SOUND SOURCE DISTANCE ESTIMATION IN DIVERSE AND DYNAMIC ACOUSTIC CONDITIONS

Saksham Singh Kushwaha[1,2], *Iran R. Roman*[2*], *Magdalena Fuentes*[2,3], *Juan Pablo Bello*[2]

[1] Courant Institute of Mathematical Sciences, New York University, NY, USA
[2] Music and Audio Research Lab, New York University, NY, USA
[3] Integrated Design and Media, New York University, NY, USA

ABSTRACT

Localizing a moving sound source in the real world involves determining its direction-of-arrival (DOA) and distance relative to a microphone. Advancements in DOA estimation have been facilitated by data-driven methods optimized with large open-source datasets with microphone array recordings in diverse environments. In contrast, estimating a sound source's distance remains understudied. Existing approaches assume recordings by non-coincident microphones to use methods that are susceptible to differences in room reverberation. We present a CRNN able to estimate the distance of moving sound sources across multiple datasets featuring diverse rooms, outperforming a recently-published approach. We also characterize our model's performance as a function of sound source distance and different training losses. This analysis reveals optimal training using a loss that weighs model errors as an inverse function of the sound source true distance. Our study is the first to demonstrate that sound source distance estimation can be performed across diverse acoustic conditions using deep learning.

Index Terms— distance estimation, multichannel audio, sound source localization, mean percentage error

1. INTRODUCTION

Sound source localization (SSL) — the task of localizing the position of a sound source relative to a microphone — has been an active area of research for decades [1, 2, 3]. SSL has important downstream applications, including sound source separation [4], audio-based navigation systems [5], and urban surveillance [6]. SSL can be broken down into two subtasks: direction-of-arrival (DOA) estimation, which approximates sound directivity in terms of azimuth and elevation angles, and distance estimation, which approximates the separation between the sound source and the microphone.

Recent developments have focused in DOA estimation. This includes large open-source datasets with DOA annotations for moving sound events in real [7, 8] and simulated [9, 10] acoustic conditions. Using these datasets, researchers have developed models able to simultaneously carry out DOA estimation and classification (i.e. speech vs music vs engine, etc.) [11, 12]. In contrast, distance estimation remains understudied, partly because it is considered to be more difficult [13]. While recent DOA approaches, notably those developed in the context of the DCASE challenge [8, 9, 10, 11, 12], output 3D coordinates to localize sources, they assume those sources to be in the unit sphere, effectively only estimating azimuth and elevation (i.e. DOA). To the best of our knowledge, existing distance estimation approaches include signal processing methods that assume a room's T_{60} to contrast sounds reaching a microphone directly versus indirectly [14, 15]. Data-driven approaches have also been developed, but using small datasets that feature only a handful of rooms [16, 17], or framing the task as classification instead of directly estimating distance [18, 16, 17, 19].

While popular datasets used for DOA estimation lack distance annotations, many have metadata from where this information can be recovered. In this study we add distance annotations to existing open-source datasets. We use these to optimize a convolutional recurrent neural network (CRNN) that estimates the distance of moving sound sources from tetrahedral microphone recordings. Ours system is the first of its kind (i.e. using deep learning), being able to carry out the task of distance estimation, and evaluated in diverse acoustic conditions. We also analyze the effect of different loss functions to learn the task. Our model outperforms a recent distance estimation approach [20] evaluated on the open-source LOCATA dataset [7]. Additionally we evaluate our model's performance across other datasets. In summary, our contributions are:

1. Distance annotations for a collection of open-source datasets previously used for DOA estimation.
2. A model able to estimate the distance of sound sources in diverse environments and acoustic conditions.
3. An analysis of model performance resulting from using different loss functions.[1]

2. RELATED WORK

Sound source distance estimation is straightforward if the onset time t_o and the speed of sound c are known. In a microphone recording, the sound would appear at time $t_r > t_o$. The sound source distance d can be calculated by $d = c \times (t_r - t_o)$. In the real world, however, knowing t_o is virtually impossible.

We focus on sound source distance estimation in enclosed, reverberant environments. Early approaches were inspired by human listening. Humans use the direct-to-reverberant ratio (DRR) [21, 22], which is the ratio between the signal energy directly reaching the listener and energy from wall reflections. The DRR can be applied to multi-channel recordings to carry out sound source distance estimation [14, 15]. Alternatives include binaural cues like spectral magnitude difference [23] and signal coherence [24]. More recently, data-driven approaches have been proposed, including feedforward neural networks (FNNs) or convolutional neural networks (CNNs) with a classification output to categorize sound source distances into one of N pre-defined distance ranges [16, 18, 17, 19]. These models have been developed and evaluated

*corresponding author email: roman@nyu.edu

[1] Code and data: github.com/sakshamsingh1/sound_distance_estimation

979-8-3503-2372-6/23/$31.00 ©2023 IEEE

Dataset	Range	Avg	Ntr	Nts	L	R	M
DCASE	1.35-7.15	3.34	900	300	60.0	9	Y
STARSS	0.42-7.02	1.83	87	74	162.2	16	Y
LOCATA	0.50-3.49	1.78	27	5	18.9	1	Y
MARCo	2.6-12	4.01	5	7	78.6	1	N
METU	0.3-2.2	1.41	146	98	2.0	1	N

Table 1: Summary of datasets used in our study. Columns indicate the range of distances ("Range") and average distance ("Avg") (both in meters), the number of training ("Ntr") and test ("Nts") recordings in each dataset, the average recording duration ("L") (in seconds), the number of unique rooms featured in the dataset ("R"), and whether sound sources move ("M") (Y: yes; N: no). DCASE and STARSS are split into training, validation, and test sets, each with a unique set of rooms.

using synthetic datasets (i.e. using simulated wave propagations) [19] or recordings in a handful of rooms with specific microphone and loudspeaker configurations [18, 16, 17].

When it comes to DOA estimation, many studies have used CRNNs[2] that estimate x, y, z coordinates on an assumed unit sphere (i.e. only estimating azimuth and elevation) [12, 11, 9, 10]. These approaches benefit from open-source data, sometimes produced by generators able to yield large-scales of training data [9, 10]. Generators use real-world multi-channel impulse responses (IR) and noise samples collected in different rooms. Sound scenes can be produced where events can be stationary or move along trajectories traced along neighboring IRs. Datasets with real recordings in rooms also exist [8, 7], and are used to evaluate models in real-world contexts.

Besides DOA estimation, these datasets could also be used to develop distance estimation methods. For instance, Daniel et al. [20] developed a technique that compares higher-order ambisonics (HOA) channels (4th order; 25 total channels) to find temporal relations between a sound source's wall reflections and infer the delay of the propagating signal. This representation is called the Generalized Time-domain Velocity Vector (GTVV) [25]. While their implementation is not publicly-available, they did evaluate it on LOCATA [7], allowing for future comparison between methods using this dataset as a point of reference.

3. METHODS

3.1. Datasets

We annotate sound source distances in existing open-source datasets and a data generator featuring single sound events in real, dynamic, and diverse rooms. We select datasets that use EigenMike since it has been commonly used for DOA estimation research [3, 13, 8].

We use the open-source data generator[3] by Politis et al. [10]. It places sounds in nine unique rooms with predefined trajectories where sounds can appear featuring different power levels. We modified its code to annotate the sound source distance, which we inferred via each room's metadata files where the possible trajectories are delineated. With this generator we create a "DCASE" dataset with recordings separated into training and test splits, each using a different set of rooms. We also use four datasets featuring recordings in real-world environments. We use STARSS (2023 version) [8], which contains sound source distance annotations by its original authors [8]. It features recordings in sixteen unique rooms. We only estimate the distance of single sound sources. Therefore we masked samples where overlapping sounds are present by replacing them with the room's ambient noise. For STARSS we used the "development" set, which comes with recordings separated into training and test splits. LOCATA [7] features recordings in a single room and contains metadata files encoding the sound source distance. We split it by using all "Task 1" and "Task 5" files for training, and "Task 3 evaluation" for testing (consistent with [20] to compare performance)[4]. The 3D-MARCo dataset [26] contains recordings of musical performances inside a reverberant church. We consulted the dataset's documentation and original authors to determine the precise sound source distance. We used the "single sources" recordings for testing and the rest for training. We excluded the "trio" recording as it features simultaneous sound sources at different distances. Finally, METU-SPARG [27] features IRs recorded in an office, sampled over a 3D grid around the microphone. Distance information is present in its metadata files. We use IRs collected below the microphone's center for testing, and the rest for training. Table 1 summarizes datasets. Because of the small number of recordings in LOCATA, MARCo, and METU-SPARG we use channel-swapping [28] to augment the training set of these datasets by a factor of eight (not reflected in Table 1).

Acronym	Full name	$\mathfrak{E}$
AE	absolute error	$\lvert y-\hat{y}\rvert$
SE	squared error	$(y-\hat{y})^2$
APE	absolute percent error	$\frac{1}{y}\lvert y-\hat{y}\rvert$
SPE	squared percent error	$(\frac{1}{y}(y-\hat{y}))^2$
TAPE	thresholded APE	$\max(\delta, \frac{1}{y}\lvert y-\hat{y}\rvert)$

Table 2: Different regressors $\mathfrak{E}$ that we investigate in the loss function. We try TAPE with $\delta = 0.01$, $\delta = 0.1$, $\delta = 0.20$.

3.2. Model and loss

We train a CRNN to dynamically estimate the distance of non-simultaneous sound sources. We modify the CRNN published by Adavane et al. [3] to have two outputs: event detector $\hat{d}$ and distance estimator $\hat{y}$. $\hat{d}$ is trained with binary cross-entropy (BCE) and $\hat{y}$ with a regressor $\mathfrak{E}$. The model's loss is

$$L = \frac{1}{N}\frac{1}{T}\sum_{n=0}^{N-1}\sum_{t=0}^{T-1} d_{n,t}\mathfrak{E}(y_{n,t},\hat{y}_{n,t}) + \text{BCE}(d_{n,t},\hat{d}_{n,t}), \quad (1)$$

where $y \in \mathcal{R}^+$ and $\hat{y} \in \mathcal{R}^+$ are the true and estimated sound source distance, respectively. $d \in \{0,1\}$ and $\hat{d} \in [0,1]$ are the true and predicted sound presence, respectively. N is the batch size, and T is the corresponding model output length along the time dimension. $d_{n,t}$ multiplies $\mathfrak{E}$ to avoid distance estimates from contributing to the loss in the absence of sound events.

Reducing the model's absolute or squared error prioritizes the accurate estimation of more distant sound sources. In other words, an error of 0.1 meters is more dramatic if the target is 1 meter away

[2]The survey by Grumiax et al. [13] reviews all relevant SSL literature.

[3]github.com/danielkrause/DCASE2022-data-generator

[4]Other "tasks" in LOCATA feature overlapping sounds

Model	Exp	Best $\mathfrak{E}$	Mean ↓	Median ↓	Std ↓
CRNN	TWL	SPE	0.413	0.330	0.347
	TWA	SE	0.368	0.340	**0.244**
	FWL-S	APE	**0.337**	0.290	0.246
	FWL-D	APE	0.352	**0.269**	0.275
avg pred			0.452	0.410	0.283
[20]			0.448	0.326	0.416

Table 3: Comparison of distance estimation performance on the LOCATA test set across experiments. For each experiment we report the model using the $\mathfrak{E}$ that resulted in the best cross-validation performance. in each experiment is also shown. TWL: train with LOCATA. TWA: train with all. FWL-S: fine-tune with LOCATA from STARSS, FWL-D: fine-tune with LOCATA from DCASE.

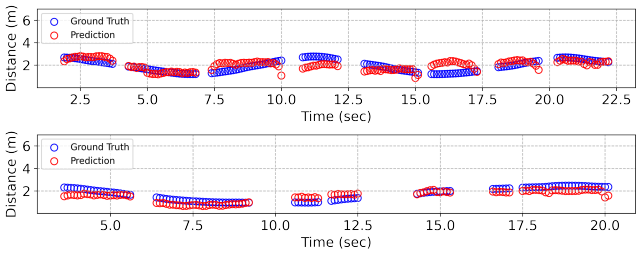

Figure 1: Qualitative comparison between the ground truth and best-model predictions on two excerpts from the LOCATA test set.

versus 10 meters away. Therefore, we also try using the absolute percent error and squared percent error, which should result in a loss that uniformly weighs errors across ground truth distances. Furthermore, we also experiment with the thresholded absolute percent error. Table 2 shows the $\mathfrak{E}$ equations we experiment with.

We keep the model's original input, consisting of a tetrahedral microphone's Log-mel spectrograms and generalized cross-correlation (GCC), capturing the difference in time of signal arrival between microphones. To obtain the tretrahedral microphone (4 channels) from EigenMike (32 channels) recordings, we selected channels 6, 10, 26, and 22, consistent with the STARSS dataset [8].

3.3. Training procedures

First we trained the CRNN to carry out sound detection using the DCASE data. During this phase the model's parameters are trained to optimally output $\hat{d}$, but the $\hat{y}$ output, the distance estimator, remains untrained. We initialized model parameters with the Kaiming method and trained using an Adam optimizer (learning rate $1e-3$) with patience of 40 epochs based on optimal cross-validation performance (15% of recordings randomly separated as a validation set). This resulted on a sound event detector with an $F_1 = 0.94$ on the DCASE test set. We refer to this model as the "pre-trained sound event detector" (PSED)[5].

Next, we initialized the CRNN with PSED parameters and we trained both $\hat{d}$ and $\hat{y}$ using LOCATA. We used Adam (learning rate $1e-3$) with patience of 40 based on optimal performance on a hold-out set consisting of the LOCATA "Task 3 training" files. We refer to this experiment as "Train with LOCATA" (TWL). To study the potential benefit of using a larger training set, we repeated this

[5]Pre-training $\hat{d}$ avoids local minima seen learning $\hat{d}$ and $\hat{y}$ from scratch.

Model	Exp	Best $\mathfrak{E}$	Mean ↓	Median ↓	Std ↓
CRNN	TWD	SE	1.032	0.903	0.838
CRNN	FWD-S	AE	**0.952**	**0.731**	0.834
avg pred			1.014	0.866	**0.596**
CRNN	TWM	SE	1.346	0.417	2.158
CRNN	FWM-S	SPE	**0.811**	**0.405**	0.508
avg pred			1.183	1.611	**0.494**
CRNN	TWT	APE	**0.148**	0.122	**0.126**
CRNN	FWT-S	TAPE*	0.167	**0.114**	0.150
avg pred			0.378	0.289	0.234

Table 4: Performance on diverse datasets compared to the "avg pred" baseline. D=DCASE. M=MARCo. T=METU-SPARG. TWD and FWD-S highlight generalization to new rooms at test time. *TAPE with threshold of 0.01

experiment but substituted the training set to be all the training data across datasets listed in Table 1. We refer to this experiment as "Train with all data" (TWA).

Compared to the DCASE and STARSS datasets, LOCATA is very small. Therefore, we also experimented with using LOCATA to fine-tune a model pre-trained with a larger dataset. We first initialized the CRNN with the PSED parameters to train both $\hat{d}$ and $\hat{y}$ using the STARSS dataset. We used Adam (learning rate $1e-3$) with patience of 40 based on optimal cross-validation performance on STARSS (15% of recordings randomly separated as a validation set). We refer to this as the "STARSS pre-trained model" (SPTM). Next, we initialized the CRNN with the SPTM parameters to train both $\hat{d}$ and $\hat{y}$ using LOCATA. We used Adam (learning rate $1e-3$) with patience of 40 based on optimal performance on a hold-out set consisting of the LOCATA "Task 3 training" files. We refer to this experiment as "Fine-tune with LOCATA from STARSS" (FWL-S). We also carried out this procedure using DCASE instead of STARSS, resulting in an experiment called "Fine-tune with LOCATA from DCASE" (FWL-D). Each experiment is run seven times with a different regressor $\mathfrak{E}$: AE, SE, APE, SPE, TAPE($\delta = 0.01$), TAPE($\delta = 0.1$), and TAPE($\delta = 0.2$), all listed in Table 2.

3.4. Baselines for comparison and metrics

We compare performance on the LOCATA test set against the average sound source distance in the LOCATA training set ("avg pred"), and the recent signal processing approach by Daniel et al. [20]. It is worth noting that [20] did not compare against a baseline since they consider their approach to be the first to not make assumptions about a room's DRR [20]. In our own literature review, we did not find other distance estimation approaches evaluated on any of the open-source datasets that we consider in this study. We use the same metrics used in the study by Daniel et al. [20], which are the mean, median, and standard deviation of the model's absolute-valued distance estimate error.

4. RESULTS

Table 3 shows that both "avg pred" and Daniel et al. [20] baselines have similar "mean" metrics. Thus it is possible that the method by Daniel et al. is correlated with the global statistics of the LOCATA training data [20]. Table 3 also shows that all our experiments resulted in a model that outperforms both baselines.

The "fine-tuning" experiments (FWL-S and FWL-D) yielded the best performance according to the "mean" and "median" metrics. To understand this pattern, let's remember that STARSS, DCASE and LOCATA consist of recordings in real rooms with humans producing sounds (i.e., speech, footsteps, etc.) around a microphone. However, STARSS and DCASE have much more room diversity (16 rooms and 9 rooms, respectively) than LOCATA (1 room). Therefore, initializing the CRNN with the SPTM model parameters (or the DCASE equivalent) may be providing with an initial representation of multi-room reverberation, from where it is easier to find the parameters to optimally perform in the acoustic conditions of the LOCATA room. Figure 1 qualitatively compares predictions made by the best FWL-S model versus ground truth. Close alignment is observed, with errors still tracing the ground truth contour.

The best models in Table 3 may be overfitting. Contrasting "fine-tuning" experiments with TWL yields insight into this issue. TWL initializes the CRNN with PSED parameters, resulting in a distance estimator $\hat{d}$ that learns this task only on the LOCATA data. This makes overfitting to the LOCATA training set likely and we do see poorer performance at test time. The better-performing TWA (based on the "mean" metric) shows the benefit of using more training data and significant mitigation of overfitting compared to TWL.

To further study this issue, we repeated the "Train with LOCATA" and "Fine-tune with LOCATA from STARSS" experiments with the other datasets:DCASE, MARCo and METU-SPARG. Table 4 shows the results. Compared to their "avg pred" baseline, we again see the benefit of initializing the CRNN with SPTM parameters vs PSED (on DCASE and MARCo according to the "mean" metric). However, this was not the case for METU-SPARG. This can be explained by its small size and statistical properties that are virtually the same across training and test splits. Thus, overfitting to train data results in good performance on the test split.

Tables 3 and 4 also show what specific $\mathfrak{E}$ resulted in the best model. In general, the "percentage" $\mathfrak{E}$s were better. For FWL-S, we analyzed the effect of different $\mathfrak{E}$ (Table 5). We observe that the "percentage" $\mathfrak{E}$s (APE, SPE, and TAPE) result in improved performance compared to AE and SE. This makes sense, as APE, SPE, and TAPE uniformly weight errors as a function of ground truth distance. Figure 2 visualizes this effect by plotting the mean FWL-S model error as a function of ground truth distance for AE, APE, TAPE($\delta = 0.01$), and TAPE($\delta = 0.2$) on the LOCATA test-set. Note also how AE tries to reduce errors associated with more distant sound sources and underperforms for sound sources that are closer to the microphone. In contrast, "percentage" $\mathfrak{E}$s reduce prediction errors for targets closer to the microphone. In general, performance deteriorates as a function of ground truth distance due to attenuation and arrival likely to be closely-followed by reverberations.

5. CONCLUSION AND FUTURE WORK

We have proposed a model and optimization routine to carry out sound source distance estimation, which is an understudied component of SSL. Our solution is a CRNN with two outputs: a distance estimator and a sound event detector. Experiments revealed the benefit of using a loss function that uniformly weighs the model's estimate error across the full range of distances by converting it into a percentage of the ground truth distance. We also observe how the model tends to overfit to specific datasets, and the benefit of training with larger datasets featuring diverse acoustic conditions. To carry out this study, we have annotated sound source distances in a large collection of open-source datasets and a data generator,

$\mathfrak{E}$	Mean ↓	Median ↓	Std ↓
AE	0.438	0.360	0.342
SE	0.374	0.319	0.256
APE	0.337	0.290	0.246
SPE	0.334	0.292	0.259
TAPE ($\delta = 0.01$)	0.322	0.248	0.261
TAPE ($\delta = 0.10$)	0.361	0.312	0.250
TAPE ($\delta = 0.20$)	0.346	0.282	0.260

Table 5: The effect of different $\mathfrak{E}$ on the best-performing CRNN on the LOCATA test split. FWL-S experiments are shown since those yielded the best model. Note how TAPE($\delta = 0.01$) has the overall best test-set performance. However, best model selection was agnostic of the test set, and was based on cross-validation performance to maximally emulate a real testing scenario.

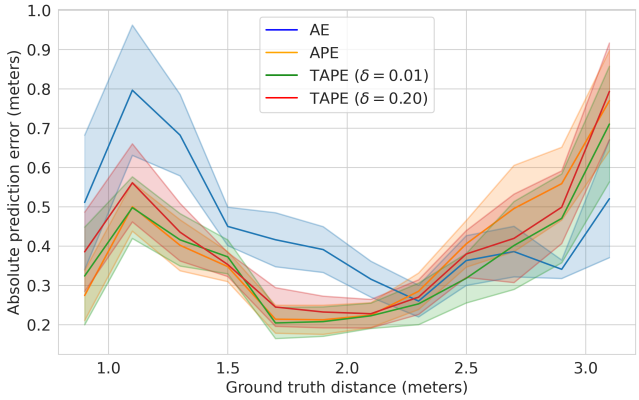

Figure 2: The effect of regressors $\mathfrak{E}$ on the CRNN distance estimate error as a function of ground truth distance (on the LOCATA test set in experiment FWL-S). Lines correspond to the CRNN's average error. The 95% confidence interval of the mean is shown.

which we openly-release for future research by the broader community. In the future, we plan to expand this study by including more open-source datasets and adding more rooms to the data generator. Future work could also investigate whether using the GTVV [25] as an additional or unique input feature to the model could improve performance. Similarly, other features like spectral magnitude difference [23] and signal coherence [24] or alternative input formats like larger microphone arrays, HOA or binaural audio could be used. Model architectures such as transformers and conformers could also be explored.

Finally, a major shortcoming of the model presented here is its inability to track the distance of simultaneously-occurring sound sources. Recent solutions to this issue have been proposed in the DOA estimation literature [12], which could be applied to expand our approach. Ultimately, we aim to develop a method that can jointly carry out the tasks of classification, localization, and distance estimation while being robust to different acoustic conditions.

6. ACKNOWLEDGEMENTS

This work is supported by the National Science Foundation grant no. IIS-1955357. The authors thank the funding source and their grant collaborators, particularly Bea Steers, who helped proofreading this manuscript.

7. REFERENCES

[1] M. Wax and T. Kailath, "Optimum localization of multiple sources by passive arrays," *IEEE trans. on Acoustics, Speech, and Signal Processing*, vol. 31, no. 5, pp. 1210–1217, 1983.

[2] Y. Huang, J. Benesty, G. W. Elko, and R. M. Mersereati, "Real-time passive source localization: A practical linear-correction least-squares approach," *IEEE trans. on Speech and Audio Processing*, vol. 9, no. 8, pp. 943–956, 2001.

[3] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "SELD of overlapping sources using CRNNs," *IEEE Journal of Sel. Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48, 2018.

[4] T. Jenrungrot, V. Jayaram, S. Seitz, and I. Kemelmacher-Shlizerman, "The cone of silence: Speech separation by localization," *Advances in Neural Information Processing Systems*, vol. 33, pp. 20 925–20 938, 2020.

[5] C. Chen, U. Jain, C. Schissler, S. V. A. Gari, Z. Al-Halah, V. K. Ithapu, P. Robinson, and K. Grauman, "Soundspaces: Audio-visual navigation in 3d environments," in *Proceedings of the Computer Vision–ECCV (Part VI 16)*, 2020, pp. 17–36.

[6] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in *2007 IEEE Conference on Advanced Video and Signal Based Surveillance*, 2007, pp. 21–26.

[7] 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 Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1620–1643, 2020.

[8] A. Politis, K. Shimada, P. Sudarsanam, A. Hakala, S. Takahashi, D. A. Krause, N. Takahashi, S. Adavanne, Y. Koyama, K. Uchida, Y. Mitsufuji, and T. Virtanen, "STARSS23: Sony-TAu Realistic Spatial Soundscapes 2023," Mar. 2023.

[9] S. Adavanne, A. Politis, and T. Virtanen, "A multi-room reverberant dataset for SELD," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop*, 2019, pp. 10–14.

[10] A. Politis, S. Adavanne, D. Krause, A. Deleforge, P. Srivastava, and T. Virtanen, "A dataset of dynamic reverberant sound scenes with directional interferers for SELD," in *Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events Workshop*, 2021, pp. 125–129.

[11] T. N. T. Nguyen, K. N. Watcharasupat, N. K. Nguyen, D. L. Jones, and W.-S. Gan, "Salsa: Spatial cue-augmented log-spectrogram features for polyphonic SELD," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1749–1762, 2022.

[12] K. Shimada, Y. Koyama, S. Takahashi, N. Takahashi, E. Tsunoo, and Y. Mitsufuji, "Multi-accdoa: Localizing and detecting overlapping sounds from the same class with auxiliary duplicating permutation invariant training," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 316–320.

[13] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "A survey of sound source localization with deep learning methods," *The Journal of the Acoustical Society of America*, vol. 152, no. 1, pp. 107–151, 2022.

[14] H. Liu, Y. Chen, Y. Huang, X. Cheng, and Q. Xiao, "Study on the localization method of multi-aperture acoustic array based on tdoa," *IEEE Sensors Journal*, vol. 21, no. 12, pp. 13 805–13 814, 2021.

[15] H. Chen, T. D. Abhayapala, P. N. Samarasinghe, and W. Zhang, "Direct-to-reverberant energy ratio estimation using a first-order microphone," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 2, pp. 226–237, 2016.

[16] M. Yiwere and E. J. Rhee, "Distance estimation and localization of sound sources in reverberant conditions using deep neural networks," *International Journal of Applied Engineering Research*, vol. 12, no. 22, pp. 12 384–12 389, 2017.

[17] R. Roden, N. Moritz, S. Gerlach, S. Weinzierl, and S. Goetze, *On sound source localization of speech signals using deep neural networks*. Technische Universität Berlin, 2019.

[18] R. Takeda and K. Komatani, "Sound source localization based on deep neural networks with directional activate function exploiting phase information," in *International conf. on acoustics, speech and signal processing*. IEEE, 2016, pp. 405–409.

[19] G. Bologni, R. Heusdens, and J. Martinez, "Acoustic reflectors localization from stereo recordings using neural networks," in *International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2021, pp. 1–5.

[20] J. Daniel and S. Kitic, "Echo-enabled direction-of-arrival and range estimation of a mobile source in ambisonic domain," in *2022 30th European Signal Processing Conference (EUSIPCO)*. IEEE, 2022, pp. 852–856.

[21] C. W. Sheeline, *An investigation of the effects of direct and reverberant signal interaction on auditory distance perception*. Stanford University, 1983.

[22] P. Zahorik, "Direct-to-reverberant energy ratio sensitivity," *The Journal of the Acoustical Society of America*, vol. 112, no. 5, pp. 2110–2117, 2002.

[23] E. Georganti, T. May, S. Van De Par, and J. Mourjopoulos, "Sound source distance estimation in rooms based on statistical properties of binaural signals," *IEEE transactions on audio, speech, and language processing*, vol. 21, no. 8, pp. 1727–1741, 2013.

[24] S. Vesa, "Sound source distance learning based on binaural signals," in *Workshop on Applications of Signal Processing to Audio and Acoustics*, 2007, pp. 271–274.

[25] S. Kitić and J. Daniel, "Generalized time domain velocity vector," in *International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2022, pp. 936–940.

[26] H. Lee and D. Johnson, "An open-access database of 3d microphone array recordings," in *Audio Engineering Society Convention 147*. Audio Engineering Society, 2019.

[27] O. Olgun and H. Hacihabiboglu, "METU SPARG Eigenmike em32 Acoustic Impulse Response Dataset v0.1.0," Apr. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.2635758

[28] 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," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1251–1264, 2023.