ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10094681

AN EXPERIMENTAL STUDY ON SOUND EVENT LOCALIZATION AND DETECTION UNDER REALISTIC TESTING CONDITIONS

Shutong Niu[1], *Jun Du*[1,*], *Qing Wang*[1], *Li Chai*[2], *Huaxin Wu*[2], *Zhaoxu Nian*[1],
Lei Sun[2], *Yi Fang*[2], *Jia Pan*[2], *Chin-Hui Lee*[3]

[1] University of Science and Technology of China, Hefei, China
[2] iFLYTEK, Hefei, China
[3] Georgia Institute of Technology, Atlanta, USA
jundu@ustc.edu.cn, chl@ece.gatech.edu

ABSTRACT

We study four data augmentation (DA) techniques and two model architectures on realistic data for sound event localization and detection (SELD). First, based on ResNet-Conformer (RC), we compare the four DA approaches on the realistic DCASE 2022 SELD test set which is often not easy to handle due to room reverberations and audio overlaps in spontaneous recordings. Experimental results show that, except for audio channel swapping (ACS), the other three data augmentation methods that work well on the simulated SELD data set are no longer effective due to mismatches between simulated and realistic conditions. Next, using ACS-based augmentation, the two improved ResNet-Conformer networks further enhance SELD performances in realistic conditions. By incorporating these two sets of techniques, our overall system ranked the first place in SELD task of the DCASE 2022 Challenge.

Index Terms— Sound event localization and detection, realistic data, data augmentation, model architecture, DCASE 2022

1. INTRODUCTION

Sound event localization and detection (SELD) is a task to identify various types of sound events and the corresponding direction-of-arrival (DOA) in a sound scene over time [1]. It has broad applications, such as detection and localization of alarms, audio surveillance, bio-diversity monitoring and virtual reality [2–4]. SELD has attracted a lot of attention since it was introduced in the Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge in 2019 [1], which provides a standard data set for researchers.

The SELD problems contain two subtasks: sound event detection (SED) and sound source localization (SSL). In the early stages, these two tasks are treated separately. For the SED, some conventional methods are developed from the automatic speech recognition (ASR) area, e.g., hidden Markov models (HMMs) [5] and Gaussian mixture models (GMMs) [3]. Some studies utilize the support vector machines (SVM) and non-negative matrix factorization (NMF) for sound event detection [6, 7]. For the SSL, most conventional methods are based on the array processing approaches, such as acoustic intensity vector analysis and steered response power [6, 8]. In recent years, neural network (NN)-based methods have been widely used in SELD [1, 4, 9–14]. According to the modeling methods, these methods can be roughly divided into two categories. The first one adopts the deep neural networks for sound event detection (SED) and still utilizes the physics-based signal processing algorithms for sound source localization (SSL) [9]. The second one uses the NN-based methods for joint modeling of detection and localization in the SELD task, which has gained more and more interest and been widely used in the DCASE Challenges [1, 4, 10–14]. In [4], two branches, namely, the SED branch and DOA branch, are adopted to perform the SED and SSL in a single network. In [11], a two-stage method is proposed, which learns SED first, and then uses the learned feature layers to estimate DOA. Then activity-coupled Cartesian DOA (ACCDOA) and multi-ACCDOA representations are proposed [12, 13], which combine SED and DOA estimation together and enable the SELD task to be solved through a single target. An event independent network V2 (EINV2) is proposed in [14] to handle the SELD task through a track-wise format.

One important factor of these NN-based methods is the model architecture. Various architectures have been explored for the SELD task [2, 4, 14, 15]. In [4], convolutional and recurrent neural network (CRNN) is adopted in SELD task and becomes a widely used architecture in DCASE Challenges. In [14], Multi-head self-attention (MHSA) is shown to be effective for SELD. In [2], the Conformer [16] is adopted in ResNet to model the global and local information, called ResNet-Conformer. Besides, densely connected multidilated DenseNet (D3Net) and time-frequency RNN (TFRNN) are utilized in [15], which helped their system achieve first place in the DCASE 2021 SELD task. The training data size is another important factor for the NN-based methods. Different data augmentation techniques have been explored, such as equalized mixture data augmentation (EMDA), Cutout, SpecAugment and Mixup [17–20]. In [2], a four-stage data augmentation approach is proposed, including audio channel swapping (ACS), multi-channel simulation (MCS), time-domain mixing (TDM), and time-frequency masking (TFM).

As described above, the SELD area has made great progress by introducing NN-based methods. However, most of them are evaluated on the simulated datasets and have not been explored in real scenes. Therefore, in this paper, we explore the two main factors in NN-based SELD methods under realistic conditions by leveraging the DCASE 2022 SELD dataset [21], namely data augmentation methods and model architectures. The key contributions of this paper are three-fold: (1) we explore different data augmentation methods on the realistic data and compare with their performances obtained on the simulated data; (2) we improve the original ResNet-Conformer for the SELD task, which can achieve better performance under realistic scenes; and (3) by incorporating these two key methods, our overall system achieved first place among all submitted systems in the SELD task of the DCASE 2022 Challenge.

[*] corresponding author

Table 1: The statistics of duration (minute) and overlap ratio ($\%$) for each class on the testing part of the DCASE 2022 SELD dataset.

Sound Class	Female Speech	Male Speech	Clapping	Telephone	Laughter	Domestic Sounds	Walk	Door	Music	Musical Instrument	Water Tap	Bell	Knock
Duration (min)	23.04	53.66	0.98	1.50	4.47	23.18	1.18	0.52	45.49	14.43	3.45	1.95	0.08
Overlap Ratio ($\%$)	60.00	54.25	83.19	51.89	84.79	69.56	54.79	66.77	67.20	85.62	79.79	100.00	82.22

2. DATA ANALYSIS

Unlike previous iterations, the SELD task in DCASE 2022 uses recordings of real sound scenes to evaluate the systems for the first time, bringing significant differences in the dataset [21]. Through analysis of the DCASE 2022 SELD dataset, the biggest difficulties can be summarized into three aspects. First of all, the class presences depend on the natural actions and interactions in real scenes, which may lead to the uneven distribution of classes. Table 1 lists the statistics of duration and overlap ratio for each class on the testing part of the released DCASE 2022 SELD dataset. As can be seen, the duration of class 'Female Speech' is 53.66 minutes, while that of class 'Knock' is only 0.08 minutes. This will pose a challenge to the SED task. Secondly, unlike the simulated data which use the captured spatial room impulse responses (SRIRs) to generate the spatial information, real SELD data contain more complex reverberations changing with rooms, sound locations, and array positions. This will increase the difficulty of the SSL task. Finally, it is very common for sound events to occur simultaneously in real scenes. From Table 1 we can see that the overlap ratios of all events are more than 50$\%$, and the overlap ratio of class 'Bell' even reaches 100$\%$. Especially, the realistic overlapping regions are not the simple addition of single sources with different signal-to-noise ratios (SNRs) like in the simulated dataset. The high overlap ratios in real data will make the SELD task face severe challenges.

In previous work [2], our team proposed a four-stage data augmentation approach and employed the ResNet-Conformer to process the simulated DCASE 2020 SELD task, which has achieved good results. However, whether these methods are effective on more difficult real data is still unknown. Therefore, based on our previous methods, we conduct an experimental exploration of data augmentation methods on real data and also improve the ResNet-Conformer for real scenarios. These two key methods bring the main improvements for our DCASE 2022 SELD system [10].

3. TECHNIQUES FOR PERFORMANCE IMPROVEMENT

3.1. Data Augmentation

Audio Channel Swapping - In the DCASE SELD task, the 4-channel first-order Ambisonics (FOA) spatial format [1] is conversed from the 32-channel recordings. Considering a sound source from the azimuth angle ϕ and elevation angle θ, we can decompose the sound field on the FOA channels as:

$$\begin{bmatrix} S_1(t,f) \\ S_2(t,f) \\ S_3(t,f) \\ S_4(t,f) \end{bmatrix} = \begin{bmatrix} 1 \\ \sin(\phi)\cos(\theta) \\ \sin(\theta) \\ \cos(\phi)\cos(\theta) \end{bmatrix} p(t,f). \tag{1}$$

where $p(t,f)$ is a point of the sound source in the time-frequency (T-F) domain. $S_i(t,f)$ corresponds to the i-th channel in FOA data. Therefore, the frequency-independent spatial responses of the FOA channels can be obtained as:

$$\begin{aligned} H_1(\phi,\theta,f) = 1, \qquad H_2(\phi,\theta,f) = \sin(\phi)\cos(\theta), \\ H_3(\phi,\theta,f) = \sin(\theta), H_4(\phi,\theta,f) = \cos(\phi)\cos(\theta). \end{aligned} \tag{2}$$

where $H_m(\phi,\theta,f)$ denotes the spatial response of the m-th channel. From Eq.2, we can see that the spatial responses of the FOA recordings can be represented as cosine functions, which correspond to the DOAs. Therefore, the angle transformations of DOAs can be easily expressed by this form. Eight DOA transformations were performed for ACS augmentation in our submitted system, as detailed in [2], and this method is also discussed in [22]. Note that the ACS method preserves the reverberation characteristics of real data and doesn't simulate new overlapping segments.

Multi-Channel Simulation - Multi-channel simulation (MCS) [2] is an augmentation technique to directly generate new data, whose procedure can be summarized as two steps: (1) extracting the spectral and spatial information from the non-moving single sound event segments in original recordings; (2) randomly selecting the extracted spectral and spatial information to simulate the multi-channel data. The details can be found in [2].

Time-Domain Mixing - We simulate the overlapping segments by mixing two single sources in realistic recordings. Previous results [2] on the simulated dataset show that TDM can improve the model's ability to handle overlapping segments.

Time-Frequency Masking - We also use the SpecAugment [19] method to improve the generalization ability of the model. The time and frequency masks are randomly applied on the extracted log Mel-spectrogram features during the training process.

3.2. Model Architecture

The ResNet-Conformer [2] can capture both the global- and local-level dependencies within the input audio sequences, which has been proven effective in the simulated dataset. In the SELD task of the DCASE 2022 Challenge, we employ the ResNet-Conformer as our main model architecture. The overall architecture of the ResNet-Conformer is shown in Fig. 1. As can be seen, we first use a ResNet18 network to extract the deep representations. Then the Conformer [16] is used to model the context dependencies within the input. Suppose the input of one Conformer layer is $\mathbf{z}$, we can obtain the output as:

$$\hat{\mathbf{z}} = \frac{1}{2}\mathrm{FFN}(\mathbf{z}) + \mathbf{z}, \tag{3}$$

$$\mathbf{z}' = \mathrm{MHSA}(\hat{\mathbf{z}}) + \hat{\mathbf{z}}, \tag{4}$$

$$\mathbf{z}'' = \mathrm{Conv}(\mathbf{z}') + \mathbf{z}', \tag{5}$$

$$\mathbf{o} = \mathrm{Layernorm}(\frac{1}{2}\mathrm{FFN}(\mathbf{z}'') + \mathbf{z}''). \tag{6}$$

where $\mathrm{FFN}(\cdot)$, $\mathrm{MHSA}(\cdot)$, $\mathrm{Conv}(\cdot)$ and $\mathrm{Layernorm}(\cdot)$ denote the feed forward network, multi-head self-attention module, convolution block, and layer normalization, respectively. The convolution block

[1] There are two data formats, namely first-order Ambisonics (FOA) and tetrahedral microphone array (MIC) in the DCASE dataset. We find that FOA can achieve slightly better results than MIC. Therefore, the data augmentation methods are all applied to the FOA data.

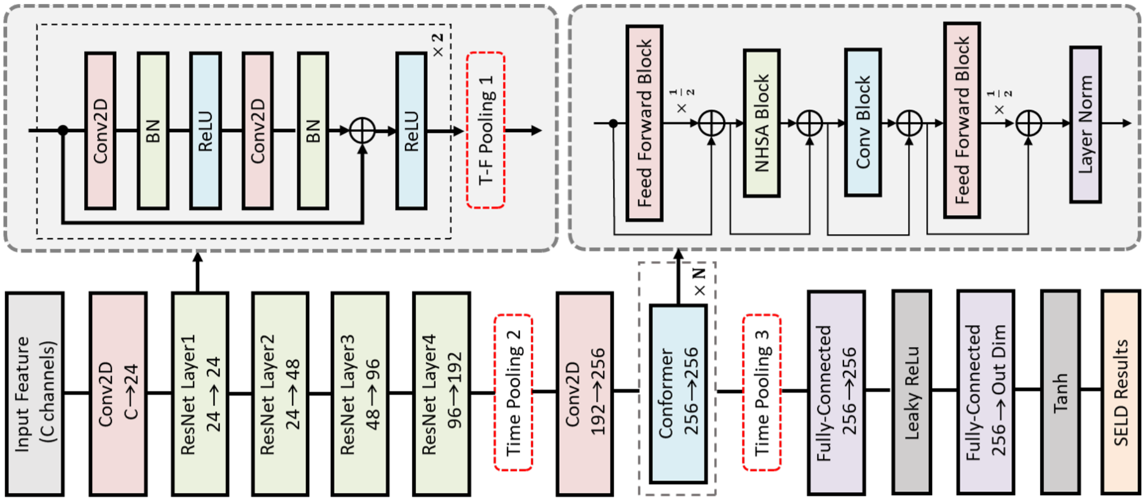


Fig. 1: The overall architecture of the ResNet-Conformer used in our DCASE 2022 SELD system, which mainly contains a ResNet18 network and N Conformer layers. 'MHSA' means multi-head self-attention. 'T-F' means time-frequency. The red dotted boxes indicate the selected positions for the time pooling, and we choose one of them to perform the time pooling.

can exploit the local features, and the self-attention can capture the global context, which is beneficial for handling the SELD tasks.

ResNet-Conformer can achieve better performance than the baseline CRNN. However, there is a disadvantage in the ResNet-Conformer. In the DCASE 2022 SELD task, time pooling is usually performed only once to make the time resolution of features the same as the labels. However, the original ResNet-Conformer applies the time pooling at the 'T-F Pooling 1' (TFP1) position in the first ResNet layer, as shown in Fig. 1. Therefore, a problem worth considering is that applying pooling in the time dimension so early may cause losses of information in subsequent neural networks. Especially in real scenes, the time duration distribution of different classes is extremely uneven, and some classes with less time duration need a higher time resolution for detection and localization. To deal with this problem, we have made different improvements to the model architecture. The first one is we only perform the frequency pooling in the ResNet, and move the time pooling operation between the ResNet and Conformer, as shown in 'Time Pooing 2' (TP2) in Fig. 1, which is called ResNet-Conformer with middle pooling (RCMP). Similarly, the second one is we only apply the time pooling after the Conformer, as shown in 'Time Pooing 3' (TP3) in Fig. 1, which is called ResNet-Conformer with late pooling (RCLP). Such adjustments help the model to obtain higher time resolution under the same inputs.

4. EXPERIMENTS AND RESULT ANALYSIS

We evaluate SELD on the official development set of the DCASE 2022 Task 3 [21]. The set totals 121 recording clips (about 5 hours), which can be split into a training part (*dev-set-train*, 67 clips) and a testing part (*dev-set-test*, 54 clips) according to the official set. We follow this configuration and also add the 1200 one-minute synthesized mixtures (*synth-set*) offered by organizers to build the basic training set (about 23 hours). All above recordings are 4-channel with a 24 kHz sampling rate. We apply the short-term Fourier transform (STFT) on 4-channel FOA audios to extract log Mel-spectrograms. Then we calculate the 3-channel acoustic intensity vector (IV). We concatenate them to get the 7-channel feature at each frame. In DCASE 2022 SELD task, the CRNN is still utilized as the model architecture of baseline. We use the ResNet-Conformer as our main model architecture [2]. The number of attention heads is 8. The dimensions of input, key and value vectors are set to 256, 32 and 32, respectively. The number of Conformer layers N is set to 8. Adam [23] is adopted as the optimizer. The tri-stage learning rate scheduler [19] is used with an upper limit of 0.001. The maximum number of training steps is set to 120,000. To make a fair comparison, the output format is unified as multi-ACCDOA in this paper. The mean square error (MSE) loss with auxiliary duplicating permutation invariant training (ADPIT) [13] is utilized to optimize the model. We evaluate all methods with SELD_{score} [24], which is calculated as:

$$\text{SELD}_{score} = \frac{1}{4}[ER_{20^{\text{o}}} + (1 - F_{20^{\text{o}}}) + LE_{\text{CD}}^{'} + (1 - LR_{\text{CD}})] \quad (7)$$

where $ER_{20^{\text{o}}}$ and $F_{20^{\text{o}}}$ are location-dependent error rate and F-score when the spatial error is within 20^{o}. $LE_{\text{CD}}^{'} = LE_{\text{CD}}/\pi$, in which LE_{CD} denotes the localization error between predictions and references of the same class. LR_{CD} is a simple localization recall metric. Note that the $F_{20^{\text{o}}}$, LE_{CD} and LR_{CD} in DCASE 2022 SELD task are calculated through macro-averaging [21], which is different from the micro-averaging used in previous challenges. However, the relative trends of these two calculation methods are usually consistent.

4.1. Evaluation of the Data Augmentation Methods

Table 2 lists the performance comparison among different DA methods on the SELD dataset from the DCASE 2022. To explore whether these methods are effective under realistic conditions, we first apply them separately to the SELD dataset based on the ResNet-Conformer. As can be observed, the performances of these methods are quite different on the realistic dataset. The ACS method still achieves effective improvement (20.0% relatively), which is consistent with its performance on the simulated dataset [2]. However, the other three methods (MCS, TDM and TFM) bring limited improvements (2.0% to 4.0% relatively), which is very different from their performances on the simulated dataset [2]. Even combing with ACS, they still cannot achieve effective improvement. According to the gap between the real and simulated data and the characteristics of DA methods, the reasons for this difference can be roughly

Table 2: The performance comparison among different data augmentation methods on the SELD dataset of the DCASE 2022 Challenge, including ACS, MCS, TDM and TFM. 'Base' denotes the basic training set. The model architecture is ResNet-Conformer.

Data	$ER_{20°}$	$F_{20°}$	LE_{CD}	LR_{CD}	$SELD_{score}$
Base	0.72	27.0%	25.40°	62.0%	0.50
ACS	0.56	42.0%	20.60°	67.0%	0.40 (20.0% ↓)
MCS	0.71	32.0%	21.26°	56.0%	0.48 (4.0% ↓)
TDM	0.69	28.0%	22.88°	59.0%	0.49 (2.0% ↓)
TFM	0.69	33.0%	22.50°	57.0%	0.48 (4.0% ↓)
ACS + MCS	0.57	42.0%	20.50°	66.0%	0.40 (20.0% ↓)
ACS + TDM	0.57	40.0%	18.61°	66.0%	0.40 (20.0% ↓)
ACS + TFM	0.55	46.0%	20.50°	64.0%	0.39 (22.0% ↓)

summarized in two aspects: (1) ACS method does not destroy the reverberation conditions and overlapping segments of real recordings, so it can bring relatively large improvement in real scenes. However, MCS and TDM change the spatial information and generate new overlapping segments, respectively, which are quite different from the realistic scenes, therefore resulting in little performance improvement; (2) the occurrences of simultaneous events are fairly common in the realistic SELD dataset as illustrated in Table 1. This leads to the reduction of single-event segments used for MCS and TDM, making the corresponding improvements not obvious. As for TFM, the experiments in [2] show that the improvement is obvious when the data size is large in the SELD task. However, due to the annotation complexity, the size of real data is usually small, making it difficult for TFM to achieve sufficient improvement in real scenarios. When combined with ACS, TFM can bring slight improvement (from 0.40 to 0.39), but this improvement is still not as obvious as that under simulated conditions. Therefore, in order to improve the training efficiency, only the ACS method is employed in our subsequent experiments.

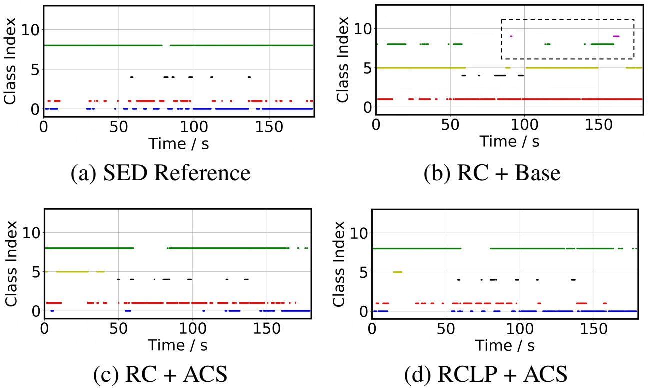
4.2. Evaluation of the Model Architectures

Table 3 compares the performances of different model architectures introduced in Section 3.2 on DCASE 2022 SELD dataset. 'Base' denotes the basic training set offered by the organizers. We first compare the different architectures using the basic training set. As can be seen, the model can achieve better performance in both detection and localization tasks as the time pooling operation gradually moves backward. This shows that with the higher time resolution, the improved ResNet-Conformer can detect events more accurately, and also can better track the movements of the sound sources in the real scene. Finally, we train the modified RC with the ACS dataset, and moving back the time pooling can achieve more significant improvements, yielding a 32.1 % relative improvement compared with the baseline. We also added some post-processing methods during the challenge, such as dynamic threshold and time-overlapped testing [10], denoted as 'RCLP + PP' in Table 3. After fusing the ACCDOA and multi-ACCDOA based systems, we obtained the final results, which achieves a 52.8 % relative improvement compared with the baseline and first place in the SELD task of DCASE 2022.

To better illustrate the impacts of different methods, we choose one recording and present the corresponding SED results in Fig. 2. As can be seen, the 'RC + Base' can roughly detect sound events, but there are several obvious errors as shown in Fig. 2 (b): (1) mistakenly classifying the 'class 8' (green line, *Music*) into the 'class 5' (yellow line, *Domestic sounds*) and 'class 9' (purple line in dotted box, *Musical instrument*); (2) cannot distinguish between the 'class 0' (blue line, *Female speech*) and 'class 1' (red line, *Male speech*). Through employing the ACS method, the model can better distinguish between these classes and the missing 'class 0' and 'class 8' are partially corrected. Meanwhile, the 'class 5' and 'class 9' are also obviously shielded. Moreover, with higher time resolution, the model can further distinguish between the 'class 0' and 'class 1', and further shield 'class 5', as shown in Fig. 2 (d).

Table 3: The performance comparison among different model architectures on the DCASE 2022 SELD dataset. The model architectures include ResNet-Conformer (RC) (as in Table 2), ResNet-Conformer with middle pooling (RCMP), and ResNet-Conformer with late pooling (RCLP). PP means post-processing.

Models	Data	$ER_{20°}$	$F_{20°}$	LE_{CD}	LR_{CD}	$SELD_{score}$
Baseline (CRNN)	Base	0.72	24.0%	26.61°	49.0%	0.53
RC (in Table 2)		0.72	27.0%	25.40°	62.0%	0.50 (5.7% ↓)
RCMP		0.70	32.0%	21.18°	58.0%	0.48 (9.4% ↓)
RCLP		0.71	31.0%	22.30°	64.0%	0.47 (11.3% ↓)
RC (in Table 2)	ACS	0.56	42.0%	20.60°	67.0%	0.40 (18.9% ↓)
RCMP		0.50	49.0%	17.47°	63.0%	0.37 (30.2% ↓)
RCLP		0.51	51.0%	17.34°	66.0%	0.36 (32.1% ↓)
RCLP + PP [10]	-	0.41	61.0%	15.30°	74.0%	0.28 (47.2% ↓)
Submission [10]	-	0.38	67.0%	14.80°	78.0%	0.25 (52.8% ↓)



(a) SED Reference

(b) RC + Base

(c) RC + ACS

(d) RCLP + ACS

Fig. 2: The visualization and comparison of SED results of different methods, including ResNet-Conformer (RC) trained on the basic training data (denoted as RC + Base), RC trained on the audio channel swapping (ACS) data (denoted as RC + ACS), and RCLP trained on the ACS data (denoted as RCLP + ACS).

5. SUMMARY

We explore four data augmentation approaches and two improved RC architectures with realistic recordings for SELD. We find that the four augmentation methods behave differently in simulated and realistic data sets. High feature resolution often benefits the model performance. Moreover, a good combination of techniques can help the models achieve further improvements. In the future, we will explore more effective techniques for SELD under realistic conditions.

6. ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China under Grant No. 62171427.

7. REFERENCES

[1] Archontis Politis, Annamaria Mesaros, Sharath Adavanne, Toni Heittola, and Tuomas Virtanen, “Overview and evaluation of sound event localization and detection in dcase 2019,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 684–698, 2020.

[2] Qing Wang, Jun Du, Hua-Xin Wu, Jia Pan, Feng Ma, and Chin-Hui 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.

[3] Giuseppe Valenzise, Luigi Gerosa, Marco Tagliasacchi, Fabio Antonacci, and Augusto Sarti, “Scream and gunshot detection and localization for audio-surveillance systems,” in *2007 IEEE Conference on Advanced Video and Signal Based Surveillance*. IEEE, 2007, pp. 21–26.

[4] Sharath Adavanne, Archontis Politis, Joonas Nikunen, and Tuomas 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, 2018.

[5] Taras Butko, Fran González Pla, Carlos Segura, Climent Nadeu, and Javier Hernando, “Two-source acoustic event detection and localization: Online implementation in a smart-room,” in *2011 19th European Signal Processing Conference*. IEEE, 2011, pp. 1317–1321.

[6] Kuba Lopatka, Jozef Kotus, and Andrzej Czyzewski, “Detection, classification and localization of acoustic events in the presence of background noise for acoustic surveillance of hazardous situations,” *Multimedia Tools and Applications*, vol. 75, no. 17, pp. 10407–10439, 2016.

[7] Jort F Gemmeke, Lode Vuegen, Peter Karsmakers, Bart Vanrumste, et al., “An exemplar-based nmf approach to audio event detection,” in *2013 IEEE workshop on applications of signal processing to audio and acoustics*. IEEE, 2013, pp. 1–4.

[8] Charles Knapp and Glifford Carter, “The generalized correlation method for estimation of time delay,” *IEEE transactions on acoustics, speech, and signal processing*, vol. 24, no. 4, pp. 320–327, 1976.

[9] Thi Ngoc Tho Nguyen, Douglas L Jones, and Woon-Seng Gan, “A sequence matching network for polyphonic sound event localization and detection,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 71–75.

[10] Qing Wang, Li Chai, Huaxin Wu, Zhaoxu Nian, Shutong Niu, Siyuan Zheng, Yuyang Wang, Lei Sun, Yi Fang, Jia Pan, et al., “The nerc-slip system for sound event localization and detection of dcase2022 challenge,” .

[11] Yin Cao, Qiuqiang Kong, Turab Iqbal, Fengyan An, Wenwu Wang, and Mark D Plumbley, “Polyphonic sound event detection and localization using a two-stage strategy,” *arXiv preprint arXiv:1905.00268*, 2019.

[12] Kazuki Shimada, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, and Yuki Mitsufuji, “Accdoa: Activity-coupled cartesian direction of arrival representation for sound event localization and detection,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 915–919.

[13] Kazuki Shimada, Yuichiro Koyama, Shusuke Takahashi, Naoya Takahashi, Emiru Tsunoo, and Yuki Mitsufuji, “Multi-accdoa: Localizing and detecting overlapping sounds from the same class with auxiliary duplicating permutation invariant training,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 316–320.

[14] Yin Cao, Turab Iqbal, Qiuqiang Kong, Fengyan An, Wenwu Wang, and Mark D Plumbley, “An improved event-independent network for polyphonic sound event localization and detection,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 885–889.

[15] Kazuki Shimada, Naoya Takahashi, Yuichiro Koyama, Shusuke Takahashi, Emiru Tsunoo, Masafumi Takahashi, and Yuki 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] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al., “Conformer: Convolution-augmented transformer for speech recognition,” *arXiv preprint arXiv:2005.08100*, 2020.

[17] Naoya Takahashi, Michael Gygli, Beat Pfister, and Luc Van Gool, “Deep convolutional neural networks and data augmentation for acoustic event detection,” *arXiv preprint arXiv:1604.07160*, 2016.

[18] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang, “Random erasing data augmentation,” in *Proceedings of the AAAI conference on artificial intelligence*, 2020, vol. 34, pp. 13001–13008.

[19] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le, “Specaugment: A simple data augmentation method for automatic speech recognition,” *arXiv preprint arXiv:1904.08779*, 2019.

[20] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz, “mixup: Beyond empirical risk minimization,” *arXiv preprint arXiv:1710.09412*, 2017.

[21] Archontis Politis, Kazuki Shimada, Parthasaarathy Sudarsanam, Sharath Adavanne, Daniel Krause, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, Yuki Mitsufuji, and Tuomas Virtanen, “Starss22: A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events,” *arXiv preprint arXiv:2206.01948*, 2022.

[22] Luca Mazzon, Yuma Koizumi, Masahiro Yasuda, and Noboru Harada, “First order ambisonics domain spatial augmentation for dnn-based direction of arrival estimation,” *arXiv preprint arXiv:1910.04388*, 2019.

[23] Diederik P Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[24] Annamaria Mesaros, Sharath Adavanne, Archontis Politis, Toni Heittola, and Tuomas Virtanen, “Joint measurement of localization and detection of sound events,” in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019, pp. 333–337.