Data Diversity for Improving DNN-based Localization of Concurrent Sound Events

Daniel Krause
Faculty of Information Technology and Communication Sciences
Tampere University
Tampere, Finland
daniel.krause@tuni.fi

Archontis Politis
Faculty of Information Technology and Communication Sciences
Tampere University
Tampere, Finland
archontis.politis@tuni.fi

Konrad Kowalczyk
Faculty of Computer Science, Electronics and Telecommunications
AGH University of Science and Technology
Kraków, Poland
konrad.kowalczyk@agh.edu.pl

***Abstract*—Sound source localization (SSL) is an actively researched topic in the field of multichannel audio signal processing with numerous practical applications. Since it is used in different acoustic contexts, ensuring a good generalization of the techniques and models to various acoustic signals and environments is of great importance. In this paper, we aim to investigate the influence of different types of sources on the training process of a model based on a deep neural network (DNN). We present several training datasets, containing different mixtures of noise, speech and sound events, and perform a comparative study in which we test the trained models on distinct target signals. The problem is analyzed in the context of localizing two simultaneously active sound sources. Additionally, two data augmentation methods are incorporated into the framework to verify their impact on model generalization. The results of experiments performed for two concurrent sources show the localization accuracy of models trained with diverse data types. In particular, training using speech and sound event data or mixtures thereof are shown to result in localization accuracy increase for a variety of source types under test.**

***Index Terms*—Sound source localization, multiple source localization, deep neural networks**

I. Introduction

Localization and tracking of acoustic sources is a widely researched topic [1], with applications ranging from acoustic monitoring to signal enhancement. Traditionally, the signal model with geometric plane wave propagation model has been assumed, and the directions-of-arrival (DOAs) of the acoustic sources are found by estimating model parameters. Such approaches make minimal assumptions about the source signals and do not require training or supervision. Recently, data-based localization studies have received an increased interest, with promising results in adverse noise and reverberant conditions [2–6]. Early works have focused on localization of a single active source, with the problem formulated as regression [7] or classification [8]. Classification-based approaches, with output positional probabilities on a pre-defined grid, have an advantage of naturally handling multiple sources, but for a high directional resolution in both azimuth and elevation a large amount of classes is required, making the training cumbersome. Regression-based approaches require the network to have as many regressors as the maximum number of sources to be localized. The latter approach leverages from an external activity detector or post-processing to define which part of the output corresponds to the source activity. Furthermore, the DoA estimates of multiple sources can be permuted at the outputs. This can be alleviated to some extent using training strategies such as the permutation invariant training [9]. One aspect that so far has not received much attention, in contrast to the input features [10] or architectural and training loss choices, is the effect of the source signal on the accuracy of the DNN-based localization. In principle, localization exploits inter-channel signal properties, which are independent from the source signal, and thus the DNN-based localization should be able to generalize well to the sources that are different from the ones used in training. This principle has been exploited in the work of [3], where spatialized noise is used for convenient training and the performance of the system is evaluated on speech recordings. Nonetheless, the majority of DNN-based methods [2, 4] use matched training and testing, in which the same type of source is used.

In this paper, we investigate the differences in performance between the models trained with different types of source signals with the aim to find optimal solutions for distinct test scenarios. For this purpose, we divide sound sources into three main categories: noise, speech, and sound events. Noise signals avoid the need of sample gathering and are the most general non-informative signal type, providing a good potential for generalization [11]. Speech combines many properties that are shared amongst other signal types, containing both harmonic and noise-like features, whilst showing a transient character [12]. It is also the most common application scenario. Sound events are the most diverse type with all kinds of signal properties, for instance a class can be both noisy or correlated [13]. To enable realistic conditions, we synthesize the data by convolving the source signals with simulated impulse responses for different source-array distances, in rooms with different reverberation times [14]. In addition, we investigate whether data augmentation techniques can be used to alternatively improve the generalization of the model [15–17].

The final contribution of this paper is investigation of the robustness of sound source localization depending on the diversity of the training data for a scenario with two simultaneously active sources.

II. DNN MODEL FOR SOUND SOURCE LOCALIZATION

In this work, we localize two simultaneously active sound sources using a DNN model. In order to compare the localization accuracy for different types of training and test data, we apply the same Convolutional Recurrent Neural Network (CRNN) model across all performed experiments. The CRNN model has been shown in [2, 18] as well as in our previous studies [10, 19] to yield a good balance between model complexity and localization accuracy. The CRNN architecture adopted in this work is summarized in Table I. Note that we perform max pooling only across the frequency axis in order to preserve the temporal resolution of the input features. The final layer consists of 6 output units representing two DOA vectors, each containing three estimated Cartesian coordinates of the vector pointing towards the sound event. Note that the unit-normalized vectors are used during training. Finally, the output values are smoothed using a median filter over the 25 frame window. Regression loss is the well known mean squared error (MSE).

Similarly to speaker-independent source separation [9], source-independent localization for more than one source based on regression can suffer from the permutation problem, where the estimates may switch outputs from frame-to-frame or sequence-to-sequence, affecting both training convergence and overall performance. To avoid such permutation issues, we apply the so-called frame-level Permutation Invariant Training (tPIT) [20]. PIT simply compares the combined error of the outputs for all permutations of outputs and ground truths (2 in our case), and only the combination that produces the minimum error is selected to perform back-propagation during the training.

Models are trained using multichannel magnitude spectrograms and the sine & cosine values of the inter-channel phase differences, which have been shown to perform better than phase spectrograms and equally good to generalized cross-correlation features in similar settings [10]. In this work, the complex spectrum is computed using the Short-Time Fourier Transform with a Hamming window of $40ms$ length and 50% overlap. The phase features between the i-th and j-th microphone channels are computed as

$$\mathrm{SI}_{i,j}[n,k] = \sin\left(\mathrm{IPD}_{i,j}[n,k]\right), \tag{1}$$

$$\mathrm{CI}_{i,j}[n,k] = \cos\left(\mathrm{IPD}_{i,j}[n,k]\right). \tag{2}$$

The inter-channel phase differences (IPDs) for each (i,j) channel pair are defined as

$$\mathrm{IPD}_{i,j}[n,k] = \arg(X_i[n,k]) - \arg(X_j[n,k]). \tag{3}$$

where $X_i[n,k]$ denotes the complex spectrum value, $\arg(X_i[n,k])$ denotes phase of the spectrum, and n and k denote the time and frequency indices, respectively. Sines and cosines of IPDs have been firstly proposed for multichannel DNN-based speech separation [21] and further investigated for localization in [10].

TABLE I
CRNN MODEL ARCHITECTURE FOR THE LOCALIZATION OF CONCURRENT SOURCES.

Layer type	Description
2D CNN	64 filters, 3x3, BN, ReLU, MaxPooling x4
2D CNN	64 filters, 3x3, BN, ReLU, MaxPooling x4
2D CNN	64 filters, 3x3, BN, ReLU, MaxPooling x2
Bi-LSTM	64 units, tanh
Bi-LSTM	64 units, tanh
Fully connected	128 units, linear
Fully connected	6 units, linear

In this work, we aim to investigate which types of sound sources and data augmentation techniques are more effective when training the DNN-based model for the localization of simultaneously active sound sources.

III. DATA GENERATION

A. Various Types of Sound Events

To provide a credible comparison of different types of sources, sufficient amounts of data of a similar type are required. Hence, we create 7 sets of audio data, differing only in the type of sound source signals. The first five datasets are used for training, while the latter are used during the tests:

- **Noise-Train:** consists of random white Gaussian noise sources,
- **Speech-Train:** uses only speech signals from the TIMIT Acoustic-Phonetic Continuous Speech Corpus, [22]
- **Event-Train:** consists of 14 individual sound event classes from the NIGENS dataset [23],
- **SN-Train:** a mixture of random noise sources and speech data with 1:1 proportion,
- **SNE-Train:** a mixture of noise, speech and sound events signals at 1:1:1 proportions,
- **Event-Test:** a test set of the same type as Event-Train, but with a different set of room and DOA combinations,
- **Speech-Test:** a test set of the same type as Speech-Train, but with a different set of room and DOA combinations.

Each of the training datasets contains 800 audio files, each 40 seconds long with 24 kHz sampling frequency and 16 bit resolution. The test sets consist of 200 audio files each. The data is divided into 4 splits to enable fold-wise cross-validation. Data is created using 1 second long chunks for each sound source, always with two overlapping sources being active at the same time. In order to provide a possibly large and general representation of the localization scenarios, we simulate a random shoe-box room for each file separately. The randomized parameters are all physical dimensions of the room and the reverberation time (RT). To estimate the RT curve, we simply randomize the values for the 125Hz and 4kHz frequency bands (denoted as RT125 and RT4k, respectively) and perform a linear interpolation in between. Room impulse responses (RIRs) for a tetrahedral microphone array that consists of four cardioid microphones are simulated using the image source method [14]. For each room we pick a

TABLE II
RANDOMIZATION OF PARAMETERS FOR DATA GENERATION.

Parameter	Random range
Room width and length	[6.0 10.0] m
Room height	[2.5 6.0] m
RT125	[0.3 0.9] s
RT4k	[0.8 0.98] * RT125
Receiver position (x, y)	[20.0 40.0] %
Receiver height	[1.0 2.0] m
Source distance	[1.5 5.0] m
Azimuth angle	[-180.0 180.0] °
Elevation angle	[-90.0 90.0] °

random position of the receiver, whereas each individual sound source is located at a position computed by randomly selecting the azimuth and elevation angles as well as the distance from the array. The receiver position is expressed as the percentage of the room's width and depth. We also ensure that each source and receiver are located at least 1 m apart from the walls, ceiling, and floor. The randomized parameters are the same for all datasets and are summarized in Table II. Finally, audio files are synthesized by convolving the signals of each source with the respective RIRs.

B. Data Augmentation

To answer the question about the actual gain achieved by utilizing more specific sound sources in the training process, it should be determined whether this approach shows any improvement over other data enhancing techniques. Therefore, apart from comparing different types of data, we study the usage of two data augmentation techniques with the aim to establish if they constitute a valuable addition to the target data types. These methods have been successfully used in monophonic applications, therefore we straightforwardly extend them to multichannel data. In this study, two augmentation techniques are exploited:

- **Background noise (BkgN):** random chunks of background noise from the DESED dataset are used in random places of the audio files [24] with a level of -5dB, -10dB or -15dB relative to the power of the original signal.
- **SpecAugment:** first proposed for automatic speech recognition [25], this method exploits the masking with zero values in the time and frequency domain. In this paper, we mask 15% of the feature input matrix.

IV. EXPERIMENTAL EVALUATION

A. Performed Experiments

During the experiments, models are trained for all five training sets listed in Sec. IIIA, i.e. Noise-Train, Speech-Train, Events-Train, SN-Train and SNE-Train. Each model is trained three times depending on the utilized augmentation method, i.e. with background noise (BkgN) augmentation, SpecAugment or without any augmentation. In the testing step, each of the trained models is tested on all three homogeneous datasets to evaluate the performance on each of the discussed target sound sources. Apart from Event-Test and Speech-Test datasets, the testing set from Noise-Train is used to evaluate the models' performance on noise signals. Finally, we compare the results obtained by all trained models on three test sets to investigate their generalization and performance for the specific data type. The procedure is repeated 5 times to obtain a mean result for all trials. Experiments are performed using the Keras library [26].

B. Evaluation Measures

In order to evaluate models' performance, the DOA error metric is used [27]. For a sequence of N frames, the measure is defined as

$$E_{\mathrm{DOA}} = \frac{1}{\sum_{n=0}^{N-1} D_E^n} \sum_{n=0}^{N-1} \mathcal{H}(\mathrm{DOA}_R^n, \mathrm{DOA}_E^n), \quad (4)$$

where D_E^n denotes the number of estimated DOAs in the n-th frame with $n = 0, 1, \ldots, N-1$ and $\mathcal{H}(\cdot)$ denotes the Hungarian algorithm for matching the reference and estimated direction vectors. The matching criterion is defined as a Cartesian distance between the respective DOAs, similarly to [28] which has shown to be a suitable regression target [29]. It is computed according to

$$\sigma = \frac{360^\circ}{\pi} \sin^{-1}\left(\frac{\sqrt{(x_E - x_R)^2 + (y_E - y_R)^2 + (z_E - z_R)^2}}{2}\right), \quad (5)$$

where x, y, z denote the coordinates of the DOA vector, and subscripts E and R stand for the estimated and reference DOA values. The final results are averaged over all folds [30].

V. RESULTS AND DISCUSSION

A. Comparison of Different Data Types

Table III presents the DOA error results obtained using 5 models trained using different training datasets for each of the three test sets.

When testing on noise signals, all models show comparable performance. The lowest DOA error equal to 8.92° is achieved when training the model on the white Gaussian noise, however the models trained on mixtures containing also speech and/or events achieve just sightly worse performance. The largest DOA errors are observed for the Speech-Train and Event-Train datasets, which can be intuitively explained by the absence of noise in the training data.

The differences between the models become much more apparent when testing on audio that contains speech or events. DNNs trained using noise only show definitely the worst performance amongst all compared input data types, with 15.69° and 15.82° DOA errors for speech and events, respectively. For the Speech-Test dataset, we can observe a similar performance of models trained on speech or events only, whereas for Event-Test dataset we can notice a moderately better DOA error for networks trained with sound events. Compared with models trained with noise, the overall improvement exceeds 3.5° in all cases, with the most clearly visible gain reaching over 5° for the Event-Train model. Interestingly, the DOA error can be even further decreased by utilizing mixtures of different signal types. While there is already an improvement for SN-Train,

TABLE III
THE DOA ERROR RESULTS OBTAINED FOR DIFFERENT TYPES OF TRAINING DATA WITHOUT AUGMENTATION.

Training data	DOA error [°]		
	Tested on noise	Speech-Test	Events-Test
Noise-Train	**8.92 ± 0.07**	15.69 ± 0.21	15.82 ± 0.22
Speech-Train	9.53 ± 0.08	11.52 ± 0.12	11.97 ± 0.12
Events-Train	9.38 ± 0.09	11.34 ± 0.11	10.90 ± 0.10
SN-Train	9.17 ± 0.08	10.99 ± 0.10	10.70 ± 0.11
SNE-Train	8.97 ± 0.09	**10.01 ± 0.10**	**10.57 ± 0.10**

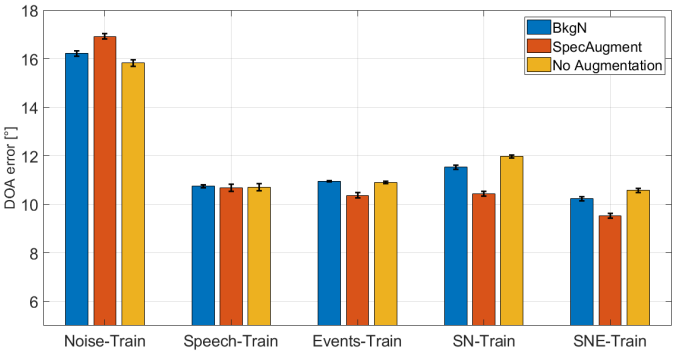

Fig. 1. Localization performance obtained using different data augmentation methods. Tested on Event-Test, averaged over 5 trials.

we observe the best performance for SNE-Train with 10.01° and 10.57° when tested on speech and events, respectively.

These results show that even for source-independent localizers, a much better DOA estimation can be achieved by including a variety of different sound source types in the training process. Comparing the results obtained for models trained only on homogeneous data, it can be observed that DNNs trained on both speech and events tend to generalize better then models trained with noise. Still, the performance can be further increased by utilizing all available signal types.

B. Data augmentation

We repeat the experiments for the training data augmented using the SpecAugment and BkgN methods. The emerging conclusions are coherent for all testing sets, hence we show only the results for the Event-Test dataset in Fig. 1.

Augmentation that relies on adding background noise to the data results in a slight improvement for models trained on the mixture sets. For Event-Train and Speech-Train we observe no significant change of the networks' performance, whereas Noise-Train shows a moderate increase in DOA error. Better results are observed when using the SpecAugment technique. The lowest overall error equal to 9.52° is shown for SNE-Train. A similar improvement can be seen for SN-Train and Event-Train, whilst for Speech-Train and Noise-Train again the technique does not lead to better results.

Overall, using augmentation techniques for all investigated training sets, does not remarkably change the differences observed between the data types. Best results are obtained using both mixture datasets, significantly outperforming the Noise-Train set. Performance can be further increased especially by using the SpecAugment technique. In this study we focused on simple augmentation strategies independent of the spatial format and spatial properties of the multichannel signals. Recently some dedicated spatial augmentation techniques have appeared for joint localisation and detection [31–33]. Such techniques are expected to bring larger benefits to localization than the ones tested here, and their comparison remains a topic for future research. Nevertheless, we expect the benefits from using a diverse dataset to remain valid.

VI. CONCLUSIONS

In this paper, we investigate the robustness of sound source localization depending on the diversity of the training data. Seven different datasets are created, each consisting of homo- and heterogeneous mixtures of noise, speech and sound events. With the aim of localizing overlapping sound events, we perform a comparative study analyzing the models' performance depending on the type of training data. Furthermore, we utilize two augmentation methods to use them as a point of reference for our results. All experiments are performed for a scenario of two simultaneously active sources.

We show that models trained with speech or events show a significant improvement over DNNs fed with noise signals only when tested on their target sound sources. Using mixtures of different types of sounds leads to even larger improvements and very good generalization between the different testing sets. Interestingly, models trained with speech and sound events showed better generalization to noise signals than *vice versa*. Compared with noise signals, the gain from using other types of sound sources ranges from 3° to 5° of the DOA error, showing that the common technique of training source localizers with noise might be not optimal. Differences between the datasets remain for scenarios utilizing augmentation methods as well. We note that the conclusions should be further investigated by performing experiments and tests on real life data, which is a challenging task due to the shortfall of data containing sound events. Our study shows however that including sound events during the training of SSL models can significantly improve their final performance.

ACKNOWLEDGMENT

The research leading to these results has received funding from the National Science Centre under grant number DEC-2017/25/B/ST7/01792.

REFERENCES

[1] M. Brandstein, *Microphone arrays: signal processing techniques and applications*. Springer Science & Business Media, 2001.

[2] L. Perotin, R. Serizel, E. Vincent, and A. Guérin, "CRNN-based multiple DoA estimation using acoustic intensity features for ambisonics recordings," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 22–33, 2019.

[3] S. Chakrabarty and E. A. P. Habets, "Multi-speaker DOA estimation using deep convolutional networks trained with noise signals," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 8–21, 2019.

[4] T. N. T. Nguyen, W.-S. Gan, R. Ranjan, and D. L. Jones, "Robust source counting and DOA estimation using spatial pseudo-spectrum and convolutional neural network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2626–2637, 2020.

[5] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "Robust sound source tracking using SRP-PHAT and 3D convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 300–311, 2020.

[6] M. J. Bianco, S. Gannot, E. Fernandez-Grande, and P. Gerstoft, "Semi-supervised source localization in reverberant environments using deep generative modeling," *The Journal of the Acoustical Society of America*, vol. 148, no. 4, pp. 2662–2662, 2020.

[7] F. Vesperini, P. Vecchiotti, E. Principi, S. Squartini, and F. Piazza, "A neural network based algorithm for speaker localization in a multi-room environment," in *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2016, pp. 1–6.

[8] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 2814–2818.

[9] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1901–1913, 2017.

[10] D. Krause, A. Politis, and K. Kowalczyk, "Feature overview for joint modeling of sound event detection and localization using a microphone array," in *28th European Signal Processing Conference (EUSIPCO 2020)*, 2020, pp. 31–35.

[11] V. Tuzlukov, *Signal Processing Noise*. CRC Press, 2010.

[12] J. Allen, M. S. Hunnicutt, and D. Klatt, *From Text to Speech: The MITalk system*. Cambridge University Press, 1987.

[13] T. Virtanen, M. D. Plumbley, and D. Ellis, *Computational Analysis of Sound Scenes and Events*. Springer, 2018.

[14] A. Politis, "Microphone array processing for parametric spatial audio techniques," Ph.D. dissertation, Aalto University, 2016, https://github.com/polarch/shoebox-roomsim.

[15] D. S. Park, Y. Zhang, C.-C. Chiu, Y. Chen, B. Li, W. Chan, Q. V. Le, and Y. Wu, "Specaugment on large scale datasets," *arXiv:1912.05533*, 2019.

[16] X. Li, Y. Zhang, X. Zhuang, and D. Liu, "Frame-level specaugment for deep convolutional neural networks in hybrid ASR systems," *arXiv:2012.04094*, 2020.

[17] A. Pervaiz, F. Hussain, H. Israr, M. A. Tahir, F. R. Raja, N. K. Baloch, F. Ishmanov, and Y. B. Zikria, "Incorporating noise robustness in speech command recognition by noise augmentation of training data," *Sensors*, vol. 20, no. 8, 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/8/2326

[18] S. Kapka and M. Lewandowski, "Sound source detection, localization and classification using consecutive ensemble of CRNN models," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*. New York University, 2019.

[19] D. Krause, A. Politis, and K. Kowalczyk, "Comparison of convolution types in CNN-based feature extraction for sound source localization," in *28th European Signal Processing Conference (EUSIPCO 2020)*, 2020, pp. 820–824.

[20] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multitalker speech separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 241–245.

[21] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *Proceedings of The 43rd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, 2018, pp. 1–5.

[22] J. S. Garofolo, "TIMIT acoustic phonetic continuous speech corpus," *Linguistic Data Consortium*, 1993.

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

[24] N. Turpault and R. Serizel, "DESED synthetic (version v2.2)." *Zenodo*, 2020, http://doi.org/10.5281/zenodo.3713328.

[25] 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. Interspeech 2019*, 2019, pp. 2613–2617. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2680

[26] F. Chollet *et al.*, "Keras," GitHub, Web Download, 2015.

[27] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *26th European Signal Processing Conference (EUSIPCO)*, 2018, pp. 1462–1466.

[28] ——, "Localization, detection and tracking of multiple moving sound sources with a convolutional recurrent neural network," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, 2019, pp. 20–24.

[29] L. Perotin, A. Défossez, E. Vincent, R. Serizel, and A. Guérin, "Regression versus classification for neural network based audio source localization," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019.

[30] G. Forman and M. Scholz, "Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement," *ACM SIGKDD Explorations Newsletter*, vol. 12, no. 1, pp. 49–57, 2010.

[31] L. Mazzon, Y. Koizumi, M. Yasuda, and N. Harada, "First order ambisonics domain spatial augmentation for DNN-based direction of arrival estimation," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, New York University, NY, USA, October 2019, pp. 154–158.

[32] P. Pratik, W. J. Jee, S. Nagisetty, R. Mars, and C. Lim, "Sound event localization and detection using crnn architecture with mixup for model generalization," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, New York University, NY, USA, October 2019, pp. 199–203.

[33] K. Shimada, N. Takahashi, S. Takahashi, and Y. Mitsufuji, "Sound event localization and detection using activity-coupled cartesian DOA vector and RD3NET," DCASE2020 Challenge, Tech. Rep., July 2020.