SPATIAL DATA AUGMENTATION WITH SIMULATED ROOM IMPULSE RESPONSES FOR SOUND EVENT LOCALIZATION AND DETECTION

Yuichiro Koyama⋆, *Kazuhide Shigemi*†∗, *Masafumi Takahashi*⋆, *Kazuki Shimada*⋆,
Naoya Takahashi⋆, *Emiru Tsunoo*⋆, *Shusuke Takahashi*⋆, *Yuki Mitsufuji*⋆

⋆ Sony Group Corporation, Tokyo, Japan †The University of Tokyo, Japan

ABSTRACT

Recording and annotating real sound events for a sound event localization and detection (SELD) task is time consuming, and data augmentation techniques are often favored when the amount of data is limited. However, how to augment the spatial information in a dataset, including unlabeled directional interference events, remains an open research question. Furthermore, directional interference events make it difficult to accurately extract spatial characteristics from target sound events. To address this problem, we propose an impulse response simulation framework (IRS) that augments spatial characteristics using simulated room impulse responses (RIR). RIRs corresponding to a microphone array assumed to be placed in various rooms are accurately simulated, and the source signals of the target sound events are extracted from a mixture. The simulated RIRs are then convolved with the extracted source signals to obtain an augmented multi-channel training dataset. Evaluation results obtained using the TAU-NIGENS Spatial Sound Events 2021 dataset show that the IRS contributes to improving the overall SELD performance. Additionally, we conducted an ablation study to discuss the contribution and need for each component within the IRS.

Index Terms— Sound event localization and detection, deep neural networks, data augmentation, room impulse responses

1. INTRODUCTION

Sound event localization and detection (SELD) involves identifying both the direction of arrival (DOA) and the type of sound [1–5]. As many combinations of DOA and types of sound are possible, recording and annotating real sound events for a SELD task is time consuming. Therefore, numerous methods have been utilizing data augmentation techniques based on given datasets [6–9].

The multi-channel simulation framework (MCS), which convolves extracted covariance matrices with enhanced source signals, enables us to create new combinations by randomly combining extracted spectral information and spatial information [9]. Compared with other spatial data augmentations such as the rotation augmentation method [10], the framework can change both the directional information and the reverberation pattern of sound events. Furthermore, the MCS does not need clean sources since it extracts enhanced source signals. Experimental results obtained with the TAU-NIGENS Spatial Sound Events 2020 dataset, consisting of only target sound events and diffusive background noise, showed that the MCS helps to improve the SELD performance [3, 9].

Directional interference events, which are sound events not included in target sound event classes, can be recorded unintentionally during the recording process because it is generally difficult to perfectly control all sound events generated in a real environment. Contamination by such interference events decreases the number of events that are available as non-overlapped events, i.e., clean events. As a result, extracting spatial information becomes challenging, and this could lead to degradation in the performance of existing data augmentation methods such as the MCS. Therefore, we assume that generating the target spatial information with an acoustic simulation could improve the augmentation quality.

In this paper, we propose an impulse response simulation framework (IRS) that augments spatial characteristics using simulated room impulse responses (RIR), which is not affected by directional interference events. RIRs corresponding to a microphone array assumed to be placed in various rooms are accurately simulated using image source methods [11] and spherical-harmonic-domain representation of its frequency response [12, 13]. The source signals of the target sound events are extracted from a mixture on the basis of annotated information and a proposed interference elimination process. The simulated RIRs are then convolved with the extracted source signals to obtain an augmented multi-channel training dataset. Evaluation results obtained using the TAU-NIGENS Spatial Sound Events 2021 dataset [9] show that the IRS contributes to improving the SELD score and the need for each component within the IRS. We also show that the IRS is usable with other typical data augmentation techniques and contributes to achieving state-of-the-art performance.

2. RELATED WORK

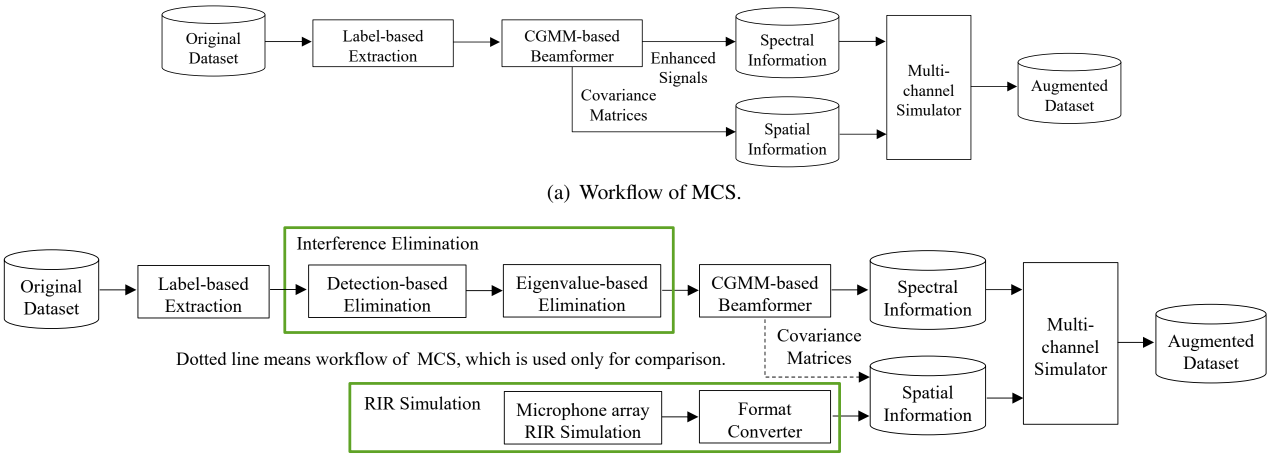
2.1. Supervised approaches for SELD

The primary problem in the SELD task is how to associate sound event detection (SED) predictions with DOA predictions (or vice versa) when multiple sound events overlap, which is called the data association problem [1]. To solve this problem, supervised approaches using deep neural network such as the convolutional recurrent neural network (CRNN) are typically used.

Spectral information, e.g., multi-channel spectrogram, log-mel spectrogram, and spatial information, e.g., GCC-PHAT, inter-channel phase differences (IPDs), are combined and used as input features. These features are fed into several two-dimensional convolutional layers followed by recurrent layers such as a gated recurrent unit (GRU). For instance, the assigning of a densely connected dilated DenseNet (D3Net) [14] to the convolutional layers achieved state-of-the-art performance in previous DCASE challenges [8, 15]. Finally, the output of the recurrent layers is transformed by fully-connected layers into an output representation. While two-branch representation, e.g., SELDnet [1], uses two branches for two targets, an SED target and a DOA target, activity-coupled Cartesian

∗Work done during an internship at Sony Group Corporation.

978-1-6654-0540-9/22/$31.00 ©2022 IEEE



(a) Workflow of MCS.

(b) Workflow of IRS. Dotted line represents workflow of MCS with interference elimination block, which is used only for comparison in section 4.

Fig. 1. Comparison of MCS and IRS. Interference elimination block and RIR simulation block are main differences between MCS and IRS. Interference elimination block is necessary for dealing with dataset including directional interference events.

DOA (ACCDOA) representation assigns a sound event activity to the length of a corresponding Cartesian DOA vector [8]. The ACCDOA representation enables us to handle a SELD task as a single task with a single network.

2.2. Data augmentation

Some data augmentation techniques have been widely used for the SELD task. EMDA [16, 17] mixes audio events with random amplitudes, delays, and the modulation of frequency characteristics, i.e., equalization. The rotation augmentation method [10] rotates an observed signal represented in the first-order Ambisonics (FOA) format and enables us to increase the number of DOA labels without losing the physical relationships between steering vectors and observations. SpecAugment [18], which was originally proposed for the speech recognition task and recently also showed efficacy on the SELD task [8], applies frequency masking to input features.

In addition to these augmentation methods, the MCS is also effective for creating new combinations of spectral information and spatial information [9]. Fig. 1(a) describes the workflow of the MCS. First, non-overlapping and static, i.e., not moving, events are extracted from an original dataset on the basis of label information. Then, the complex Gaussian mixture model (CGMM)-based beamformer [19] block takes the extracted events and outputs enhanced signals and covariance matrices. The covariance matrices are calculated with the spectrograms enhanced by CGMM-based masking. The enhanced signals are stored as spectral information, and the covariance matrices are stored as spatial information. They are independently and randomly sampled to create new combinations of spectral information and spatial information. The multi-channel simulator block finally takes them and simulates multi-channel input signals.

The MCS actually contributed to improving the performance on the TAU-NIGENS Spatial Sound Events 2020 dataset [3, 9], which does not include directional interference events. However, it assumes no directional interference events and highly relies on non-overlapping events (i.e., clean events) in the original dataset. Therefore, using the MCS for datasets including directional interference events such as the TAU-NIGENS Spatial Sound Events 2021 dataset could lead to performance degradation.

3. PROPOSED METHOD

Inspired by the MCS, we propose the IRS, which augments spatial characteristics using RIRs. Fig. 1(b) describes the workflow of the IRS. The interference elimination block and RIR simulation block are the main differences from the MCS. The interference elimination block is necessary for dealing with datasets including directional interference events such as the TAU-NIGENS Spatial Sound Events 2021 dataset [5]. Also, we assume that the covariance matrix of an event extracted from the original dataset is still not clean enough for use as spatial information even if the interference elimination block is applied. Therefore, we simulate RIRs, which are used for spatial information.

3.1. Interference elimination

The interference elimination block is composed of two blocks: the detection-based elimination block and eigenvalue-based elimination block. The motivation for using two types of elimination block is to eliminate events with different ranges of signal-to-interference ratio (SIR).

First, the detection-based elimination block uses a model pre-trained without the IRS, whose performance is shown as ID 0 in Table 1 . The non-overlapping static events extracted by the label-based extraction block are processed with the pre-trained model. Events that the pre-trained model cannot detect are regarded as events overlapped with interference events, which are eliminated.

Then, the eigenvalue-based elimination block first applies a short-time Fourier transform (STFT) to the events extracted by the detection-based elimination block and computes a spatial covariance matrix from the obtained spectrogram. Its eigenvalues are then computed and normalized such that the maximum eigenvalue becomes 1. These eigenvalues can be considered to reflect how many sound sources the observation signals include. Let $\gamma_c (c = 1, \ldots, C)$ be the normalized eigenvalues, where C is the number of input channels. We introduce two types of thresholds, α and β. We define the frequency bin where the number of eigenvalues that satisfy $\gamma_c > \alpha$ is more than 1 as an *overlapped bin*. Then, focusing on the limited range from the minimum frequency f_{min} to maximum frequency f_{max}, let K_{focus} be the number of total focused bins and

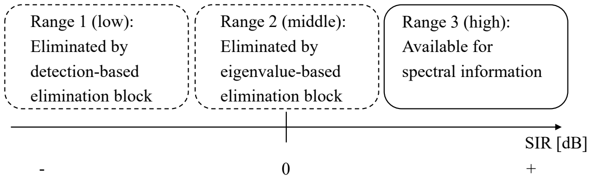


Fig. 2. Assumption in terms of SIR. We propose using events from range 3 for spectral information.

K_{overlap} be the number of *overlapped bins*. Extracted events that satisfy $(K_{\text{overlap}}/K_{\text{focus}}) > \beta$ are regarded as events overlapped by interference events, which are also eliminated.

Fig. 2 illustrates our assumption regarding the relationship between the interference elimination block and SIR, where we divide the SIR range into three levels: “range 1 (low),” “range 2 (middle),” and “range 3 (high).” We assume that the events in range 1 shown in Fig. 2 can be eliminated by the detection-based elimination block because the events here are heavily overlapped by interference events; thus, the pre-trained model cannot detect such events. The events in range 2 are expected to be eliminated by the eigenvalue-based elimination block because the power of a target sound event and that of an interference event are almost the same and $K_{\text{overlap}}/K_{\text{focus}}$ tends to be higher. The events in range 3 are fed into the CGMM-based beamformer block, and the SIRs are improved by the beamformer. As a result, the enhanced signals are clean enough to be used as spectral information.

3.2. RIR simulation

The RIR simulation block is composed of two blocks, the microphone-array RIR simulation block and format converter block, by which spatial aliasing caused by the discrete sampling of a sound field can be simulated to reflect the actual recording condition.

In the microphone-array RIR simulation block, an image source method [11] is utilized to define the positions for all image sources depending on given reverberant conditions. RIRs are obtained as linear combinations of the frequency responses H for all image sources. We assume that the array can be regarded as a rigid spherical array [20, 21]. By considering the waves from all image sources as plane waves, the frequency response of h-th microphone with a wave number of k on a rigid baffle of radius R for l-th image source is obtained as

$$H_{hl}(k,\psi_{hl}) = g_{hl}e^{-\mathrm{i}kd_{hl}} \sum_{n=0}^{\infty} \mathrm{i}^n (2n+1) b_n\left(kR\right) P_n\left(\cos\psi_{hl}\right), \tag{1}$$

where ψ_{hl} denotes the angle between the DOA of the l-th plane wave and the orientation of the h-th microphone, $P_n(\cdot)$ denotes the Legendre polynomial [20, 22], g_{hl} and d_{hl} denote attenuation factor and delay time respectively, both of which are caused by the number of reflections and distance. The imaginary unit is denoted by i. The b_n is a radial function for a rigid baffle array written as

$$b_n(kR) = \frac{\mathrm{i}}{(kR)^2 h_n^{(1)'}(kR)}, \tag{2}$$

where $h_n^{(1)'}(\cdot)$ denotes the derivative of the n-th-order spherical Hankel function of the first kind. Computing the linear combinations of $H_{hl}(k,\psi_{hl})$ in terms of all image sources, the simulated RIR of h-th microphone can be obtained as

$$x_h(k) = \sum_{l} H_{hl}(k,\psi_{hl}). \tag{3}$$

Then, in the format converter block, the simulated RIRs are converted to the intended format, e.g., higher-order Ambisonics (HOA). The n-th-order and m-th-degree spherical harmonic function is defined with the angle $\Omega = \{\theta, \phi\}$ [22] as

$$Y_{nm}(\Omega) \equiv \sqrt{\frac{2n+1}{4\pi}\frac{(n-m)!}{(n+m)!}} P_{nm}(\cos\theta)e^{\mathrm{i}m\phi}, \tag{4}$$

where θ, ϕ, and $P_{nm}(\cdot)$ denote the elevation, azimuth, and associated Legendre function, respectively. The spherical-harmonic representation of the RIR can be computed by using the following encoding process [12, 13]:

$$\mathbf{a}(k) = \mathbf{B}(k)^{-1}\mathbf{Y}^{\dagger}\mathbf{x}(k), \tag{5}$$

with

$$\mathbf{B}(k) = \begin{pmatrix} b_0 & 0 & 0 & 0 & \dots & 0 \\ 0 & b_1 & 0 & 0 & \dots & 0 \\ 0 & 0 & b_1 & 0 & \dots & 0 \\ 0 & 0 & 0 & b_1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & b_N \end{pmatrix}, \tag{6}$$

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}(\Omega_1) & \mathbf{y}(\Omega_2) & \cdots & \mathbf{y}(\Omega_M) \end{bmatrix}^T, \tag{7}$$

where $\mathbf{y}(\Omega) \in \mathbb{C}^{(N+1)^2}$ is a column vector containing $Y_{nm}(\Omega)$, N is the spherical-harmonic order, e.g., $N = 1$ corresponds to FOA, M is the number of microphones, $\mathbf{x}(k) \in \mathbb{C}^M$ is a column vector containing $x_h(k)$, and $(\cdot)^{\dagger}$ represents the Moore–Penrose pseudoinverse.

4. EXPERIMENT

4.1. Experimental settings

We evaluated our approach on the development set of the TAU-NIGENS Spatial Sound Events 2021 dataset using the suggested setup [5]. The dataset was composed of 6 folds, which contained 600 one-minute Eigenmike recordings with the FOA format: 400 for training (folds 1 to 4), 100 for validation (fold 5), and 100 for testing (fold 6). The sound event samples were from the NIGENS general sound events database [23], which consists of 12 event classes such as footsteps and barking dog. Each event had an equal probability of being either static or moving. The signal-to-noise ratios ranged from 6 dB to 30 dB. The sampling frequency was 24 kHz.

We prepared two types of CRNNs: the conventional CRNN used in [2] and RD3Net [14]. Both networks have four convolutional blocks followed by a GRU. They take frame-wise multi-channel amplitude spectrograms and inter-channel phase differences (IPDs) as frame-wise features and output frame-wise ACCDOA. The STFT was applied with a configuration having a 20-ms frame length and 10-ms frame hop.

The IRS was applied to the training data by adding new training folds created by the IRS. In this experiment, we added 2 folds, which increased the amount of training data by 50%. The thresholds for our interference elimination block, α and β, were 0.3 and 0.4 respectively. The f_{min} and f_{max} were 100 Hz and 4 kHz, respectively. A minimum variance distortionless response (MVDR) beamformer [24] was applied to the CGMM-based beamformer block. In the RIR simulation block, the reverberation time (RT60) was randomly set to be within 100 to 500 ms. We only simulated static

Table 1. Experimental results with CRNN-based architecture on TAU-NIGENS Spatial Sound Events 2021 dataset. D-based and E-based represents detection-based elimination and eigenvalue-based elimination respectively. CM stands for covariance matrix.

ID	IRS	Interference elimination		Spatial	Other	Metrics in test set				
		D-based	E-based	information	augmentation	ER$_{20}$ ↓	F$_{20}$ ↑	LE$_{\mathrm{CD}}$ ↓	LR$_{\mathrm{CD}}$ ↑	SELD$_{\mathrm{score}}$ ↓
0	None				None	0.68	41.2	18.1	40.8	0.489
1	✓(MCS)	None	None	CM	None	0.66	43.6	19.0	44.4	0.471
2	✓(MCS)	✓	✓	CM	None	0.64	45.6	18.0	46.1	0.457
3	✓	None	None	RIR	None	0.64	46.6	17.2	44.9	0.455
4	✓	✓	None	RIR	None	0.64	46.4	18.3	45.7	0.455
5	✓	None	✓	RIR	None	0.63	47.7	17.7	47.1	0.445
6	✓	✓	✓	RIR	None	0.62	49.2	16.8	48.1	**0.436**
7	None				✓	0.56	55.0	18.4	59.5	0.380
8	✓(MCS)	None	None	CM	✓	0.58	53.6	18.0	59.4	0.386
9	✓(MCS)	✓	✓	CM	✓	0.57	54.9	18.1	59.4	0.381
10	✓	None	None	RIR	✓	0.56	55.2	18.0	60.3	0.377
11	✓	✓	None	RIR	✓	0.56	56.1	17.3	60.2	0.372
12	✓	None	✓	RIR	✓	0.54	56.0	16.2	57.0	0.376
13	✓	✓	✓	RIR	✓	0.54	57.8	17.0	61.5	**0.360**

Table 2. Performance of RD3Net

IRS	Metrics for test set				
	ER$_{20}$ ↓	F$_{20}$ ↑	LE$_{\mathrm{CD}}$ ↓	LR$_{\mathrm{CD}}$ ↑	SELD$_{\mathrm{score}}$ ↓
None [15]	0.48	64.1	13.2	63.2	0.321
✓	0.49	65.0	14.0	70.7	**0.302**

impulse responses. The SN3D normalization scheme of Ambisonics was used in the format converter block [3]. For comparison, the MCS was also applied by using covariance matrices for the spatial information instead of RIRs. In addition, we investigated the efficiency of combining the IRS and other augmentation techniques, that is, EMDA, rotation augmentation methods, and a multi-channel version of SpecAugment, whose efficacy were already shown in [15]. These augmentation techniques were applied on-the-fly [25].

The frame length of the network input during training was 128 frames except for the experiment to show the performance of RD3Net. The batch size for the training was 32. The learning rate was linearly increased from 0.0 to 0.001 with 50,000 iterations [26]. After the warm-up, the learning rate was decreased by 10% if the SELD score of the validation did not improve in 40,000 consecutive iterations. We used the Adam optimizer with a weight decay of 10^{-6}. We validated and saved model weights every 10,000 iterations up to 400,000 iterations. Finally, we averaged the model weights from the last 5 models as in [27].

Four metrics were used for the evaluation [28]: LE$_{\mathrm{CD}}$, LR$_{\mathrm{CD}}$, ER$_{20^\circ}$, and F$_{20^\circ}$. LE$_{\mathrm{CD}}$ is a localization error that indicates the average angular distance between predictions and references of the same class. LR$_{\mathrm{CD}}$ is a simple localization recall metric that expresses the true positive rate of how many of these localization predictions are correctly detected in a class out of the total number of class instances. ER$_{20^\circ}$ and F$_{20^\circ}$ are the location-dependent error rate and F-score, where predictions are considered as true positives only when the distance from the reference is less than 20°. To evaluate the overall performance, we adopted SELD$_{\mathrm{score}}$, which is defined as

$$\mathrm{SELD}_{\mathrm{score}} = [\mathrm{ER}_{20^\circ} + (1 - \mathrm{F}_{20^\circ}) + \mathrm{LE}_{\mathrm{CD}}/\pi + (1 - \mathrm{LR}_{\mathrm{CD}})]/4. \quad (8)$$

4.2. Results

Table 1 shows the experimental results for the IRS using CRNN. ID 0 represents the training method without any data augmentation techniques, and ID 7 represents that with other data augmentation techniques (i.e., EMDA, rotation augmentation method, multi-channel version of SpecAugment). IDs 1 and 8 are equivalent to the MCS, and IDs 2 and 9 are the MCS with the interference elimination block.

The SELD$_{\mathrm{score}}$ of ID 0, 3, 4, 5, and 6 shows that the IRS consistently improved the SELD$_{\mathrm{score}}$ and that using both the eigenvalue-based elimination block and detection-based elimination block was the most effective. In this experiment, the number of events extracted by the label-based extraction was 1243, the number of events eliminated by the eigenvalue-based elimination block was 244, and the number of events eliminated by the detection-based elimination block was 16. This result suggests that using the events from range 1 in Fig. 2 for the spectral information degrades the performance even if the number of events is relatively small. This is because such events have an extremely low SIR as shown in Fig. 2. Comparing IDs 1, 2, 3, and 6 (i.e., comparing the IRS with the MCS), it is shown that the RIR was more suitable for the spatial information than the covariance matrix calculated from the extracted events. This is because the events from range 3 in Fig. 2 still had interference events with low energy, which disturbed the spatial information of the target sound events. The results for IDs 7 to 13 show the same tendency as the results for IDs 0 to 6, which suggests that the IRS can be used with other augmentation techniques.

Table 2 shows the performance of RD3Net. We trained RD3Net with other augmentation techniques and 1024 input frames. It is shown that the IRS contributed to further improving the performance of the state-of-the-art method (i.e., RD3Net).

5. CONCLUSION

We proposed the impulse response simulation framework (IRS), which augments spatial characteristics using simulated room impulse responses. Experimental results obtained using the TAU-NIGENS Spatial Sound Events 2021 dataset indicated that each block in the IRS contributed to improving the SELD performance. In addition, it was shown that combining the IRS with other typical augmentation techniques lead to further improvement. Finally, RD3Net trained with the IRS achieved state-of-the-art performance on the SELD task. In future work, we will explore a new framework that does not depend on the label-based extraction block, which will help us to create new combinations of spectral and spatial information without supervised annotations.

6. 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 of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48, 2019.

[2] Y. Cao, Q. Kong, T. Iqbal, F. An, W. Wang, and M. Plumbley, "Polyphonic sound event detection and localization using a two-stage strategy," in *Proc. of the Detection and Classification of Acoustic Scenes and Events (DCASE) 2019 Workshop*, 2019, pp. 30–34.

[3] A. Politis, S. Adavanne, and T. Virtanen, "A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection," *arXiv preprint arXiv:2006.01919*, 2020.

[4] T. N. T. Nguyen, D. L. Jones, and W.-S. Gan, "A sequence matching network for polyphonic sound event localization and detection," in *Proc. of IEEE ICASSP*, 2020, pp. 71–75.

[5] A. Politis, S. Adavanne, D. Krause, A. Deleforge, P. Srivistavana, 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.

[6] Y. Cao, T. Iqbal, Q. Kong, F. An, W. Wang, and M. D. Plumbley, "An improved event-independent network for polyphonic sound event localization and detection," in *Proc. of IEEE ICASSP*, 2021, pp. 885–889.

[7] T. N. T. Nguyen, N. K. Nguyen, H. Phan, L. Pham, K. Ooi, D. L. Jones, and W.-S. Gan, "A general network architecture for sound event localization and detection using transfer learning and recurrent neural network," in *Proc. of IEEE ICASSP*, 2021, pp. 935–939.

[8] 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. of IEEE ICASSP*, 2021, pp. 915–919.

[9] 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.

[10] L. Mazzon, Y. Koizumi, M. Yasuda, and N. Harada, "First order ambisonics domain spatial augmentation for DNN-based direction of arrival estimation," in *Proc. of the Detection and Classification of Acoustic Scenes and Events (DCASE) 2019 Workshop*, 2019.

[11] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. of IEEE ICASSP*, 2018, pp. 351–355.

[12] S. Moreau, J. Daniel, and S. Bertet, "3D sound field recording with higher order Ambisonics–objective measurements and validation of a 4th order spherical microphone," in *Audio Engineering Society Convention*, 2006, pp. 20–23.

[13] A. Politis and H. Gamper, "Comparing modeled and measurement-based spherical harmonic encoding filters for spherical microphone arrays," in *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017, pp. 224–228.

[14] N. Takahashi and Y. Mitsufuji, "Densely connected multidilated convolutional networks for dense prediction tasks," in *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 993–1002.

[15] K. Shimada, N. Takahashi, Y. Koyama, S. Takahashi, E. Tsunoo, M. Takahashi, and Y. Mitsufuji, "Ensemble of ACCDOA-and EINV2-based Systems with D3Nets and Impulse Response Simulation for Sound Event Localization and Detection," *arXiv preprint arXiv:2106.10806*, 2021.

[16] N. Takahashi, M. Gygli, B. Pfister, and L. V. Gool, "Deep convolutional neural networks and data augmentation for acoustic event detection," in *Proc. of Interspeech*, 2016.

[17] N. Takahashi, M. Gygli, and L. Van Gool, "AENet: Learning deep audio features for video analysis," *IEEE Transactions on Multimedia*, vol. 20, pp. 513–524, 2017.

[18] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. of Interspeech*, 2019, pp. 2613–2617.

[19] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," in *Proc. of IEEE ICASSP*, 2016, pp. 5210–5214.

[20] A. Politis, "Microphone array processing for parametric spatial audio techniques," PhD thesis, Aalto University, 2016.

[21] Y. Mitsufuji, N. Takamune, S. Koyama, and H. Saruwatari, "Multichannel blind source separation based on evanescent-region-aware non-negative tensor factorization in spherical harmonic domain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 607–617, 2021.

[22] E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography*. Academic Press, 1999.

[23] I. Trowitzsch, J. Taghia, Y. Kashef, and K. Obermayer, "The NIGENS general sound events database," *arXiv:1902.08314*, 2019.

[24] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 260–276, 2010.

[25] H. Erdogan and T. Yoshioka, "Investigations on data augmentation and loss functions for deep learning based speech-background separation," in *Proc. of Interspeech*, 2018, pp. 3499–3503.

[26] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He, "Accurate, large minibatch sgd: Training imagenet in 1 hour," *arXiv preprint arXiv:1706.02677*, 2017.

[27] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, R. Yamamoto, X. Wang *et al.*, "A comparative study on transformer vs RNN in speech applications," in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 449–456.

[28] A. Mesaros, S. Adavanne, A. Politis, T. Heittola, and T. Virtanen, "Joint measurement of localization and detection of sound events," in *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019.