
			20	0.74	31.1	0.78	26.2
			30	0.61	45.0	0.63	43.4
			40	0.54	**54.2**	0.54	**54.2**

Table 2: Location-sensitive detection performance in 1 s segments

Class-sensitive localization				**Location-sensitive detection**				
AL_seg		SE_seg			AL_seg		SE_seg	
DE$_{CL}$	F$_{CL}$	DE$_{CL}$	F$_{CL}$	θ	ER$_{LD}$	F$_{LD}$	ER$_{LD}$	F$_{LD}$
12.7	95.3	**11.3**	**95.3**	**10**	0.37	**63.2**	0.37	**63.2**
				20	0.18	83.6	0.21	81.1
				30	0.13	89.8	0.15	87.3
				40	0.10	92.1	0.13	89.7

Table 3: Performance metrics in 1 s segments, moving source

DCASE baseline performance for detection is included for comparison as error rate ER$_D$ and F-score F$_D$. We observe the same consistent behavior in the measured performance, with the system improving during training. When a smaller threshold is imposed, naturally, a smaller number of true positives results in a lower F-score F$_{LD}$ and higher error rate ER$_{LD}$. Here also, the two approaches for estimating the segment-based localization error result in similar performance.

To gain more insight into the proposed metrics, we compare the class-sensitive localization and location-sensitive detection performance for similar angular error values (highlighted in Tables 1 and 2). In 1 s segments, DE$_{CL}$ is 36.7 and 38.1, so we compare the corresponding F$_{CL}$ of 79% (Table 1) with F$_{CL}$ for a threshold θ of 40 (Table 2), which is about 54%. The large difference shows that F$_{CL}$ takes into account many sounds that are detected very far away from the reference location, but the system also detects some sounds very precisely, bringing the average spatial error to under 40 degrees. In fact, only 54% of the predicted events are within 40 degrees from their reference, not 79% as indicated by F$_{CL}$.

For the dataset with moving sources, the proposed metrics and all variants exhibit similar behavior as for stationary sources. Table 3 presents results of the fully-trained system evaluated in 1 s segments. Calculated performance is similar for the two approaches of estimating segment-based location, and measured performance is higher for a more permissive threshold. DE$_{CL}$ of around 12 is comparable to the threshold of 10, with F$_{LD}$ showing that only 63 of the 95% F$_{CL}$ consists of events detected within 10 degrees from their reference position. The much better performance compared to Tables 1 and 2 is due to the lower complexity of the dataset.

5. DISCUSSION

The proposed measures combine two different ways of evaluation, therefore the segment-based formulation may seem questionable. One consequence of the change of time resolution from frames to the segment-based evaluation is allowing extreme cases where detection in a single frame within the segment drives the decision. However, similar cases are accepted in the segment-based evaluation of sound event detection, under the assumption that they happen only at the onset/offset times, and for large amount of data this should not have a considerable effect on the calculated performance. With a similar judgment to the localization and detection, the estimated spatial error in a segment can be based on a single correctly predicted output, and subsequent decisions are based on that.

Regarding the different estimation approaches for the segment-based spatial error, the one calculated based on average location hides much of the actual error, which is counter-intuitive for a localization measurement. In this respect, calculating directly the average error within the segment is more natural, as it follows closely the location changes. On the other hand, this imposes a frame-wise comparison before the actual evaluation, which is counter-intuitive to evaluating the performance at a coarser time resolution. The main difference between the two is the use of false positives in estimation of the average location, which means that the differences will be small when sources are not moving (average location of reference is the same as in each frame) and when the system is very good (predicted locations follow closely the reference). As observed in our multiple experiments, this difference is indeed rather small.

As a general recommendation, the metric for performance measurement should be chosen based on the aspect that needs to be emphasized. For characterizing the system in terms of spatial errors, DE$_{CL}$ and F$_{CL}$ are useful to indicate the average spatial error and detection performance. As mentioned, DE$_{CL}$ hides the actual distribution of the spatial errors, so if detection within a certain extent of spatial error is required, F$_{LD}$ or ER$_{LD}$ with the given threshold θ provides a better characterization of performance.

6. CONCLUSIONS AND FUTURE WORK

This paper introduced a novel approach to jointly evaluate the localization and detection performance of systems aimed at sound event localization and detection[3]. Different experiments showed that the proposed measures behave consistently, and they can be used for joint evaluation at any temporal resolution. The extension of these joint measures to event-based measurements was not yet considered, as event-based metrics are still quite rarely used in sound event detection due to the onset/offset uncertainty. Nevertheless, such extension is planned for future work, to evaluate localization and detection correctness of individual event instances.

[3] https://github.com/sharathadavanne/seld-metric

7. REFERENCES

[1] T. Butko, F. G. Pla, C. Segura, C. Nadeu, and J. Hernando, "Two-source acoustic event detection and localization: Online implementation in a smart-room," in *European Signal Processing Conference (EUSIPCO)*, 2011.

[2] R. Chakraborty and C. Nadeu, "Sound-model-based acoustic source localization using distributed microphone arrays," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[3] T. Hirvonen, "Classification of spatial audio location and content using convolutional neural networks," in *Audio Engineering Society Convention 138*, 2015.

[4] C. Grobler, C. Kruger, B. Silva, and G. Hancke, "Sound based localization and identification in industrial environments," in *IEEE Industrial Electronics Society (IECON)*, 2017.

[5] K. Lopatka, J. Kotus, and A. Czyzewsk, "Detection, classification and localization of acoustic events in the presence of background noise for acoustic surveillance of hazardous situations," *Multimedia Tools and Applications Journal*, vol. 75, no. 17, 2016.

[6] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48, March 2019.

[7] A. Mesaros, T. Heittola, and D. Ellis, *Datasets and Evaluation*. Cham: Springer International Publishing, 2018, pp. 147–179.

[8] S. Krstulovic, *Audio Event Recognition in the Smart Home*. Cham: Springer International Publishing, 2018, pp. 335–371.

[9] A. Mesaros, A. Diment, B. Elizalde, T. Heittola, E. Vincent, B. Raj, and T. Virtanen, "Sound event detection in the DCASE 2017 challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 6, pp. 992–1006, June 2019.

[10] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016. [Online]. Available: http://www.mdpi.com/2076-3417/6/6/162

[11] G. Poliner and D. Ellis, "A discriminative model for polyphonic piano transcription," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 1, p. 048317, 2007.

[12] E. D. Di Claudio and R. Parisi, "Multi-source localization strategies," in *Microphone Arrays*. Springer, 2001, pp. 181–201.

[13] H. W. Kuhn, "The hungarian method for the assignment problem," in *50 Years of Integer Programming*, 2010.

[14] D. Schuhmacher, B.-T. Vo, and B.-N. Vo, "A consistent metric for performance evaluation of multi-object filters," *IEEE transactions on signal processing*, vol. 56, no. 8, pp. 3447–3457, 2008.

[15] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The CLEAR MOT metrics," *EURASIP Journal on Image and Video Processing*, vol. 2008, no. 1, p. 246309, May 2008. [Online]. Available: https://doi.org/10.1155/2008/246309

[16] J. Bento and J. J. Zhu, "A metric for sets of trajectories that is practical and mathematically consistent," *arXiv preprint arXiv:1601.03094*, 2016.

[17] S. Adavanne, A. Politis, and T. Virtanen, "A multi-room reverberant dataset for sound event localization and detection," *arXiv preprint arXiv:1905.08546*, 2019.

[18] ——, "TAU Spatial Sound Events 2019 - Ambisonic and Microphone Array, Development Datasets," Feb. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.2599196

[19] ——, "TAU Moving Sound Events 2019 - Ambisonic, Anechoic, Synthetic IR and Moving Source Dataset," Apr. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.2636586

[20] ——, "Localization, detection and tracking of multiple moving sound sources with a convolutional recurrent neural network," *arXiv preprint arXiv:1904.12769*, 2019.

978-1-7281-1123-0/19/$31.00 ©2019 IEEE