DIMINISHING DOMAIN MISMATCH FOR DNN-BASED ACOUSTIC DISTANCE ESTIMATION VIA STOCHASTIC ROOM REVERBERATION MODELS

Tobias Gburrek, Adrian Meise, Joerg Schmalenstroeer and Reinhold Haeb-Umbach

Paderborn University, Department of Communications Engineering, Germany
{gburrek, schmalen, haeb}@nt.uni-paderborn.de

ABSTRACT

The room impulse response (RIR) encodes, among others, information about the distance of an acoustic source from the sensors. Deep neural networks (DNNs) have been shown to be able to extract that information for acoustic distance estimation. Since there exists only a very limited amount of annotated data, e.g., RIRs with distance information, training a DNN for acoustic distance estimation has to rely on simulated RIRs, resulting in an unavoidable mismatch to RIRs of real rooms. In this contribution, we show that this mismatch can be reduced by a novel combination of geometric and stochastic modeling of RIRs, resulting in a significantly improved distance estimation accuracy.

Index Terms— acoustic distance estimation, room impulse response simulation, stochastic room impulse responses

1. INTRODUCTION

Knowing the distance between a speaker and the recording device can be valuable information for downstream signal processing tasks, e.g., for geometry calibration in wireless acoustic sensor networks (WASNs) [1], signal processing in hearing aids [2] or source extraction [3]. Common approaches estimate the distance between an acoustic source and a compact recording device with multiple microphones by evaluating the power ratio between the coherent signal part, originating from the direct signal propagation path, and the diffuse signal part which summarizes the propagation paths with multiple reflections [4, 5].

In real environments each room has individual acoustic transfer functions, that depend not only on the distance between the recording device and the acoustic source but also on the room's shape, the positions of the device and the source, furniture, and materials on walls, ceiling and floor. Hence, either training data of the room under consideration or at least data from rooms with similar characteristics are required to finetune the parameters of a distance estimator and thus increase the model's precision [6–8].

Collecting and annotating recordings from real environments with diverse room sizes and reverberation conditions is a tedious task. Publicly available data is usually limited in one of the required variabilities: Meeting data often lacks ground truth positioning information, while data intended for comparing localization techniques usually stem only from a very limited number of rooms. As shown in [6], this limited size of these data sets also limits the performance of data-driven distance estimation methods.

Recent approaches to distance estimation are based on deep neural networks (DNNs), be it single-channel [8] or multi-channel [9], and require a large amount of training data to reliably generalize to unknown data. In [9] we proposed to use a convolutional recurrent neural network (CRNN) trained on simulated data, which leveraged the problem by using diverse room setups for generating training data, that were similar to the rooms during tests. To this end, synthetic room impulse responses were generated via an image source method (ISM) [10] and subsequently convolved with speech data. This synthetic data models sources and microphones with omnidirectional characteristics, which directly influences the distance-related features, e.g., coherent-to-diffuse power ratio (CDR), and thus leads to a mismatch between synthetic data and recordings from real environments. A DNN trained with synthetic room impulse responses (RIRs) with omnidirectional sources will have difficulties dealing with real-world data with typically directional sources [1, 8]. To reduce the resulting systematic errors on real-world data, direct-to-reverberant energy ratio (DRR) augmentation techniques can be used [1, 11]. Alternatively, synthetic data may be enriched by recordings from real environments or pre-trained models may be fine-tuned to environments, as for example proposed by the authors of [6]. However, the generalization ability between different data sets typically is limited as shown in [8].

In order to be able to create large synthetic data sets for training a distance estimator, the modeling of the RIRs must become more realistic and, for example, include directional characteristics of sources and microphones to enable the applicability of the models to arbitrary scenarios. For example, there are approaches such as [12] that directly learn to map synthetic RIRs to real RIRs. However, the approach from [12] is not suitable for the problem at hand, since the general structure of a real RIR may be adapted, but the exact parameters of the simulated scenario, e.g., the distance, are not preserved.

In this paper, we propose an approach to RIR simulation with the aim of improving the performance of a data-driven distance estimator, trained with synthetic RIRs, on data from real environments. While the signal propagation paths with only a few reflections are simulated using a geometric approach to model the mainly specular characteristics of the reflection, the signal propagation paths with more reflections, which are mainly diffuse, are simulated based on a stochastic approach. Thereby, the power of the stochastic part of the simulated RIRs is chosen so that the resulting DRR, as distance carrying information, meets the relation between the source-microphone distance and the critical distance, which results from the parameters of the simulated room. Furthermore, the directivity of the sources is taken into account in the geometrical part of the simulated RIR and the calculation of the power of the stochastic part. Experiments have shown that training a DNN-based distance estimator solely on the proposed simulated data improves its generalization ability to real data from the MIRaGe [13] and MIRD [14] data sets.

The paper is organized as follows: In Sec. 2 we briefly review common techniques for simulating RIRs before we present our approach to generating RIRs in Sec. 3. After a short explanation of the used distance estimator in Sec. 4, experimental results are presented and discussed in Sec. 5. Finally, we end with some conclusions in Sec. 6.

2. REVIEW ON RIR SIMULATION TECHNIQUES

Common simulation software for RIRs employs either the image source method, that approximately considers the geometrical setup of microphones and sound sources in a shoe box-shaped room [15–17], ray/cone tracing algorithms utilizing 3D models [18] or a combination of both. Although ray/cone tracing algorithms promise more realistic simulations than the image source method by considering furniture and different wall materials, it remains a tool for special purposes. The generation of diverse and detailed 3D models is time-consuming and the computational complexity of calculating the reflections and tracing the sound geometrically is intractable for the large amount of data required for DNN training.

2.1. Directivity of sources and microphones

Many acoustic sources have a directivity pattern that significantly differs from an omnidirectional directivity pattern. As reported in [19] the directivity pattern of human speakers is frequency-dependent and depends on the type (vocal or fricative) of uttered phonemes. It can be roughly approximated by a cardioid characteristic, which can also be found in monitor loudspeakers.

This implies that depending on the direction of view the acoustic source and each mirrored image of the source get an extra image-dependent weighting factor in the image source method [20]. So the summation of all weighted image signals impinging on the microphone's position is taken as an approximate recording of a directive audio source. If the microphones also have a directivity pattern the impinging mirror signals have to be weighted in accordance to the direction of view of the microphone.

Although the image method has proven its usefulness in many publications, it tends to deliver sparse sequences of impulses that, when convolved with clean audio snippets, do not provide a natural sound perception for the human listener. A RIR recorded in a real environment shows a much noisier and random structure than the RIRs generated by the image method, especially for the late reverberation.

2.2. Stochastic RIR models

Some approaches, e.g., [21], suggest to model RIRs statistically as a random process with an exponentially decaying envelope, that is influenced by some basic acoustic parameters, to better capture general characteristics of a RIR and to ignore the exact geometrical propagation of the sound. We extend the model from [21] with a delay N_d, i.e., the integer rounded time of flight between the acoustic source and the microphone, and approximate the RIR $h_s(n)$ by an exponentially decaying Gaussian process:

$$h_s(n) := \begin{cases} b(n)\, e^{-\Delta \frac{n-N_d}{f_s}} & \text{for } n \geq N_d \\ 0 & \text{else} \end{cases} \tag{1}$$

with n as the sample index, $\Delta{=}3{\cdot}\ln(10)/T_{60}$, f_s as the sampling frequency and $b(n)$ as zero-mean, white Gaussian noise with variance σ_b^2, i.e., $b(n) \sim \mathcal{N}(0, \sigma_b^2)$. However, if the model should reflect distance information and simultaneously consider the directivity patterns of the source and microphone, it has to be further extended.

3. PROPOSED RIR SIMULATION TECHNIQUE

As mentioned in [22] the image source method is suitable to model the early reflections of sound, which are mainly specular, but the late reflections, which are mainly diffuse, are not modeled appropriately by the image source method. Hence, we propose to model only the early reflections using the image source method while the late reflections are modeled using the stochastic approach presented in Section 2.2. Therefore, we combine the image source method and the stochastic model of (1) to generate a RIR $h(n)$ of length N as follows (see Fig. 1)[1]: First, we simulate the early part of the RIR $h_I(n)$ based on the image source method using the image sources up to order K. Thereby, a cardioid pattern is utilized for the source's directivity. We choose $K{=}3$ since reflections of higher order are nearly completely diffuse as reported in [22]. Additionally, a high pass filter is applied to the early part of the RIR $h_I(n)$ as proposed in [10].

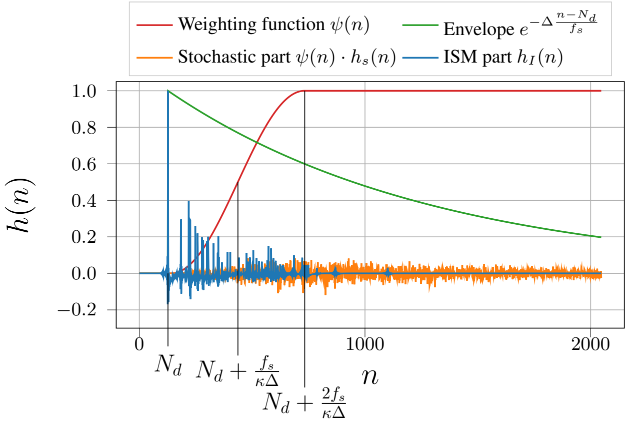

Fig. 1. Visualization of the proposed approach to RIR simulation

The diffuse, late reflections should follow (1) while preserving the distinct reflections modeled by $h_I(n)$. This is achieved by weighting $h_s(n)$ with the function

$$\psi(n) = \begin{cases} 0 & n \leq N_d \\ \frac{1}{2}\left(1 - \cos\left(\frac{\pi(n-N_d)}{2f_s/(\kappa\Delta)}\right)\right) & \text{for} \quad N_d < n \leq N_d + \frac{2f_s}{\kappa\Delta} \\ 1 & n > N_d + \frac{2f_s}{\kappa\Delta} \end{cases}. \tag{2}$$

First experiments on the choice of κ have shown that a better generalization to recorded RIRs is achieved for $\kappa{=}1$. This choice results in a smoother fade-in of the stochastic part of the RIR, i.e., less disturbance of the early reflections. The resulting RIR $h(n)$ is given by

$$h(n) = h_I(n) + \psi(n) \cdot h_s(n). \tag{3}$$

Finally, the power of the Gaussian process in (1) is chosen so that the RIR $h(n)$ exhibits a desired DRR. The desired DRR η is calculated based on the relation between geometrical as well as acoustic properties of the room and the distance between the source and the microphone [2]. In addition to that, we extend the relation from [2] by taking into account that the direct path component of the RIR is scaled by the source's directional response $D(\varphi, \varrho)$, where φ and ϱ are the azimuth and elevation angles between look direction of the source and relative position of the microphone, respectively. Thus, the desired DRR is given by

$$\eta = D^2(\varphi, \varrho) \cdot \frac{d_c^2}{d^2}, \tag{4}$$

[1] Code is available at https://github.com/fgnt/paderwasn

with d as the distance between the source and microphone. d_c denotes the critical distance with

$$d_c = 0.1\,\mathrm{m} \cdot \sqrt{\alpha \cdot \beta} \cdot \sqrt{\frac{V_R/\mathrm{m}^3}{\pi\, T_{60}/\mathrm{s}}}, \tag{5}$$

where α and β are the directivity factors of the acoustic source and the microphone, respectively, and V_R is the room volume. We consider omnidirectional microphones, i.e., $\beta = 1$. The directivity factor of the source α is drawn from the uniform distribution $\mathcal{U}(2.5, 5.5)$, which corresponds to the interval around the directivity factor of the cardioid pattern, in order to account for fluctuations of the DRR of recorded RIRs for different positions.

Given the desired DRR η, the variance σ_b^2 of the Gaussian process $b(n)$ is chosen such that the DRR of the RIR $h(n)$, i.e.,

$$\widehat{\eta} = \frac{\sum\limits_{n=N_d-w}^{N_d+w} h^2(n)}{\sum\limits_{n=N_d+w+1}^{N} h^2(n)}, \tag{6}$$

matches the DRR η, with w defining the length of a small window around the impulse at delay N_d, which belongs to the direct path. Here, we use w=40 as proposed in [11].

4. DISTANCE ESTIMATOR

We use our CRNN from [9] with the short-time Fourier transform (STFT) of the signals of a microphone pair as input for distance estimation. The STFT is represented in the form of its absolute value and the sine and cosine of its phase for each microphone signal. Note that the STFT as input feature comes with the advantage that it does not only contain information about the source microphone distance in the form of the DRR-related inter-level differencess (ILDs), which can be derived from it, but also useful side information for distance estimation, as discussed in [9]. Before calculating the STFT, all signals are normalized to the range $[-1, 1]$. The model is trained to solve distance estimation as a classification problem with a class granularity of $0.1\,\mathrm{m}$.

Since only simulated data should be involved in the training procedure, also the best-performing checkpoint can only be determined based on an independent validation data set of simulated RIRs. However, the best-performing checkpoint for simulated data might not correspond to the best-performing checkpoint for real-world data. We solve this issue via stochastic weight averaging (SWA) [23]. Thereby, the model weights of the last 25 % of the checkpoints are averaged. As mentioned in [24], this might also lead to flatter minima of the error plane, which can lead to a better generalization to other domains, e.g., a better generalization from simulated training data to real-world data.

5. EXPERIMENTS

We simulated a data set of 100 k RIRs to train the distance estimator. Thereby, 10 k different rooms are simulated. The length and width of the rooms are drawn from $\mathcal{U}\,(5\,\mathrm{m}, 7\,\mathrm{m})$ and their ceiling height from $\mathcal{U}\,(2.4\,\mathrm{m}, 3.0\,\mathrm{m})$. Moreover, the sound decay times T_{60} of the rooms were drawn from $\mathcal{U}\,(0.2\,\mathrm{s}, 0.7\,\mathrm{s})$

Ten constellations consisting of a source and a microphone pair with 8 cm inter-microphone distance are generated for each room. Therefore, the microphone pair was placed in the room with random position and orientation. Next, the source is placed relative to the microphone pair with a randomly drawn direction-of-arrival (DoA) and distance so that a minimum distance of $0.3\,\mathrm{m}$ and a maximum distance of $5\,\mathrm{m}$ (or the largest possible distance that would fit into the area considered for source placement) was maintained. Hereby, a minimal distance of $0.5\,\mathrm{m}$ to the walls and $1\,\mathrm{m}$ to the ceiling and floor is kept for each microphone and acoustic source. If the acoustic source would have to be placed outside the considered area for the drawn distance and DoA, the DoA is increased until the source position is within the considered area. The azimuth of the source's direction of view is randomly drawn from $\mathcal{U}\,(-90^\circ, 90^\circ)$ relative to the direct line of sight between the source and the microphone pair while the corresponding elevation is randomly drawn from $\mathcal{U}\,(-15^\circ, 15^\circ)$. All simulated RIRs have a length of N=$16\,384\,\mathrm{samples}$. The image source method was realized using pyroomacoustics (PRA) [16].

In order to evaluate the ability of a distance estimator, which is trained solely using simulated RIRs, to generalize to real-world data, we utilize two data sets of recorded RIRs, namely MIRaGe [13] and MIRD [14] as test sets. Both RIR data sets were recorded in a room of size $6\,\mathrm{m} \times 6\,\mathrm{m} \times 2.4\,\mathrm{m}$ with configurable reverberation times. From the data sets, we selected only those microphone pairs that have a spacing of 8 cm. The acoustic sources of MIRD are placed on a regular grid at either $1\,\mathrm{m}$ or $2\,\mathrm{m}$ distance from a single microphone array with DoAs between $\pm 90^\circ$ with 15° steps in between. Here, we used all examples with a sound decay time T_{60} of 360 ms and 610 ms, which results in a total of 364 test samples. In contrast, the MIRaGe data set has a cube-shaped volume, the so-called grid (46cm×36cm×32cm), in which the sound source is positioned and from which we have selected 100 positions. The microphone arrays are placed at defined distances ($1\,\mathrm{m}$, $2\,\mathrm{m}$, $3\,\mathrm{m}$), three facing the acoustic source and three at an angle of 45°. Additionally, 25 outside of the grid (OOG) source positions are distributed across the room. From the available sound decay times T_{60} we selected 300 ms and 600 ms, which resulted in 1200 test samples for source positions from the grid and 300 test samples for source positions outside of the grid.

Microphone signals with a length of 1 s are generated by convolving clean speech from the LibriSpeech data set [25] with the RIRs. During training the speech samples are randomly drawn from the train-clean-100 subset of LibriSpeech. For the evaluation, ten speech samples from the test-clean subset of LibriSpeech were used per constellation of source and microphone pair to mitigate the influence of the speech on the distance estimates. Moreover, additive white Gaussian noise (AWGN) is added to the microphone signals in order to simulate sensor noise with a signal-to-noise ratio (SNR), which is randomly drawn from $\mathcal{U}\,(40\,\mathrm{dB}, 60\,\mathrm{dB})$.

The distance estimators were trained for 500 k iterations utilizing the Adam optimizer [26] with a batch size of 16 and a learning rate of $3{\cdot}10^{-4}$. Thereby, a checkpoint is created every 10 k iterations. The STFT for feature extraction uses a Blackman window with a length of 25 ms and shift of 10 ms.

We evaluate the performance of the distance estimators by calculating the mean-absolute error (MAE) of the I distance estimates per data set with

$$\mathrm{MAE} = \frac{1}{I}\sum_{i=1}^{I} \left|d_i - \hat{d}_i\right|, \tag{7}$$

where d_i denotes the ground truth distance and $\hat{d}_i$ its estimate. Note that estimated distance classes are mapped to the distance estimate $\hat{d}_i$ before calculating the MAE.

Table 1. Comparison of the proposed approach to RIR simulation and the image source method (ISM) with different source directivity patterns. Additionally to the results on MIRD and MIRaGe results on a simulated version of MIRaGe (Sim.) are reported. We use the same approach to RIR simulation for the simulated version of MIRaGe and the training of the corresponding distance estimator.

Method	**Source Directivity**	**MAE / m**		
		MIRD	**MIRaGe**	**Sim.**
ISM	Omnirectional	0.75	0.54	0.20
ISM	Subcardioid	0.51	0.47	0.17
ISM	Cardioid	0.27	0.46	0.16
ISM	Supercardiod	0.49	0.61	0.21
ISM	Hypercardioid	0.32	0.54	0.21
Proposed	Cardioid	0.20	0.31	0.26

Table 1 compares the performance of a distance estimator trained with RIRs which are simulated using the proposed method to the the performance of distance estimators trained with RIRs which are simulated via the the image source method using different directivity patterns for the source. It can be seen that the proposed RIR simulation method leads to a significantly better distance estimation performance on MIRD and MIRaGe compared to the image source method (ISM). Moreover, it can be seen that the model which is trained with RIRs from the image source method with cardioid source directivity exhibits the best performance of all models whose training data were generated using the image source method. In contrast, the distance estimators which are trained with RIRs generated with less pronounced source directivities perform worst.

In addition to that, the distance estimators are evaluated on a simulated version of MIRaGe, which was generated by the same RIR simulator as the one used to generate the training data for the respective model. While there is a large gap between the performance on simulated and recorded RIRs for distance estimators whose training data was simulated using the image source method, this gap is relatively small for a distance estimator trained with data for the proposed method. This means that the proposed method improves the generalization ability of a distance estimator from simulated training data to real data by far.

Results for the distance estimation performance, which can be achieved by a combination of RIRs from the image source method and the DRR augmentation technique proposed in [1], can be found in Table 2. Thereby, the DRR augmentation method varies the DRR of the RIRs by scaling the part of the RIRs belonging to the direct path propagation. Compared to the random scaling of the direct path component, which we proposed in [1] to increase the DRR, better distance estimates can be achieved by scaling the direct path so that the RIRs show a DRR which is calculated based on (4) as in the proposed method. Hereby, the influence of the directivity on the direct path $D(\varphi, \varrho)$ in (4) is calculated for the cardioid pattern. Further, the distance estimation performance is better when simulating sources with a cardioid directivity instead of omnidirectional sources. From this we hypothesize that the distance estimator benefits from incorporating the source's directivity into the model of the early specular reflections. However, the performance which can be achieved by using the proposed RIR simulator cannot be reached.

Table 2. Comparison of the proposed approach to RIR simulation and the DRR augmentation technique from [1], which scales the impulse belonging to the direct path with the factor α. α is either randomly drawn as in [1] or calculated so that the resulting RIR meets the target DRR from (4).

Method	**Source Directivity**	**DRR aug.**	**MAE / m**	
			MIRD	**MIRaGe**
ISM	Omnidirectional	$\alpha \sim \mathcal{U}(1, 3)$	0.27	0.56
ISM	Omnidirectional	α based on (4)	0.23	0.43
ISM	Cardioid	α based on (4)	0.24	0.37
Proposed	Cardioid	-	0.20	0.31

Table 3. Ablation study of the proposed approach to RIR simulation by varying the source's directivity pattern, the maximum order of the image sources K used to simulate $h_I(n)$ and the method used calculate the late part of the RIRs. The last line corresponds to the proposed parametrization.

Source Directivity	**Order K**	**Late RIR**	**MAE / m**	
			MIRD	**MIRaGe**
Cardioid	0	Stochastic	0.52	0.54
Omnidirectional	3	Stochastic	0.36	0.36
Cardioid	3	ISM	0.24	0.42
Cardioid	3	Stochastic	0.20	0.31

An ablation study for the proposed RIR simulator is given in Table 3. It is shown that the distance estimation performance degrades a lot if only the direct path is simulated via the image source method, i.e., K=0, which again speaks for the importance of a correct simulation of the specular early reflections. Moreover, it can be seen that the stochastic process from (5) models the diffuse reflections of higher order better than the image source method. In addition to that, simulating omnidirectional sources and also choosing $D(\varphi, \varrho)$=1 in (4) degrades the performance of the distance estimator.

6. SUMMARY

In this paper, we presented a new approach to simulate RIRs for the training of a DNN-based acoustic distance estimator to improve its performance in real-world scenarios. Thereby, the image source method was utilized to simulate the reflections of lower order, which mainly show a specular character. A cardioid pattern is used to simulate the source's directivity in the image source method because in real-world scenarios acoustic sources typically exhibit a directivity pattern, which largely differs from an omnidirectional directivity pattern. In addition to that, the mainly diffuse reflections of higher order are modeled via an exponentially decaying stochastic process. The power of the latter is scaled such that the DRR of the RIR fits to the distance between the source and the microphone. Experiments on recorded RIRs show that our contribution improves the simulated training data of a distance estimator to match the characteristics present in real data better than previous approaches.

In future works we will investigate the suitability of the proposed approach to RIR simulation to generate training data for data-driven models for other purposes, like dereverberation, speech enhancement or room parameter estimation, e.g., DRR and sound decay time T_{60} estimation.

7. REFERENCES

[1] Tobias Gburrek, Joerg Schmalenstroeer, and Reinhold Haeb-Umbach, "Geometry calibration in wireless acoustic sensor networks utilizing DoA and distance information," *EURASIP Journal on Audio, Speech, and Music Processing*, 2021.

[2] Mehdi Zohourian and Rainer Martin, "Binaural direct-to-reverberant energy ratio and speaker distance estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 92–104, 2020.

[3] Darius Petermann and Minje Kim, "Hyperbolic distance-based speech separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[4] A. Brendel and W. Kellermann, "Distributed Source Localization in Acoustic Sensor Networks using the Coherent-to-Diffuse Power Ratio," *IEEE Journal of Selected Topics in Signal Processing*, 2019.

[5] Andreas Brendel, Andy Regensky, and Walter Kellermann, "Probabilistic modeling for learning-based distance estimation," in *Proc. International Congress on Acoustics (ICA)*, Aachen, Germany, Sept. 2019.

[6] Saksham Singh Kushwaha, Iran R. Roman, Magdalena Fuentes, and Juan Pablo Bello, "Sound source distance estimation in diverse and dynamic acoustic conditions," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2023.

[7] Michael Neri, Archontis Politis, Daniel Krause, Marco Carli, and Tuomas Virtanen, "Single-channel speaker distance estimation in reverberant environments," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2023.

[8] Michael Neri, Archontis Politis, Daniel Aleksander Krause, Marco Carli, and Tuomas Virtanen, "Speaker distance estimation in enclosures from single-channel audio," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2242–2254, 2024.

[9] Tobias Gburrek, Joerg Schmalenstroeer, and Reinhold Haeb-Umbach, "On source-microphone distance estimation using convolutional recurrent neural networks," in *Proc. 14th ITG-Symposium Speech Communication*, 2021.

[10] Jont Allen and David Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, pp. 943–950, Apr. 1979.

[11] Nicholas J. Bryan, "Impulse response data augmentation and deep neural networks for blind room acoustic parameter estimation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[12] Anton Ratnarajah, Zhenyu Tang, and Dinesh Manocha, "TS-RIR: Translated synthetic room impulse responses for speech augmentation," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021.

[13] Jaroslav Čmejla, Tomáš Kounovský, Sharon Gannot, Zbyněk Koldovský, and Pinchas Tandeitnik, "MIRaGe: Multichannel database of room impulse responses measured on high-resolution cube-shaped grid," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2021.

[14] Elior Hadad, Florian Heese, Peter Vary, and Sharon Gannot, "Multichannel audio database in various acoustic environments," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2014.

[15] Emanuel AP Habets, "Room impulse response generator," *Eindhoven University of Technology, Technical Report*, vol. 2, no. 2.4, 2006.

[16] Robin Scheibler, Eric Bezzam, and Ivan Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[17] Zeyu Xu, Adrian Herzog, Alexander Lodermeyer, Emanuël A. P. Habets, and Albert G. Prinn, "Simulating room transfer functions between transducers mounted on audio devices using a modified image source method," *The Journal of the Acoustical Society of America*, vol. 155, no. 1, pp. 343–357, Jan. 2024.

[18] Rene Glitza, Luca Becker, Alexandru Nelus, and Rainer Martin, "Database of simulated room impulse responses for acoustic sensor networks deployed in complex multi-source acoustic environments," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2023.

[19] Christoph Pörschmann and Johannes Arend, "Analyzing the directivity patterns of human speakers," in *DAGA - Jahrestagung für Akustik*, Mar. 2020.

[20] Sina Hafezi, Alastair H. Moore, and Patrick A. Naylor, "Modelling source directivity in room impulse response simulation for spherical microphone arrays," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2015.

[21] K Lebart, Jean-Marc Boucher, and P Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica united with Acustica*, vol. 87, pp. 359–366, May 2001.

[22] Dirk Schröder, *Physically based real-time auralization of interactive virtual environments*, Ph.D. thesis, RWTH Aachen, Berlin, 2011.

[23] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson, "Averaging weights leads to wider optima and better generalization," in *Proc. Conf. on Uncertainty in Artificial Intelligence (UAI)*, 2018.

[24] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park, "SWAD: Domain generalization by seeking flat minima," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[25] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[26] Diederik Kingma and Jimmy Ba, "Adam: A Method for Stochastic Optimization," in *Proc. International Conference on Learning Representations (ICLR)*, Banff, Canada, Apr. 2014.