How to (Virtually) Train Your Speaker Localizer

Prerak Srivastava, Antoine Deleforge, Archontis Politis, Emmanuel Vincent

HOW TO (VIRTUALLY) TRAIN YOUR SPEAKER LOCALIZER

Prerak Srivastava[1]*, Antoine Deleforge*[1]*, Archontis Politis*[2]*, Emmanuel Vincent*[1]

[1]Université de Lorraine, CNRS, Inria, Loria, F-54000 Nancy, France
[2]Audio and Speech Processing Research Group, Tampere University, Finland
[1]{prerak.srivastava, antoine.deleforge, emmanuel.vincent}@inria.fr
[2]archontis.politis@tuni.fi

Abstract

Learning-based methods have become ubiquitous in speaker localization. Existing systems rely on simulated training sets for the lack of sufficiently large, diverse and annotated real datasets. Most room acoustics simulators used for this purpose rely on the image source method (ISM) because of its computational efficiency. This paper argues that carefully extending the ISM to incorporate more realistic surface, source and microphone responses into training sets can significantly boost the real-world performance of speaker localization systems. It is shown that increasing the training-set realism of a state-of-the-art direction-of-arrival estimator yields consistent improvements across three different real test sets featuring human speakers in a variety of rooms and various microphone arrays. An ablation study further reveals that every added layer of realism contributes positively to these improvements.

Index Terms: source localization, direction-of-arrival, image source, directivity, room acoustic simulation

1. Introduction

Far-field deep learning based speech processing systems often require large amount of training data, as demonstrated by their application to various tasks such as speech recognition [1, 2], speaker localization [3, 4], speech enhancement [5, 6] and diarization [7, 8]. For increased generalization, the study in [9] suggest the use of extensive training sets that cover the range of variability found in real test sets. Due to the difficulty of obtaining enough real data with enough diversity, these training sets are typically simulated, an approach called *virtually supervised learning* in, e.g., [10, 11]. For those applications on which the effect of reverberation is detrimental, such as speaker localization or multi-microphone speech enhancement, the use of room acoustics simulators is widespread since it allows physical modeling of the room impulse response (RIR) from the source to the receivers, which includes the inter-channel differences utilized by the methods as well as reverberation. Depending on the type of simulation method used, diverse data can be generated in terms of source and receiver position, acoustic material characteristics, or room geometry, which is necessary for the generalization of the methods to a wide range of real acoustic scenes.

Among the various room acoustics simulators under use, the most widespread by far rely on the image source method (ISM) for shoebox rooms [12, 13]. Its implementation simplicity and speed allow rapid generation of RIRs for thousands of rooms with randomized dimensions, wall absorption coefficients, and source-receiver positions. More elaborate geometrical [14] or wave-based room acoustics simulators [15], that allow complex geometries and more accurate modeling of propagation effects, are not currently fast enough for this. In the context of supervised learning, they have been used to pre-compute a few large-scale single-channel datasets covering a limited set of conditions only [16, 17]. Even though shoebox ISM simulation employs stronger acoustical simplifications than more advanced room acoustics simulation methods, it has proven useful in training localization models that then perform adequately well on real datasets [18, 19, 20, 21, 22]. Most of these studies are using the simplest form of shoebox ISM, namely, broadband omnidirectional sources and receivers and a global wall- and frequency-independent absorption coefficient, which is also the fastest to simulate. More realistic simulation conditions, such as directional sources and receivers and frequency-dependent surface absorption profiles, can be integrated into shoebox ISM with little computational overhead, at the expense of more complex implementation and specification of simulation parameters. Very few studies perform simulations under these more realistic conditions, except when specialized multichannel receivers are used, e.g., spherical [23], Ambisonic [19, 20], or binaural [24, 25].

Despite the widespread use of shoebox ISM-based simulators for training speaker localization systems, the effect of integrating more realistic simulation conditions at training time on the localization performance on real recordings at test time has hardly been studied. Receiver directivity does not only concern specialized multichannel receivers, but also common arrays of omnidirectional microphones above some frequency, due to the microphone mounting or the size of the capsule. Source directivity seems also crucial considering that most methods focus on speech signals, with human speakers being highly directive [26]. Realistic surface absorption profiles can also have a drastic effect on the reverberation reaching the microphones, in terms of spatial distribution and power spectrum. A previous work by the authors showed a significant performance increase on an acoustic parameter estimation task when measured source and receiver directivities and a natural distribution of frequency-dependent absorption coefficients were integrated in the simulations [11]. In the context of speaker localization, we are only aware of one recent study [27] which analyses the effects of speaker directivity and diffuse late reverberation modeling in the simulations. While diffusion showed no significant impact, the source directivity was found to have a positive impact when testing on speech signals convolved with measured directive RIRs.

In this work, we claim that more realistic simulation at training time can significantly improve real-world localization performance, in a wide variety of scenarios. Results on three multichannel datasets of real human speakers captured in various rooms and with various microphone arrays support this claim. The effects of realistic source and receiver directivi-

ties and surface absorption profiles at training time are carefully studied in isolation and in combination for each dataset.

The structure of the rest of the paper is as follows. Section 2 introduces the *naive* and *advanced* ISM-based simulation modes and Section 3 introduces the real test sets considered in this paper. The experimental setup is described in Section 4 and the results are analyzed in Section 5. We conclude in Section 6.

2. Image Source Simulation

2.1. Generalized Image Source Method

Simulation of multichannel reverberant speech can be expressed as

$$\mathbf{x}[t] = (\mathbf{h} * s)[t] + \mathbf{n}[t] \tag{1}$$

where $*$ denotes convolution, $\mathbf{h}(t) = [h_1(t), ..., h_M(t)]^{\mathrm{T}}$ the vector of RIRs from the source to M microphones, $\mathbf{n}(t) = [n_1(t), ..., n_M(t)]^{\mathrm{T}}$ additive noise and s the dry source signal. RIR generation based on the ISM for shoebox geometries requires at least the dimensions of the (cuboid) room, the source and receiver coordinates, and a global wall absorption coefficient. An extended version of the ISM that takes into account frequency-dependent walls, propagation, and directivity effects can be expressed in the frequency domain as

$$\widehat{\mathbf{h}}(f) = \sum_{k=0}^{K} \frac{\exp(-\jmath 2\pi f r_k / c F_\mathrm{s})}{r_k} \cdot d_\mathrm{air}(r_k, f) \cdot d_k(f) \cdot \widehat{g}_\mathrm{src}(-\widetilde{\mathbf{r}}_k, f) \cdot \widehat{\mathbf{g}}_\mathrm{mic}(\widetilde{\mathbf{r}}_k, f). \tag{2}$$

Here, c is the speed of sound and F_s is the sampling frequency. $\mathbf{r}_k$ is the position vector of the k-th image source w.r.t. the microphone array center, $r_k = ||\mathbf{r}_k||$ is the image-source-to-array distance and $\widetilde{\mathbf{r}}_k = \mathbf{r}_k / r_k$ is the direction unit vector. $d_\mathrm{air}(r, f)$ is the distance-dependent air attenuation while $d_k(f)$ is the compound reflection coefficient of all the surfaces that the k-th image has been reflected from. $\widehat{\mathbf{g}}_\mathrm{mic}(\widetilde{\mathbf{r}})$ denotes the vector of M microphone directivity responses defined with their phase center at the array center, for direction-of-arrival (DOA) $\widetilde{\mathbf{r}}$. $\widehat{g}_\mathrm{src}(-\widetilde{\mathbf{r}})$ denotes the source directivity response for direction-of-departure $-\widetilde{\mathbf{r}}$. An alternative implementation of the multichannel receivers can instead utilize individual positions, distances, and DOAs for each microphone in the array, and integrate its local directivity excluding inter-channel array propagation effects. That implementation is suitable for open arrays of directional microphones of known directivity, but it is not suitable for more complex directional arrays that include scattering effects, such as spherical arrays or head-related transfer functions.

2.2. Advanced v/s Naive Simulation

In the following, we distinguish the *naive* mode of ISM simulation often used in practice for model training, where *a)* wall absorption is assumed to be frequency-independent and equal for all 6 room surfaces, i.e., $d_k(f) = d^{o_k}$ with o_k being the reflection order and *b)* sources and receivers are assumed to be omnidirectional, i.e., $\widehat{g}_\mathrm{src}(\widetilde{\mathbf{r}}, f) = \widehat{\mathbf{g}}_\mathrm{mic}(\widetilde{\mathbf{r}}, f) = 1$. Additionally we define an *advanced* mode of ISM simulation that incorporates more informed choices on the directivity and absorption components. Regarding absorption, the coefficients in 6 octave bands are drawn from a naturally balanced mix of distributions corresponding to reflective and absorptive wall, ceiling, and floor materials, as described in [10, 11]. These coefficients are then interpolated using half-cosine octave bands in the discrete Fourier domain and turned into minimum-phase responses. Note that both modes of simulation are tuned to yield comparable distributions of reverberation time. Regarding source directivity, the spatially interpolated measured directivities of a head-and-torso-with-mouth simulator (Brüel & Kjaer HATS 4128-C) and two directive loudspeakers (Genelec 8020 and YAMAHA DXR8) taken from the DIRPAT dataset [28] are integrated into the simulation. Regarding microphone array directivities, scenario-based informed choices are made, as detailed in Section 4. These extensions of the ISM are detailed in [11] and are currently available as open source code in the branch `dev/dirpat` of the pyroomacoustics simulator [29].

3. Real Test Sets

To examine the impact of increased ISM realism at training time, we evaluate virtually-supervised localization performance on three real datasets of human speakers with spatio-temporal annotations of their activity and position with respect to the microphone array. The selected publicly available datasets are captured in a variety of rooms, and with a different microphone array each. We focus specifically on datasets featuring real human speakers as opposed to ones generated by convolving dry speech signals with measured RIRs. While the latter are commonly used in the speaker localization literature, the former are closer to real world conditions. This study focuses on the most elementary localization task, namely, single-source DOA estimation in $[0, 180]^\circ$ from a two-second signal recorded using a single microphone pair.

3.1. VoiceHome-2 [30]

This dataset is specifically made for distant speech processing applications in domestic environments. It consists of short commands for smart home devices in French, collected in reverberant conditions and uttered by 12 native French speakers. The data is recorded in 12 different rooms corresponding to 4 houses, with fully annotated geometry, under quiet or noisy conditions. It is captured by a microphone array consisting of 8 MEMS placed near the corner of a cubic baffle. For this study, a two-channel sub-array with aperture 10.4 cm is selected, and 360 two-second speech recordings in quiet conditions are used.

3.2. DIRHA [31]

This multichannel dataset consists of recordings done in the living room and kitchen of a typical apartment. Microphones in different configurations are placed on the walls and ceiling of the 2 rooms. 6 native English speakers are chosen to speak sentences taken from the Wall Street Journal news text corpus. Annotations of individual microphone positions and speaker positions are provided. For this study, a wall-mounted two-channel microphone array with aperture 30 cm placed in the living room is selected, and 410 two-second speech recordings from the living room are used.

3.3. STARSS22 [32]

This dataset contains recordings of naturally acted scenes of human interaction with spatio-temporal annotations of events belonging to 13 target classes, of which speech is a dominant one. This corpus is part of the development set for Task 3 of the DCASE Challenge 2022. It was captured in facilities at Tampere University and at Sony, with the recording, annotation,

and organization of acoustic scenes kept similar on both sites. The Eigenmike spherical microphone array is used to deliver the dataset in two spatial formats, one of which is a tetrahedral sub-array of omnidirectional microphones mounted on a rigid spherical baffle. The corpus is more challenging than the other two in the sense that speakers are free to move and turn naturally during discussions, and that it contains intentional and unintentional sound events other than speech with diffuse and directional ambient noise at substantial levels. We carefully preprocessed the data to extract 2,100 two-second non-overlapping speech excerpts from microphones 6 and 10 out of the tetrahedral sub-array, with an aperture of 6.8 cm.

The three curated test sets add up to a total of 95 minutes DOA-annotated, two-channel, real human speech recordings.

4. Scenario-Based Simulation and Training

4.1. Model

To select a state-of-the-art learning-based localization method for this study, we examined the extensive literature review in [4]. We selected the convolutional neural network architecture proposed by He et al. [33], specifically its most recent version in [34], due to its multiple distinguishing features. Namely, it works with arbitrary microphone arrays, it is initially designed for DOA estimation, it has been successfully applied to real speech recordings, it is readily extendable to multiple localization, detection and counting (not addressed in this study), and its architecture is available through open source code. The model is trained over different simulated training sets using the ADAM optimizer with a learning rate of 10^{-4} and batches of size 16 for a maximum of 110 epochs, with early stopping on validation sets. We used the same input features as in [34], namely, concatenated short-time Fourier transforms with 50% overlap and 42.7 ms windows, except that all the signals considered in this study are down-sampled to 16 kHz, for consistency.

4.2. Scenario Based Training

For all the simulated training sets considered, the default air absorption model of pyroomacoustics is used for d_{air} and image sources are simulated up to order 20. A total of 40k shoebox rooms of sizes uniformly drawn at random in $[3, 10] \times [3, 10] \times [2, 4.5]$ m are simulated, each containing a source and a two-microphone array placed uniformly at random with a minimum source-array and device-wall distance of 30 cm. The obtained RIRs are convolved with speech excerpts from the Librispeech corpus [35] according to (1), yielding 40k two-second two-channel reverberated speech samples, of which 38k are used for training and 2k for validation. We experimented with supervising the model with sets containing 10k to 60k samples. While a strong performance improvement was observed from 10k to 60k, diminishing returns were hit around 40k. Further improvements may nonetheless be achievable using even larger sets.

Uncorrelated white Gaussian noise and diffuse speech-shaped noise convolved with the late part of a random RIR in the same room are added to the reverberated signals. As detailed in [11], noise levels are tuned based on reference scenarios according to the source and receiver considered. This yields consistent bell-shaped signal-to-noise ratio distributions in the range $[15, 75]$ dB with a peak at 40 dB for all of the training sets considered in this study. While a detailed analysis of the specific impact of noise at training time would be of interest, this is out of the scope of this paper. For each of the three real test sets described in Section 3, one *naive* and one *advanced* simulated training set is built based on the two modes of simulation described in Section 2 for walls, sources and receivers. For naive sets, ideal omnidirectional receivers are placed at random inside the room using the same apertures as the ones of their corresponding test sets. For VoiceHome-2, the simulated arrays are identical in the naive and advanced sets. This is because MEMS are known to be close to omnidirectional and the directivity of the VoiceHome-2 array is not available. Moreover, early experiments revealed that using mismatched microphone responses at training time was detrimental to the results, a phenomenon also reported in [11]. For DIRHA, the advanced simulation places the arrays *on* the room walls, which is equivalent to simulating microphones with a half-sphere directivity. For STARSS22, the advanced simulation uses the measured directivity pattern of the relevant sub-array of the Eigenmike.[1]

5. Experiments and Results

To evaluate the virtually-supervised localization systems, two complementary metrics are used for each test set, namely, the mean angular error (MAE, in degrees) and the ratio of sources localized with an error less than 10° (Recall, in %), which showed to be an adequate threshold to prune out outliers. Models trained using naive and advanced simulations are compared to the classical learning-free steered response power with phase transform (SRP-PHAT) localization method, as implemented in [29].

5.1. Simulated Test Sets

We start by comparing the three methods on two test sets simulated in naive and advanced modes under the VoiceHome-2 scenario. The results are shown in Table 1. First, while all the methods perform well on the naive test set, with nearly perfect recall achieved by both trained models, their performance drastically drops on the advanced test set. This suggests that the presence of realistic wall, source and receiver responses significantly hardens the localization task, even under identical noise and reverberation-time distributions. We have not found direct evidence of this in prior literature and believe this could provide a helpful guideline to improve the evaluation of localization methods on synthetic datasets. Second, as expected, the learning-based methods strongly outperform the learning-free one and the advanced model generalizes significantly better to advanced conditions than the naive one. Moreover, the former seems to perform nearly as well as the latter on naive conditions, despite the mismatch. This strengthens the evidence that speaker localization is inherently more challenging in more realistic conditions.

Table 1: *Localization results on naive and advanced simulated test sets following the VoiceHome-2 scenario.*

	Simulated Test Sets			
	Naive		Advanced	
Training	↑ Recall	↓ MAE	↑ Recall	↓ MAE
Naive	96%	2.6°	74%	8.5°
Advanced	95%	3.0°	80%	6.7°
SRP-PHAT	75%	11.1°	50%	20.8°

[1] We used the measured directivity made available by Franz Zotter et al. from Graz University: `https://phaidra.kug.ac.at/o:69292`.

Table 2: *Localization results on three real test sets achieved by the SRP-PHAT baseline and by the supervised model [34] trained using various simulation modes. Mean angular errors (MAE) are displayed with their* 95% *confidence interval. Bold numbers indicate the best system in each column and the systems statistically equivalent to it. Statistical significance was assessed using McNemar's test for the Recall metric and* 95% *confidence intervals over angular error differences for the MAE metric.*

Real Test Sets $\rightarrow$	**VoiceHome-2** [30]		**DIRHA** [31]		**STARS22** [32]	
Methods	$\uparrow$ Recall	$\downarrow$ MAE ($^\circ$)	$\uparrow$ Recall	$\downarrow$ MAE ($^\circ$)	$\uparrow$ Recall	$\downarrow$ MAE ($^\circ$)
SRP-PHAT	70%	9.9 ± 1.5	61%	15.0 ± 2.3	45%	14.9 ± 0.6
Naive Training	78%	7.6 ± 1.2	77%	8.4 ± 1.4	57%	12.9 ± 0.6
Advanced Training	$\mathbf{85}\%$	$\mathbf{5.8 \pm 0.8}$	$\mathbf{84}\%$	$\mathbf{6.3 \pm 1.0}$	61%	$\mathbf{11.4 \pm 0.5}$
Ablation study						
without wall realism	83%	6.2 ± 0.8	81%	7.5 ± 1.4	59%	12.1 ± 0.6
without source realism	82%	7.1 ± 1.1	80%	7.8 ± 1.2	$\mathbf{63}\%$	$\mathbf{11.4 \pm 0.6}$
without receiver realism	N/A	N/A	78%	8.3 ± 1.5	53%	13.4 ± 0.6

5.2. Real Test Sets

The three methods are compared on the three real datasets. As can be seen in the top part of Table 2, advanced training significantly outperforms naive training by 4 to 7 recall points and 2° MAE margins across all three datasets, despite using the exact same network architecture. It also largely outperforms the classical SRP-PHAT baseline by 15 to 23 recall points and 3° to 9° MAE margins. As predicted in Section 3, the STARS22 dataset proved most challenging for all the methods. Note that even under the quiet and static conditions of VoiceHome-2 and DIRHA, the baseline results are far from perfect. This shows that two-channel DOA estimation remains a challenging task in real-world settings.

The second half of Table 2 presents an ablation study on the proposed advanced simulation strategy. It reveals that removing any of the three considered layers of realism results in noticeable performance loss. One exception is the use of measured source directivity on STARS22, which does not seem to affect performance. One explanation could be that the human speakers in STARS22 perform significant head rotations, which is not modeled by our framework. The use of a measured array directivity seems to have the strongest impact on this dataset, an observation which we have not found in previous literature for such a simple two-element array. On the other two datasets, the positive impact of source directivity previously reported in [27] is confirmed, while realistic wall absorptions seem to offer a comparable boost in performance. This is new to the best of our knowledge, and may be explained by the presence of diverse real-world rooms in these datasets. The Python code to reproduce these experiments from training data simulation to evaluation is available here: `github.com/prerak23/Dir_SrcMic_DOA`.

6. Conclusion

This paper revealed that simulating realistic wall absorption and source/receiver directivities at training time can significantly boost the performance of a virtually supervised speaker localization model across a large test corpus, featuring real human speakers and a variety of microphone arrays and rooms. While these aspects have been mostly ignored in the speaker localization literature, we argue that they can critically benefit both the evaluation of models and their real world applicability. Several research avenues are left to explore. Our preliminary findings revealed that the results are sensitive to the noise distribution at training time, calling for a careful dedicated study of such effects. We also observed that the validation learning curves of models trained on advanced simulation tended to peak earlier than the naive ones. This leads us to believe that there is still room for improvement on the reported results, e.g., by enriching the diversity of directivity profiles through data augmentation. Finally, the inclusion of source and receiver movements at training time or the use of more efficient stochastic late reverberation models constitute worthwhile research directions.

7. Acknowledgements

This work was made with the support of the French National Research Agency through project HAIKUS "Artifical Intelligence applied to augmented acoustic scenes" (ANR-19-CE23-0023). Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see https://www.grid5000.fr).

8. References

[1] A. Gusev, V. Volokhov, T. Andzhukaev, S. Novoselov, G. Lavrentyeva, M. Volkova, A. Gazizullina, A. Shulipa, A. Gorlanov, A. Avdeeva, A. Ivanov, A. Kozlov, T. Pekhovsky, and Y. Matveev, "Deep speaker embeddings for far-field speaker recognition on short utterances," *arXiv preprint arXiv:2002.06033*, 2020.

[2] J. Huang and T. Bocklet, "Intel far-field speaker recognition system for VOiCES Challenge 2019," in *Interspeech*, 2019, pp. 2473–2477.

[3] W. Xue, Y. Tong, C. Zhang, G. Ding, X. He, and B. Zhou, "Sound event localization and detection based on multiple DOA beamforming and multi-task learning," in *Interspeech*, 2020, pp. 5091–5095.

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

[5] Y. Chen, Y. Hsu, and M. R. Bai, "Multi-channel end-to-end neural network for speech enhancement, source localization, and voice activity detection," *arXiv preprint arXiv:2206.09728*, 2022.

[6] W. Rao, Y. Fu, Y. Hu, X. Xu, Y. Jv, J. Han, Z. Jiang, L. Xie, Y. Wang, S. Watanabe, Z.-H. Tan, H. Bu, T. Yu, and S. Shang, "ConferencingSpeech Challenge: Towards far-field multi-channel speech enhancement for video conferencing," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 679–686.

[7] S. Horiguchi, Y. Fujita, S. Watanabe, Y. Xue, and K. Nagamatsu, "End-to-end speaker diarization for an unknown number of speakers with encoder-decoder based attractors," *arXiv preprint arXiv:2005.09921*, 2020.

[8] N. Ryant, P. Singh, V. Krishnamohan, R. Varma, K. Church, C. Cieri, J. Du, S. Ganapathy, and M. Liberman, "The third DIHARD diarization challenge," *arXiv preprint arXiv:2012.01477*, 2020.

[9] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech and Language*, vol. 46, pp. 535–557, 2017.

[10] C. Foy, A. Deleforge, and D. Di Carlo, "Mean absorption estimation from room impulse responses using virtually supervised learning," *Journal of the Acoustical Society of America*, vol. 150, no. 2, pp. 1286–1299, 2021.

[11] P. Srivastava, A. Deleforge, and E. Vincent, "Realistic sources, receivers and walls improve the generalisability of virtually-supervised blind acoustic parameter estimators," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2022.

[12] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[13] P. M. Peterson, "Simulating the response of multiple microphones to a single acoustic source in a reverberant room," *Journal of the Acoustical Society of America*, vol. 80, no. 5, pp. 1527–1529, 1986.

[14] S. Siltanen, T. Lokki, S. Kiminki, and L. Savioja, "The room acoustic rendering equation," *Journal of the Acoustical Society of America*, vol. 122, no. 3, pp. 1624–1635, 2007.

[15] S. Bilbao, "Modeling of complex geometries and boundary conditions in finite difference/finite volume time domain room acoustics simulation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1524–1533, 2013.

[16] C. Chen, C. Schissler, S. Garg, P. Kobernik, A. Clegg, P. Calamia, D. Batra, P. W. Robinson, and K. Grauman, "SoundSpaces 2.0: A simulation platform for visual-acoustic learning," *arXiv preprint arXiv:2206.08312*, 2022.

[17] Z. Tang, R. Aralikatti, A. J. Ratnarajah, and D. Manocha, "GWA: A large high-quality acoustic dataset for audio processing," in *ACM SIGGRAPH Conference*, 2022.

[18] S. Chakrabarty and E. A. 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.

[19] 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, 2018.

[20] 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.

[21] 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.

[22] 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.

[23] Y. Koyama, K. Shigemi, M. Takahashi, K. Shimada, N. Takahashi, E. Tsunoo, S. Takahashi, and Y. Mitsufuji, "Spatial data augmentation with simulated room impulse responses for sound event localization and detection," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2022, pp. 8872–8876.

[24] J. Ding, Y. Ke, L. Cheng, C. Zheng, and X. Li, "Joint estimation of binaural distance and azimuth by exploiting deep neural networks," *Journal of the Acoustical Society of America*, vol. 147, no. 4, pp. 2625–2635, 2020.

[25] C. Gaultier, S. Kataria, and A. Deleforge, "VAST: The virtual acoustic space traveler dataset," in *Int. Conf. on Latent Variable Analysis and Signal Separation*, 2017, pp. 68–79.

[26] R. Gonzalez, T. McKenzie, A. Politis, and T. Lokki, "Near-field evaluation of reproducible speech sources," *Journal of the Audio Engineering Society*, vol. 70, no. 7/8, pp. 621–633, 2022.

[27] F. B. Gelderblom, Y. Liu, J. Kvam, and T. A. Myrvoll, "Synthetic data for DNN-based DoA estimation of indoor speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2021, pp. 4390–4394.

[28] M. Brandner, M. Frank, and D. Rudrich, "DIRPAT—database and viewer of 2D/3D directivity patterns of sound sources and receivers," in *Audio Engineering Society Convention 144*, 2018.

[29] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 351–355.

[30] N. Bertin, E. Camberlein, R. Lebarbenchon, E. Vincent, S. Sivasankaran, I. Illina, and F. Bimbot, "VoiceHome-2, an extended corpus for multichannel speech processing in real homes," *Speech Communication*, vol. 106, pp. 68–78, 2019.

[31] M. Ravanelli, L. Cristoforetti, R. Gretter, M. Pellin, A. Sosi, and M. Omologo, "The DIRHA-English corpus and related tasks for distant-speech recognition in domestic environments," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 275–282.

[32] A. Politis, K. Shimada, P. Sudarsanam, S. Adavanne, D. Krause, Y. Koyama, N. Takahashi, S. Takahashi, Y. Mitsufuji, and T. Virtanen, "STARSS22: A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," in *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2022.

[33] W. He, P. Motlicek, and J.-M. Odobez, "Deep neural networks for multiple speaker detection and localization," in *International Conference on Robotics and Automation (ICRA)*, 2018, pp. 74–79.

[34] ——, "Neural network adaptation and data augmentation for multi-speaker direction-of-arrival estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1303–1317, 2021.

[35] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.