LOCATE THIS, NOT THAT: CLASS-CONDITIONED SOUND EVENT DOA ESTIMATION

Olga Slizovskaia, Gordon Wichern, Zhong-Qiu Wang, Jonathan Le Roux

Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA

{wichern,leroux}@merl.com

ABSTRACT

Existing systems for sound event localization and detection (SELD) typically operate by estimating a source location for all classes at every time instant. In this paper, we propose an alternative class-conditioned SELD model for situations where we may not be interested in localizing all classes all of the time. This class-conditioned SELD model takes as input the spatial and spectral features from the sound file, and also a one-hot vector indicating the class we are currently interested in localizing. We inject the conditioning information at several points in our model using feature-wise linear modulation (FiLM) layers. Through experiments on the DCASE 2020 Task 3 dataset, we show that the proposed class-conditioned SELD model performs better in terms of common SELD metrics than the baseline model that locates all classes simultaneously, and also outperforms specialist models that are trained to locate only a single class of interest. We also evaluate performance on the DCASE 2021 Task 3 dataset, which includes directional interference (sound events from classes we are not interested in localizing) and notice especially strong improvement from the class-conditioned model.

Index Terms— Sound event localization and detection, conditioned neural networks, class-conditioned embeddings

1. INTRODUCTION

Estimating the direction of arrival (DOA) of a sound while also classifying the type of event is an important type of front-end processing for a wide variety of monitoring and robotic applications. Due to its wide applicability, this task, often referred to in the literature as sound event localization and detection (SELD), has recently seen a surge of interest [1–4]. However, SELD remains challenging because sound sources can move, cease to produce sound, have their positions obscured by room reverberation, and are often mixed with interfering sounds. Furthermore, many sound events are easily confused, further complicating the SELD task.

A typical SELD pipeline consists of two main stages. The first stage extracts spectral and spatial features from the microphone array input. In the second stage, a deep neural network (DNN) is used to learn a mapping from the multi-channel input features to two output targets: (1) a vector of
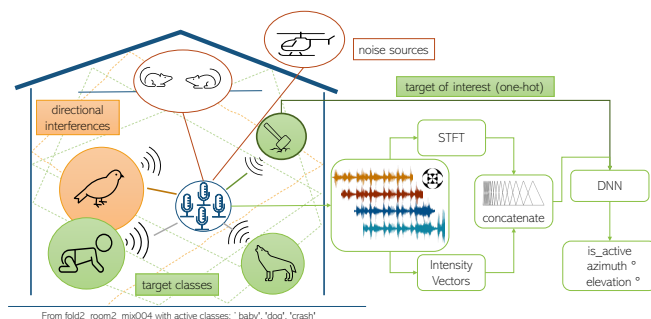
Fig. 1. Task illustration. Given a target-class identifier and a multi-channel signal recorded in conditions with multiple simultaneously active target sources, stationary noises, and directional interferences, we aim at estimating the DOA of the source of interest. Example taken from the training split of TAU-NIGENS SSE 2021 dataset.

class probabilities indicating the presence of each class in each time instance; and (2) a DOA vector containing location information for each class. In this case, DNN training can be challenging as the contributions from the two output branches must be balanced. To overcome this difficulty, Shimada et al. [5] recently proposed the activity coupled Cartesian direction of arrival (ACCDOA) representation. ACCDOA represents the DOA information for each class as a vector of Cartesian coordinates, where the class activity probability is encoded by the length of the vector, i.e., it should be one when that class is active and zero otherwise.

Existing SELD systems, including those based on the ACCDOA representation, typically assume a small, fixed set of sound event classes to be detected and localized. In practice, if there are 12 classes in the vocabulary, this means a DNN will output 12 ACCDOA representations at each time instant. This approach may become impractical for large class vocabularies, and, in certain situations, we may not be interested in localizing all classes all of the time. Training a class-specific system to localize only sound events from a single class would allow focusing on specific classes. However, this will then require C class-specific models as opposed to a single one, if we are interested in localizing C class types, and all models will also need to be run at inference time in order to detect all classes. Furthermore, it may be difficult to train class-specific

models, as we may not have enough data for each class of interest to properly train each model.

Inspired by the concepts of auditory attention [6,7], where the human auditory system tunes its internal models based on the sound sources it wants to listen to, in this paper we propose a class-conditioned approach for sound event DOA estimation, illustrated in Fig. 1. In our system, there is only a single ACCDOA vector output at each time instant, and the class represented by this output is determined based on an input describing the type of sound event we want to locate. Such conditioning-based approaches are often used in audio source separation [8] to isolate the sound from a specific musical instrument [9, 10], speaker [11, 12], or sound event [13]. Compared to a system that outputs all classes simultaneously, conditioning-based approaches require multiple passes at inference-time (one per class) in order to detect all classes, similarly to the class-specific models, although they do not require training multiple separate models. However, in situations where we may not want to detect all classes all of the time, they present several advantages. First, they can easily scale up to large numbers of classes without having to add more parameters (i.e., additional output layers), and they also lend themselves to query-by-example [14] or few-shot [15] settings where the conditioning input is not based on a class label embedding, but rather on an example of the type of sound we want to localize.

One other benefit of class-conditioned localization, which we demonstrate experimentally, is increased robustness to directional interference, i.e., sound events from classes we are not interested in localizing. We hypothesize that this increased robustness is a consequence of the class-conditioned training scheme, where all classes besides the one we are currently conditioning on are essentially treated as interfering sources, and this increased exposure allows the network to become better at ignoring directional interference. Through experiments on the DCASE 2020 Task 3 dataset [16], which does not contain directional interference, we demonstrate that our proposed class-conditioned model performs better in terms of a set of common SELD metrics than a model that outputs all classes, and a set of class-specific models. On the DCASE 2021 Task 3 dataset [17], which does contain directional interference, the class-conditioned model outperforms more clearly the all-class model, while the class-specific models fail dramatically.

The rest of this paper is organized as follows. Section 2 describes the baseline system, the class-specific SELD models, and our proposed class-conditioned SELD model. Experimental setup and evaluation results are presented and discussed in Section 3, followed by conclusions in Section 4.

2. METHOD

We first review the baseline SELD architecture with ACCDOA output representation, and then describe the proposed class-conditioned modification.

### 2.1. Baseline: All-class and Class-specific SELD models

The ACCDOA representation [5] uses Cartesian coordinates for the class-dependent azimuth and elevation angles, which are the estimation targets. Formally, let $\mathbf{d}_{t,c} = [x_{t,c}, y_{t,c}, z_{t,c}]$ represent the DOA for sound event class c at frame t. If no sound events from class c are active at frame t, $\|\mathbf{d}_{t,c}\|_2 = 0$, whereas if class c is active, $\|\mathbf{d}_{t,c}\|_2 = 1$. In the all-class baseline model with C classes of interest for a T-frame sound file, the tensor $\mathbf{D} \in \mathbb{R}^{T \times C \times 3}$ represents all ACCDOA vectors, and the goal is to estimate DNN parameters θ to estimate $\mathbf{D}$, i.e., $\hat{\mathbf{D}} = \mathcal{F}^{\mathrm{ac}}_{\theta}(\mathbf{X})$, where $\mathbf{X} \in \mathbb{R}^{M \times F \times T}$ represents the M-microphone, F-dimensional input features. Parameters θ are estimated by minimizing the mean squared error (MSE) between $\mathbf{D}$ and $\hat{\mathbf{D}}$. At inference time, class c is declared active at time t if $\|\hat{\mathbf{d}}_{t,c}\|_2 > \tau$, where τ is a threshold typically chosen to optimize performance on the validation set.

For the class-specific SELD models, the goal is to learn, for each class c, model parameters θ_c in order to obtain DOA estimates $\hat{\mathbf{D}}_c = \mathcal{F}^{\mathrm{cs}}_{\theta_c}(\mathbf{X}) \in \mathbb{R}^{T \times 3}$ for class c. Each θ_c is learned by minimizing the MSE between $\mathbf{D}_c = \mathbf{D}[:, c, :]$ and $\hat{\mathbf{D}}_c$. At inference time, we use the same threshold for all models.

2.2. Class-conditioned SELD model

One potential weakness of the baseline all-class SELD scheme is a high sparsity of the target representation tensor $\mathbf{D}$ and the need to predict all DOAs at once. In this case, we ask the network to not only focus on all locations but also identify all classes of interest, solving two different but interconnected problems. Instead, to solve the simpler task of predicting a DOA of a particular target class, it may be beneficial to focus the network on that class through an auxiliary conditioning input. Formally, we first define a C-dimensional one-hot vector $\mathbf{o}_c = [o_1, \ldots, o_C]$, where $o_c = 1$ for the class c we are interested in localizing, and $o_i = 0, \forall i \neq c$. Our goal is now to learn model parameters θ in order to estimate $\hat{\mathbf{D}}_c = \mathcal{F}^{\mathrm{cc}}_{\theta}(\mathbf{X}, \mathbf{o}_c)$, where $\hat{\mathbf{D}}_c \in \mathbb{R}^{T \times 3}$ represents the DOA estimates for class c only. Contrary to the class-specific models which each have their own parameter space, the class-conditioned model shares the same set of parameters for all classes. During training, we concatenate the outputs from evaluating $\mathcal{F}^{\mathrm{cc}}_{\theta}(\mathbf{X}, \mathbf{o}_c)$ for all C classes, such that we can obtain the full $\hat{\mathbf{D}} \in \mathbb{R}^{T \times C \times 3}$, and as in the all-class system we can learn θ by minimizing the mean squared error between $\mathbf{D}$ and $\hat{\mathbf{D}}$. Note that we have here C training examples for each multi-channel signal instead of just one, i.e., a training example is now a pair $(\mathbf{X}, \mathbf{o}_c)$.

Conditioning mechanisms, in particular feature-wise transformations, have proven to be a simple and effective technique for providing a model with context information.
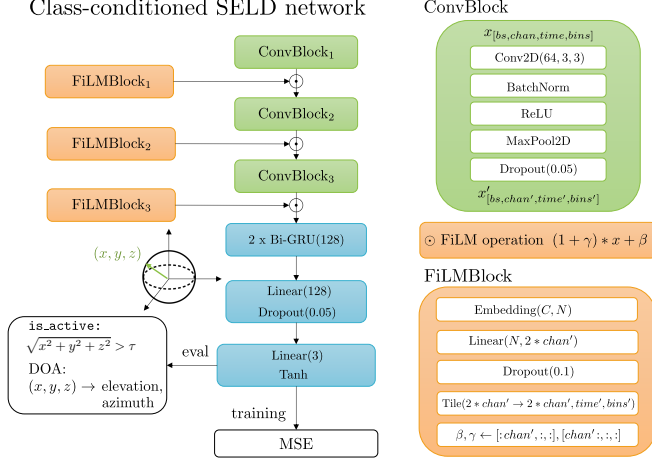
Fig. 2. Illustration of class-conditioned SELD architecture. Feature maps $x'_{[bs,chan',time',bins']}$ of each convolutional block of the baseline SELD model are shifted and scaled using a set of parameters (γ, β) learned from a conditioning one-hot vector.

For our class-conditioned SELD model, we employ the feature-wise linear modulation (FiLM) [18] scheme. We first pass the one-hot vector $\mathbf{o}_c$ through a learned embedding layer, to obtain a class embedding $\mathbf{e} \in \mathbb{R}^N$ that encodes the target class information. A FiLM layer f outputs a set of parameters $(\gamma, \beta) = f(\mathbf{e})$ that are used for scaling and shifting the learned feature maps $\mathbf{M}$ of the baseline network: $\text{FiLM}(\mathbf{M}_i | \gamma_i, \beta_i) = \gamma_i \mathbf{M}_i + \beta_i$, where i refers to a channel index. As shown in Fig. 2, we apply the FiLM operation to the feature maps of each convolutional block in the baseline SELD network. While a FiLM transformation can consist of any arbitrary function, in practice, we use a combination of an embedding, linear, and dropout layers as shown in Fig. 2.

3. EXPERIMENTAL VALIDATION

3.1. Experimental Setup

We evaluate our method on two official development datasets for the DCASE SELD task: TAU-NIGENS Spatial Sound Events 2020 [16] and TAU-NIGENS Spatial Sound Events 2021 [17]. For both datasets, we follow the official split into training (400 minutes), validation (100 minutes), and test (100 minutes) sets. The two key differences between the datasets are a reduced number of classes and the presence of directional interferences in the 2021 edition. The TAU-NIGENS 2021 dataset consists of mixtures of sound events belonging to 12 different classes, while the TAU-NIGENS 2020 dataset includes two more target classes, namely "engine" and "fire." As reported in [19], the 2021 edition has more simultaneously active events (up to three parallel tracks of different or same categories), and additional directional

interferences from non-target classes. In all experiments, we use a recommended set of features extracted from the first-order ambisonic (FOA) data split that includes 4-channel 64-dimensional mel-spectrograms and 3-channel FOA intensity vectors.

Our network architecture follows the convolutional recurrent architecture used as the DCASE SELD baseline [19] and shown in Fig. 2, but the FiLM layers are only included in the class-conditioned models. Both the class-conditioned and class-specific models have only a single ACCDOA output, while the all-class baseline has C ACCDOA outputs. We train all models for 50 epochs (note that an epoch is C times larger for the class-conditioned model, because each class and sound file pair counts as a data sample as discussed in Section 2.2) using the ADAM optimizer with a learning rate of $1e^{-3}$. We note that our goal for the experiments in this section is not to explore the data augmentation [20] and model ensembles [21] necessary to obtain the best scores on the DCASE SELD datasets, but rather explore the impact of class-conditioned SELD models in a controlled manner.

Following the DCASE SELD setup, we use four metrics and an aggregated score to evaluate different aspects of each SELD system. The four metrics are (1) the location-sensitive error rate $ER_{20°}$ that only accepts the predicted event as true positive if its location is at most $20°$ away from the ground truth and accounts for substitutions, deletions, and insertions from other classes, (2) the location-sensitive F-measure $F_{20°}$, (3) the class-sensitive localization error LE_{CD} that is computed separately for each class, and (4) the class-sensitive localization recall LR_{CD}. The aggregated SELD score [22] provides a simplified metric for the overall evaluation and is computed as follows:

$$SELD = \frac{ER_{20°}}{4} + \frac{1 - F_{20°}}{4} + \frac{LE_{CD}}{4 \times 180} + \frac{1 - LR_{CD}}{4}. \quad (1)$$
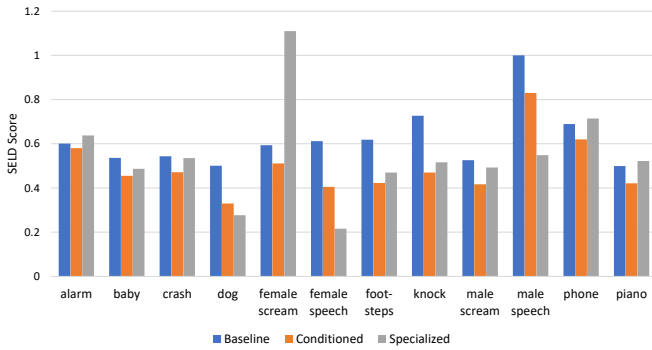
The ACCDOA format uses a norm of the DOA vector $\|\mathbf{d}_{t,c}\|_2$ to decide if class c is active at frame t or not. The default threshold τ on the norm is set to 0.5. However, the optimal choice of the threshold is a practical issue that is raised in [21]. Our empirical evaluation shows that the difference in SELD scores can be as large as 0.04 points depending on the threshold. For a fair comparison, we select an optimal threshold in terms of the SELD score for each experiment based on the validation set, and report the best results that a system can achieve together with the chosen threshold.

3.2. Results

Results for the baseline all-class SELD network, the class-specific models, and the proposed class-conditioned SELD model are presented in Table 1. As the original baseline is allegedly not fully reproducible, we ran it six times with different random seeds for the DCASE 2021 dataset to make sure there was not too much variance in the results, and report averaged performance. As we found that the measured

Table 1. Experimental results for baselines and the proposed system evaluated on the testing fold of the development datasets for DCASE 2021 Task 3 and DCASE 2020 Task 3.

	DCASE 2020						DCASE 2021					
System	τ	$ER_{20°}\downarrow$	$F_{20°}\uparrow$	$LE_{CD}\downarrow$	$LR_{CD}\uparrow$	$SELD\downarrow$	τ	$ER_{20°}\downarrow$	$F_{20°}\uparrow$	$LE_{CD}\downarrow$	$LR_{CD}\uparrow$	$SELD\downarrow$
Baseline FOA [17]	0.4	0.75	37.50	23.30	53.00	0.49	0.3	0.71	36.24	23.20	47.34	0.50
Specialized models	0.5	0.79	35.10	22.30	46.90	0.52	0.5	2.22	21.70	23.20	54.40	0.90
Class-conditioned SELD	0.3	0.72	47.90	17.80	55.90	0.45	0.3	0.64	49.72	19.16	56.60	0.42

Fig. 3. Comparison of per-class performance for different models on TAU-NIGENS 2021 data.

Fig. 4. Per-class comparison of the number of frames estimated as substitutions, deletions, and insertions for the baseline (B), class-conditioned (C) and specialized (S) systems on the DCASE 2021 Task 3 dataset.

standard deviation on the DCASE 2021 data was in fact low, we only report the results from a single run for the DCASE 2020 dataset. For the class-specific models, to obtain the aggregated scores shown in Table 1, we collect predictions from the model for each class and compare them against the ground truth. Similarly, we collect predictions of the proposed class-conditioned SELD model ran with class-conditionings for each class.

When comparing performance between the two baselines and the proposed class-conditioned SELD model, we observe that the class-conditioned model outperforms both baselines in terms of the SELD score and all individual metrics on both the DCASE 2021 Task 3 and DCASE 2020 Task 3 datasets. The relative reduction in terms of the SELD score is 16% for the DCASE 2021 dataset and 8.2% for the DCASE 2020 dataset compared to the baseline. We attribute a higher gain obtained for the DCASE 2021 data to the fact that the class-conditioned model is more robust to directional interferences which are present in the TAU-NIGENS 2021 dataset but not present in the TAU-NIGENS 2020 dataset.

Figure 3 illustrates per-class performance in terms of the SELD score for different systems on the DCASE 2021 Task 3 dataset. Interestingly, despite the fact that the aggregated SELD metrics for the specialized models demonstrate worse results, we still observe that class-specific systems can yield better scores than both the baseline and class-conditioned models for "dog," "female speech," and "male speech" categories. To further investigate this issue, we consider individual substitution, deletion, and insertion scores for each system and class. Figure 4 shows that the specialized models have a significantly higher number of insertions compared to the baseline and class-conditioned systems. While the effect
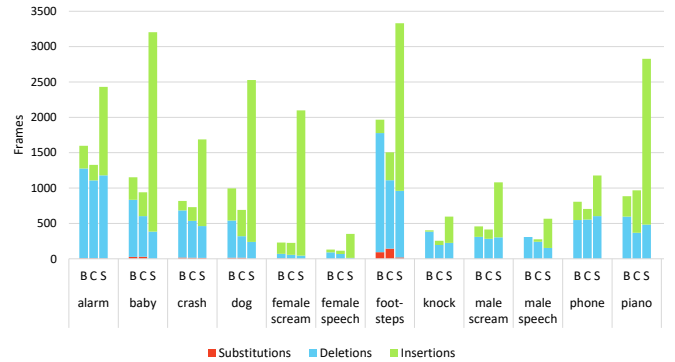
of insertions on individual per-class SELD scores is less noticeable in Fig. 3 (with the only distinct exception being the "female scream" class), the accumulated effect of all insertions makes the location-sensitive error rate three times worse compared to the baseline. We also note that the specialized model performance in Table 1 for the DCASE 2020 dataset does not exhibit the same large drop observed for the DCASE 2021 dataset. Since the DCASE 2021 dataset contains more overlapping sources, training a model that only uses labels for a single-class may be difficult if that class is rarely observed in isolation.

4. CONCLUSION

In this study, we focus on sound event detection and localization in the presence of directional interferences. We propose a class-conditioned SELD network that outperforms the baseline on the DCASE 2020 Task 3 and DCASE 2021 Task 3 datasets. The proposed system demonstrates greater improvement for the DCASE 2021 Task 3 dataset, supporting our hypothesis that a conditioned network is more robust against directional interferences such as those included in this dataset. The class-conditioned SELD network can be easily extended to accommodate a larger number of target categories which is one of the possible directions for future research. Other possible extensions of this work include incorporation of data augmentation techniques as proposed in [20], zero-shot sound event localization with pre-trained class-specific embeddings, as well as integration of a source separation module as proposed in [23] or a source counting module.

5. REFERENCES

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

[2] A. Politis, A. Mesaros, S. Adavanne, T. Heittola, and T. Virtanen, "Overview and evaluation of sound event localization and detection in DCASE 2019," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 684–698, 2021.

[3] C. Evers, H. W. Löllmann, H. Mellmann, A. Schmidt, H. Barfuss, P. A. Naylor, and W. Kellermann, "The LOCATA challenge: Acoustic source localization and tracking," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1620–1643, 2020.

[4] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "A survey of sound source localization with deep learning methods," *arXiv preprint arXiv:2109.03465*, 2021.

[5] K. Shimada, Y. Koyama, N. Takahashi, S. Takahashi, and Y. Mitsufuji, "ACCDOA: Activity-coupled Cartesian direction of arrival representation for sound event localization and detection," in *Proc. ICASSP*, Jun. 2021, pp. 915–919.

[6] J. B. Fritz, M. Elhilali, S. V. David, and S. A. Shamma, "Auditory attention—focusing the searchlight on sound," *Current opinion in neurobiology*, vol. 17, no. 4, pp. 437–455, 2007.

[7] N. Huang and M. Elhilali, "Push-pull competition between bottom-up and top-down auditory attention to natural soundscapes," *ELife*, vol. 9, p. e52984, 2020.

[8] P. Seetharaman, G. Wichern, S. Venkataramani, and J. Le Roux, "Class-conditional embeddings for music source separation," in *Proc. ICASSP*, May 2019, pp. 301–305.

[9] G. Meseguer-Brocal and G. Peeters, "Conditioned-U-Net: Introducing a control mechanism in the U-Net for multiple source separations," in *Proc. ISMIR*, Nov. 2019.

[10] O. Slizovskaia, G. Haro, and E. Gomez Gutierrez, "Conditioned source separation for musical instrument performances," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2083–2095, 2021.

[11] M. Delcroix, K. Zmolikova, K. Kinoshita, A. Ogawa, and T. Nakatani, "Single channel target speaker extraction and recognition with speaker beam," in *Proc. ICASSP*, Apr. 2018, pp. 5554–5558.

[12] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, "VoiceFilter: Targeted voice separation by speaker-conditioned spectrogram masking," in *Proc. Interspeech*, Sep. 2019.

[13] T. Ochiai, M. Delcroix, Y. Koizumi, H. Ito, K. Kinoshita, and S. Araki, "Listen to what you want: Neural network-based universal sound selector," in *Proc. Interspeech*, Oct. 2020.

[14] J. H. Lee, H.-S. Choi, and K. Lee, "Audio query-based music source separation," in *Proc. ISMIR*, Nov. 2019.

[15] Y. Wang, J. Salamon, N. J. Bryan, and J. P. Bello, "Few-shot sound event detection," in *Proc. ICASSP*, May 2020, pp. 81–85.

[16] A. Politis, S. Adavanne, and T. Virtanen, "TAU-NIGENS Spatial Sound Events 2020," May 2020. [Online]. Available: https://doi.org/10.5281/zenodo.4064792

[17] ——, "TAU-NIGENS Spatial Sound Events 2021," Feb. 2021. [Online]. Available: https://doi.org/10.5281/zenodo.4844825

[18] V. Dumoulin, E. Perez, N. Schucher, F. Strub, H. de Vries, A. Courville, and Y. Bengio, "Feature-wise transformations," *Distill*, 2018, https://distill.pub/2018/feature-wise-transformations.

[19] A. Politis, S. Adavanne, D. Krause, A. Deleforge, P. Srivastava, and T. Virtanen, "A dataset of dynamic reverberant sound scenes with directional interferers for sound event localization and detection," *arXiv preprint arXiv:2106.06999*, 2021.

[20] Q. Wang, J. Du, H.-X. Wu, J. Pan, F. Ma, and C.-H. Lee, "A four-stage data augmentation approach to ResNet-Conformer based acoustic modeling for sound event localization and detection," *arXiv preprint arXiv:2101.02919*, 2021.

[21] "Ensemble of ACCDOA- and EINV2-based systems with D3Nets and impulse response simulation for sound event localization and detection."

[22] A. Mesaros, S. Adavanne, A. Politis, T. Heittola, and T. Virtanen, "Joint measurement of localization and detection of sound events," in *Proc. WASPAA*, Oct. 2019, pp. 333–337.

[23] M. Yasuda, Y. Koizumi, S. Saito, H. Uematsu, and K. Imoto, "Sound event localization based on sound intensity vector refined by DNN-based denoising and source separation," in *Proc. ICASSP*, May 2020, pp. 651–655.