Improving multichannel speech enhancement through accurate room-acoustic simulations

Georg Götz [1,**], *Alessia Milo* [1], *Steinar Guðjónsson* [1], *Daniel Gert Nielsen* [1], *Jesper Pedersen* [1], *Finnur Pind* [1]

[1] Treble Technologies, Reykjavík, Iceland

georg.goetz@treble.tech, am@treble.tech, sg@treble.tech, dgn@treble.tech, jp@treble.tech, fp@treble.tech

Abstract

Room-acoustic simulations are widely used to augment training data for deep-learning-based speech enhancement. While most pipelines rely on simplified geometrical acoustics, wave-based approaches offer greater physical accuracy. In this work, we examine how simulation fidelity affects multichannel speech enhancement performance. To this end, we train SpatialNet on datasets augmented with different room-acoustic simulation methods and evaluate the resulting models on measured data. We compare lower-fidelity datasets based on geometrical acoustics with a high-fidelity dataset using advanced acoustic modelling and a hybrid combination of wave-based and geometrical acoustics simulations. Training on the high-fidelity dataset results in an up to 38 % relative reduction in median word error rate compared to the lower-fidelity alternatives. These results show that augmentation with high-fidelity room-acoustic simulations directly translates into improved multichannel speech enhancement performance.

Index Terms: Multichannel speech enhancement, far-field speech recognition, room acoustics, geometrical acoustics, wave-based simulation, data augmentation

1. Introduction

As smart devices become deeply embedded in everyday life, natural language user interfaces (NLUIs) have emerged as a key paradigm for enabling seamless and intuitive human-device interaction. However, this interaction frequently occurs in acoustically complex environments. Across homes, public spaces, and in-vehicle settings, the intelligibility of user instructions can be significantly degraded by background noise, reverberation, or overlapping speech. To maintain reliable performance under such conditions, automatic speech recognition (ASR) systems rely on speech enhancement techniques like denoising and dereverberation to isolate and preserve linguistic information.

Recent advances in speech enhancement have been largely driven by deep-learning-based methods. Early neural architectures relied on convolutional [1] and recurrent networks [2] to learn time-frequency masks or directly estimate clean speech spectra [3]. More recent models incorporate attention mechanisms and Transformer-based architectures to capture long-range temporal dependencies and complex acoustic patterns. For instance, deep filtering approaches like DeepFilterNet [4] combine deep neural networks with classical signal processing principles for real-time enhancement. In contrast, hybrid architectures like SpatialNet integrate convolutional layers with multi-head self-attention mechanisms and leverage both local spectral structure and global contextual modelling [5]. These data-driven techniques have substantially improved robustness in challenging acoustic conditions and now represent an important paradigm in contemporary speech enhancement research.

Going beyond the single channel paradigm, multichannel speech enhancement exploits additional information captured by microphone arrays. This additional information enables systems to leverage spatial diversity between target speech and interfering sources to improve robustness in noisy environments [6, 7]. By jointly exploiting spatial and spectral characteristics of the recorded signals, multichannel approaches can achieve improved separation and enhancement performance compared to single-channel methods [6, 7].

Deep learning models are trained on increasingly large and heterogeneous datasets, such as those introduced in the Deep Noise Suppression (DNS) challenges [8]. In practice, data augmentation is typically performed by convolving clean speech with room impulse responses (RIRs) to simulate different acoustic environments. However, much of the state-of-the-art still relies on relatively simplistic synthetic RIRs, which are often derived from shoebox geometries rendered with the image-source method (ISM) using a single frequency-independent absorption coefficient [9, 10, 11]. However, such simple RIRs may not sufficiently capture the acoustic complexity of real environments.

Prior evidence in related tasks suggests that increasing RIR realism can improve downstream performance [12, 13, 14, 15, 16]. Most previous studies focused on the effect of source and receiver realism [13, 14, 15] and wall material realism [12, 13, 14, 15]. Additionally, Bezzam et al. investigate the performance difference between ISM- and ray-tracing-based training datasets [12], while Tang et al. also include single-channel numerical simulations in their study [16]. However, none of the referenced prior studies have explored the effect of simulation fidelity for multichannel speech enhancement with rigid arrays.

This study aims to close this gap by examining how room-acoustic simulation fidelity affects multichannel speech enhancement performance in real-world conditions. We train SpatialNet [5] with several training datasets of varying fidelity and evaluate the resulting models on measured data. The remainder of this paper is structured as follows. Section 2 summarizes different room-acoustic simulation paradigms. Section 3 describes the experimental setup of this study. Section 4 presents the results and Section 5 concludes the paper.

** indicates the corresponding author.

2. Simulation paradigms for acoustic modelling

In the context of data augmentation for speech enhancement, differences in simulation methods/fidelity may lead to different spatial and spectral characteristics in the generated RIRs. Room-acoustic simulation methods are commonly divided into geometrical acoustics (GA) and wave-based approaches, with hybrid methods combining elements of both. GA techniques, including the ISM and ray-based methods such as ray tracing, approximate sound propagation using specular reflections and energy transport assumptions that become increasingly valid at higher frequencies [17, 18, 19, 20, 21]. However, GA does not inherently model wave phenomena such as room modes, or diffraction around edges and obstacles, and diffraction in particular remains difficult to represent accurately within GA-based frameworks [22, 23]. These limitations are also reflected in broader discussions of simulation uncertainty and cross-tool variability [24, 25].

In contrast, wave-based methods (e.g., finite-difference time-domain, finite-/spectral-element, and discontinuous Galerkin formulations) solve the acoustic wave equation directly, enabling the representation of diffraction, modal behaviour, and complex, frequency-dependent boundary conditions. These effects are particularly relevant at low and mid frequencies where wavelengths are comparable to room dimensions [26, 27, 28, 29]. While traditionally associated with higher computational cost, recent advances in time-domain formulations and high-performance implementations have made large-scale wave-based or partially wave-based simulations more feasible [30, 31, 32].

Hybrid approaches combine these paradigms by applying a wave solver below a crossover frequency and GA above it, thereby capturing low-/mid-frequency wave effects while relying on GA assumptions at higher frequencies. The trade-offs between "rays or waves" have been discussed extensively in the room-acoustics literature [33]. In the context of data augmentation for multichannel speech enhancement, hybrid simulation is intended to reduce discrepancies related to low-frequency modal behaviour, diffraction, and frequency-dependent decay, while maintaining broadband coverage.

3. Experiment setup

3.1. Network configuration

We train SpatialNet [5] to investigate the relationship between room-acoustic simulation accuracy and neural network downstream performance. Inspired by the Conformer architecture [34], SpatialNet integrates convolutional modelling and multi-head self-attention [35] for end-to-end multichannel speech enhancement in the STFT domain. It combines narrow-band blocks for speaker clustering and temporal filtering with cross-band blocks for learning correlations across frequencies [5]. Our experiment uses the SpatialNet-small configuration at 16 kHz sampling rate.

SpatialNet is microphone-array-dependent, i.e., speech enhancement on different arrays requires retraining the network for each array geometry [5]. In this work, we focus on a six-channel subset of the em32 Eigenmike array. More precisely, we choose the channels 1, 19, 11, 27, 21, and 9, as defined in the manufacturer's data sheet[1], corresponding approximately to locations in the front, back, right, left, top, and bottom of the array, respectively.

Although the Eigenmike is not a typical array for speech enhancement, we specifically chose it for two reasons. First, we want to investigate whether accurate receiver modelling improves speech enhancement compared to simplified open array simulations. Speech enhancement is commonly used in smart speakers consisting of several microphones mounted on a rigid scatterer, and the Eigenmike is a reasonable approximation of this setup. Second, we want to investigate the simulation-to-real performance achievable when training datasets are generated with modern simulation tools. Practical speech enhancement algorithms must operate well under real-world conditions, and evaluation datasets should be designed to assess such scenarios. Several datasets containing Eigenmike measurements are publicly available, thus making the array suitable for reproducible studies of real-world performance.

3.2. Training datasets

We train SpatialNet with several datasets augmented using different room-acoustic simulation methods. To this end, we compare lower-fidelity datasets based on GA with a high-fidelity dataset employing hybrid simulations. For the GA-based datasets, we further analyse the effect of dataset design by contrasting an uninformed dataset, where target reverberation times and room dimensions are randomly sampled from predefined ranges, with an informed dataset, where these parameters are matched to realistic material properties and room configurations drawn from curated libraries. Figure 1a shows the resulting distributions of the reverberation time T_{20} for all training datasets.

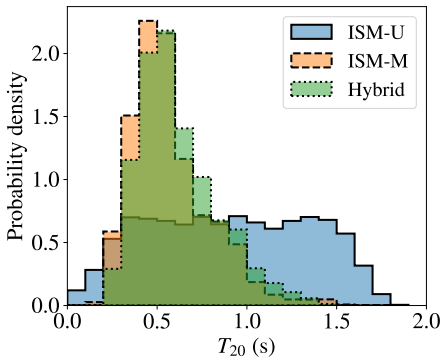
3.2.1. ISM: Image-source dataset

The original SpatialNet paper [5] augments the training data using RIRs simulated with the open-source toolbox gpuRIR [36]. This toolbox implements GPU-accelerated image-source simulations, and combines them with diffuse late reverberation tails [36]. Following the simulation setup of the original SpatialNet paper, we transition between ISM and diffuse reverberation after the first 15 dB of energy decay. This ensures a faithful simulation of early reflections, while efficiently modelling the high reflection density characteristic of late reverberation.

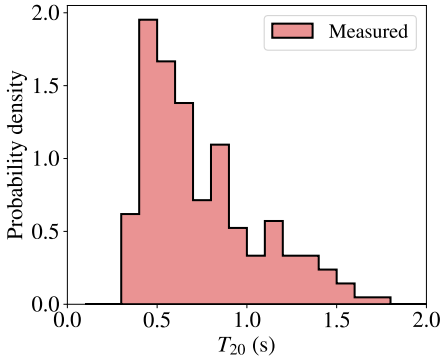
For this study, we simulate two ISM-based datasets. The first dataset, ISM-U, randomly samples target reverberation times and room dimensions from predefined intervals. These intervals are chosen to match the minimum and maximum values of the hybrid dataset (see Sec. 3.2.2) to ensure comparability: $x \in [3\,\mathrm{m}, 33\,\mathrm{m}]$, $y \in [3\,\mathrm{m}, 26\,\mathrm{m}]$, $z \in [2.5\,, 4.7\,\mathrm{m}]$, $T_{20} \in [0.2\,\mathrm{s}, 1.6\,\mathrm{s}]$, where x, y, and z are the respective room dimensions. Randomly sampling reverberation times and room dimensions is common practice and widely used in state-of-the-art speech enhancement studies. This setup represents an uninformed simulation scenario in which no prior knowledge about typical materials or room volumes is incorporated.

The second dataset, ISM-M, incorporates additional information to obtain more realistic material and room dimension choices. Specifically, we match the setup of the hybrid dataset as closely as possible. To this end, we compute bounding boxes for each room in the hybrid dataset, and use the resulting dimensions in gpuRIR. We also replicate the exact source and receiver positions from the hybrid dataset and set the T_{20} values

[1] https://eigenmike.com/sites/default/files/documentation-2023-10/EigenStudio%20User%20Manual%20R02D.pdf



(a) *Training datasets* ($N_{\text{Scenes}} = 4801$)



(b) *Evaluation dataset* ($N_{\text{Scenes}} = 60$)

Figure 1: *Distribution of the reverberation time T_{20} for the training and evaluation datasets used in this study, averaged over octave bands from 63 Hz to 4 kHz.*

from the hybrid simulation as the target reverberation times for the gpuRIR simulations. This way, the ISM-M dataset closely matches the hybrid dataset, with the only remaining differences being the presence of scattering objects in the rooms and the simulation and source–receiver modelling paradigm.

For both ISM datasets, the Eigenmike array is modelled as an open microphone array. This approach is common practice when working with gpuRIR.

3.2.2. Hybrid: High-fidelity dataset

The hybrid dataset was generated entirely using the Treble SDK simulation tool, with all geometry models and boundary-condition materials taken from the included libraries. The dataset consists of three subsets representing different room types and size ranges. The first subset contains 133 living-room geometries with volumes ranging from 40 to 180 m^3. The second subset contains 103 classroom geometries with volumes ranging from 90 to 400 m^3. The third subset contains 88 restaurant geometries with volumes ranging from 300 to 1600 m^3. All geometries are furnished and vary in shape and complexity.

The materials applied to the surfaces in the room models are carefully curated to represent realistic scenarios. All materials are frequency-dependent and characterised by complex surface impedances. Each surface is assigned an appropriate material. For example, windows use glass materials, furniture uses wood or plastic materials, and walls and ceilings use gypsum or concrete materials.

Each room contains four sources placed randomly within the space. Most sources are directive and represent either speech sources or loudspeakers, with their orientations randomised. In larger rooms, one speech source is replaced by an omnidirectional source to increase variability. The number of receivers ranges from 20 to 30 per room, distributed randomly and positioned at least 1 m from any source and 0.5 m from any surface. During training, SpatialNet uses scenes of up to 3 overlapping speakers from the same environment. We randomly chose them from the dataset while ensuring that they are all in the same room, resulting in 4801 scenes in total. The distribution of reverberation times is shown in Figure 1a.

The crossover frequency, i.e., the upper frequency limit of the wave solver, is set between 1 and 2 kHz depending on the room size. The remaining part of the audible spectrum, from the crossover frequency up to 12 kHz, is simulated using the GA solver of the Treble SDK, which combines the ISM with a ray-radiosity approach. For this dataset, a maximum ISM order of three is used.

Finally, the effect of the Eigenmike array is incorporated by simulating a full-wave free-field device-related transfer function (DRTF) up to 12 kHz with the Treble SDK. Using the 16th-order Ambisonics RIRs from the hybrid simulations, the Eigenmike DRTF can be rendered in post-processing. This approach accounts for the full scattering effects of the Eigenmike sphere geometry. From the simulated Eigenmike response, we extract six channels evenly distributed around the sphere and use them for training, as outlined in Section 3.1.

3.3. Evaluation dataset

We evaluate all trained models on a dataset of Eigenmike measurements to avoid bias towards either simulation paradigm and to assess performance under real-world conditions. To this end, we introduce LibriCSS-EM6, a dataset inspired by the structure of LibriCSS [37], but based on the six-channel Eigenmike subset described in Section 3.1. Analogous to LibriCSS, the proposed EM6 variant comprises approximately 5000 utterances across six overlap conditions: 0S, 0L, OV10, OV20, OV30, and OV40. The conditions 0S and 0L denote 0 % speaker overlap with short and long inter-speaker silences, respectively, whereas OV10–OV40 correspond to 10 % to 40 % speaker overlap. Reverberant speech is generated by convolving dry LibriSpeech utterances from the clean test set [38] with Eigenmike room impulse responses (RIRs). The LibriCSS metadata provide the necessary time stamps and utterance identifiers to reconstruct the continuous streams, and segmentation follows the original LibriCSS procedure.

The measured Eigenmike RIRs were randomly drawn from two publicly available datasets: Motus [39] and Arni6DoF [40]. The Motus dataset comprises Eigenmike RIRs recorded in a room with various furniture configurations, making it suitable for assessing model performance under different reverberation scenarios and with varying scattering objects. Arni6DoF, recorded in a variable-acoustics room, further extends the evaluation set with additional reverberation conditions. Figure 1b illustrates the resulting band-averaged T_{20} distribution for LibriCSS-EM6, demonstrating broad coverage of reverberation times representative of medium-sized rooms.

Following the LibriCSS structure, the proposed EM6 subset comprises 60 sessions (10 sessions $\times$ 6 overlap conditions). Each session uses randomly drawn RIRs from either the Motus or Arni6DoF dataset, with equal contributions from both and

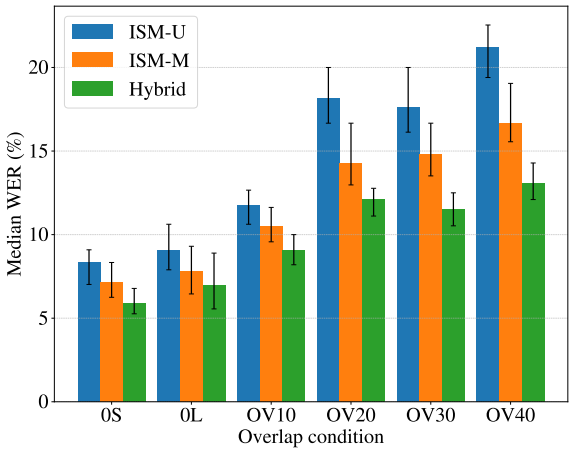


Figure 2: *Median word error rate (WER) and bootstrapped 95 % confidence intervals for speech enhanced with different SpatialNet models across several speaker overlap conditions. The models were trained on datasets augmented with room-acoustic simulations of varying fidelity.*

a balanced assignment across overlap conditions. Arni6DoF comprises five room configurations, each repeated across six sessions, while Motus provides enough configurations to assign a distinct room to every session. Prior to segmentation, we add multichannel diffuse noise to the speech mixture at a random SNR between 0 to 20 dB. The noise signals are generated from REVERB Challenge recordings [41] using the method proposed in [42].

3.4. Evaluation setup

We train one SpatialNet model per training dataset. In all cases, training was running for 30 epochs, after which the validation loss no longer improved substantially. Following the evaluation procedure proposed in the SpatialNet paper [5], we average the network weights over the last ten epochs to stabilise ASR performance.

We evaluate the downstream performance of the different SpatialNet variants by transcribing the enhanced speech signals with a Kaldi [43] speech recognition pipeline implemented in pyKaldi2 [44], adopting the same evaluation pipeline that was proposed for the original LibriCSS dataset [37]. More specifically, acoustic scores are generated by a 3-layer bidirectional long short term memory (BLSTM) [45] acoustic model (512 units per direction) trained on LibriSpeech with cross-entropy and maximum mutual information (MMI) [46] sequence training, and decoded with Kaldi using the standard LibriSpeech 4-gram language model.

4. Results

We enhance and transcribe all utterances of LibriCSS-EM6 with each of the trained SpatialNet models and calculate the corresponding word error rates (WERs). Figure 2 shows the median WER of all models under the investigated overlap conditions, with bootstrapped 95 % confidence intervals. Across all overlap conditions, the model trained with hybrid room-acoustic simulations performs best, whereas the models trained with image-source-based augmentation perform worse. Among the image-source datasets, the variant using informed materials and room dimensions performs considerably better than the one where both parameters were randomly drawn from typical ranges.

Table 1: *Absolute and relative median WER improvement achieved by the Hybrid training dataset compared to the ISM datasets, with 95 % bootstrapped confidence intervals. Positive numbers indicate that the Hybrid dataset achieves a better WER than the reference dataset.*

Reference system	Absolute median WER improvement	Relative median WER improvement (%)
Overlap condition: 0S		
ISM-U	2.17 [1.26, 3.03]	26.7 [17.2, 34.2]
ISM-M	1.19 [0.46, 2.07]	16.7 [6.7, 25.5]
Overlap condition: 0L		
ISM-U	1.91 [0.66, 3.21]	21.4 [7.7, 34.6]
ISM-M	0.72 [−0.55, 1.88]	8.7 [−7.7, 23.5]
Overlap condition: OV10		
ISM-U	2.78 [1.67, 3.87]	23.5 [15.0, 31.4]
ISM-M	1.50 [0.60, 2.48]	14.2 [5.8, 22.3]
Overlap condition: OV20		
ISM-U	6.15 [4.57, 7.50]	33.3 [27.3, 39.0]
ISM-M	2.21 [1.07, 4.17]	15.6 [8.2, 25.0]
Overlap condition: OV30		
ISM-U	6.14 [4.73, 7.64]	34.6 [28.1, 40.3]
ISM-M	3.20 [2.25, 4.67]	22.2 [15.6, 29.0]
Overlap condition: OV40		
ISM-U	8.18 [6.67, 9.56]	38.3 [33.3, 43.2]
ISM-M	4.00 [2.77, 5.70]	23.5 [17.2, 30.0]
Overall		
ISM-U	4.29 [3.33, 4.53]	30.0 [25.0, 31.7]
ISM-M	1.93 [1.43, 2.50]	16.3 [12.9, 20.0]

While Figure 2 illustrates the overall WER trends across overlap conditions, Table 1 provides a direct comparison of the models. The table reports the absolute and relative median WER improvements of the model trained with high-fidelity data (Hybrid) over the remaining models. The corresponding bootstrapped 95 % confidence intervals are computed from paired utterance-level differences, accounting for the shared variability across test utterances. The positive values indicate that the model trained with hybrid room-acoustic simulations consistently outperforms the models trained using image-source-based augmentations. All but one confidence interval lie entirely above zero, indicating that the corresponding improvements are statistically significant at the 95 % confidence level.

5. Conclusion

This paper investigated different room-acoustic simulation paradigms for augmenting training data in multichannel speech enhancement and analysed the impact of simulation fidelity on downstream performance. We observe a relative reduction in median word error rate of up to 38 % when the training dataset is augmented using high-fidelity room-acoustic simulations rather than simplified approaches. These findings demonstrate that

performance gains can be achieved without altering the network or training strategy. Increasing the physical accuracy of the training data alone is sufficient.

6. Generative AI use disclosure

The authors acknowledge the use of ChatGPT 5.2 (accessed in February 2026) to rephrase some sentences for clarity, polish the manuscript's language, and help to arrange and format tables and figures. All AI-assisted content was reviewed and revised by the authors to ensure accuracy and clarity of meaning.

7. References

[1] S. R. Park and J. W. Lee, "A fully convolutional neural network for speech enhancement," in *Proc. Interspeech*, Stockholm, Sweden, 2017, pp. 1993–1997.

[2] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *Proc. 12th Int. Conf. Latent Var. Anal. Signal Separation*, Liberec, Czech Republic, 2015, pp. 91–99.

[3] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, 2018.

[4] H. Schröter, A. N. Escalante-B., T. Rosenkranz, and A. Maier, "DeepFilterNet: A low complexity speech enhancement framework for full-band audio based on deep filtering," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Singapore, Singapore, 2022, pp. 7407–7411.

[5] C. Quan and X. Li, "SpatialNet: Extensively learning spatial information for multichannel joint speech separation, denoising and dereverberation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 32, pp. 1310–1323, 2024.

[6] R. Haeb-Umbach, T. Nakatani, M. Delcroix, C. Boeddeker, and T. Ochiai, "Microphone array signal processing and deep learning for speech enhancement: Combining model-based and data-driven approaches to parameter estimation and filtering," *IEEE Signal Processing Magazine*, vol. 41, no. 6, pp. 12–23, 2024.

[7] X. Zeng, S. Xu, and M. Wang, "A time-frequency fusion model for multi-channel speech enhancement," *EURASIP J. Audio Speech Music Process.*, vol. 2024, no. 1, pp. 47:1–12, 2024.

[8] H. Dubey, A. Aazami, V. Gopal, B. Naderi, S. Braun, R. Cutler, A. Ju, M. Zohourian, M. Tang, M. Golestaneh, and R. Aichner, "ICASSP 2023 Deep Noise Suppression Challenge," *IEEE Open J. Signal Process.*, vol. 5, pp. 725–737, 2024.

[9] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, 2017, pp. 5220–5224.

[10] C. Kim, A. Misra, K. Chin, T. Hughes, A. Narayanan, T. Sainath, and M. Bacchiani, "Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in Google Home," in *Proc. Interspeech*, Stockholm, Sweden, 2017, pp. 379–383.

[11] M. Maciejewski, G. Wichern, E. McQuinn, and J. Le Roux, "WHAMR!: Noisy and reverberant single-channel speech separation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Online conference, 2020, pp. 696–700.

[12] E. Bezzam, R. Scheibler, C. Cadoux, and T. Gisselbrecht, "A study on more realistic room simulation for far-field keyword spotting," in *Proc. Asia-Pacific Signal Information Process. Assoc. Annual Summit Conf. (APSIPA ASC)*, Auckland, New Zealand, 2020, pp. 674–680.

[13] R. Arakawa, M. Parvaix, C. Lai, H. Erdogan, and A. Olwal, "Quantifying the effect of simulator-based data augmentation for speech recognition on augmented reality glasses," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Seoul, Korea, 2024, pp. 726–730.

[14] P. Srivastava, A. Deleforge, A. Politis, and E. Vincent, "How to (virtually) train your speaker localizer," in *Proc. Interspeech*, Dublin, Ireland, 2023, pp. 1204–1208.

[15] E. Gusó, J. Luberadzka, U. Sayin, and X. Serra, "MB-RIRs: a synthetic room impulse response dataset with frequency-dependent absorption coefficients," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, Tahoe City, CA, USA, 2025, pp. 1–5.

[16] Z. Tang, R. Aralikatti, A. J. Ratnarajah, and D. Manocha, "GWA: A large high-quality acoustic dataset for audio processing," in *Proc. ACM SIGGRAPH Conf.*, Vancouver, BC, Canada, article no. 36, 2022.

[17] L. Savioja and U. P. Svensson, "Overview of geometrical room acoustic modeling techniques," *J. Acoust. Soc. Am.*, vol. 138, no. 2, pp. 708–730, 2015.

[18] A. Krokstad, S. Strøm, and S. Sørsdal, "Calculating the acoustical room response by the use of a ray tracing technique," *J. Sound Vib.*, vol. 8, no. 1, pp. 118–125, 1968.

[19] ——, "Fifteen years' experience with computerized ray tracing," *Appl. Acoust.*, vol. 16, no. 4, pp. 291–312, 1983.

[20] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.

[21] J. Borish, "Extension of the image model to arbitrary polyhedra," *J. Acoust. Soc. Am.*, vol. 75, no. 6, pp. 1827–1836, 1984.

[22] R. R. Torres, U. P. Svensson, and M. Kleiner, "Computation of edge diffraction for more accurate room acoustics auralization," *J. Acoust. Soc. Am.*, vol. 109, no. 2, pp. 600–610, 2001.

[23] C. Schissler, R. Mehra, and D. Manocha, "High-order diffraction and diffuse reflections for interactive sound propagation in large environments," *ACM Trans. Graph.*, vol. 33, no. 4, article no. 39, 2014.

[24] M. Vorländer, "Computer simulations in room acoustics: Concepts and uncertainties," *J. Acoust. Soc. Am.*, vol. 133, no. 3, pp. 1203–1213, 2013.

[25] F. Brinkmann, L. Aspöck, D. Ackermann, S. Lepa, M. Vorländer, and S. Weinzierl, "A round robin on room acoustical simulation and auralization," *J. Acoust. Soc. Am.*, vol. 145, no. 4, pp. 2746–2760, 2019.

[26] B. Hamilton, "Finite difference and finite volume methods for wave-based modelling of room acoustics," Ph.D. dissertation, University of Edinburgh, 2016.

[27] F. Pind, A. P. Engsig-Karup, C.-H. Jeong, J. S. Hesthaven, M. S. Mejling, and J. Strømann-Andersen, "Time domain room acoustic simulations using the spectral element method," *J. Acoust. Soc. Am.*, vol. 145, no. 6, pp. 3299–3310, 2019.

[28] A. G. Prinn, "A review of finite element methods for room acoustics," *Acoustics*, vol. 5, no. 2, pp. 367–395, 2023.

[29] D. Botteldooren, "Finite-difference time-domain simulation of low-frequency room acoustic problems," *J. Acoust. Soc. Am.*, vol. 98, no. 6, pp. 3302–3308, 1995.

[30] H. Wang, I. Sihar, R. Pagán Muñoz, and M. Hornikx, "Room acoustics modelling in the time-domain with the nodal discontinuous Galerkin method," *J. Acoust. Soc. Am.*, vol. 145, no. 4, pp. 2650–2663, 2019.

[31] F. Pind, C.-H. Jeong, A. P. Engsig-Karup, J. S. Hesthaven, and J. Strømann-Andersen, "Time-domain room acoustic simulations with extended-reacting porous absorbers using the discontinuous Galerkin method," *J. Acoust. Soc. Am.*, vol. 148, no. 5, pp. 2851–2863, 2020.

[32] A. Melander, E. Strøm, F. Pind, A. P. Engsig-Karup, C.-H. Jeong, T. Warburton, N. Chalmers, and J. S. Hesthaven, "Massively parallel nodal discontinous Galerkin finite element method simulator for room acoustics," *Int. J. High Perform. Comput. Appl.*, vol. 38, no. 3, pp. 154–174, 2024.

[33] S. Siltanen, T. Lokki, and L. Savioja, "Rays or waves? understanding the strengths and weaknesses of computational room acoustics modeling techniques," in *Proc. Int. Symp. Room Acoust. (ISRA)*, Melbourne, Australia, 2010.

[34] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Interspeech*, Shanghai, China, 2020, pp. 5036–5040.

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. 31st Conf. Neural Inf. Process. Syst. (NIPS)*, Long Beach, CA, USA, 2017.

[36] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "gpuRIR: A python library for room impulse response simulation with GPU acceleration," *Multimed. Tools Appl.*, vol. 80, no. 4, pp. 5653–5671, 2021.

[37] Z. Chen, T. Yoshioka, L. Lu, T. Zhou, Z. Meng, Y. Luo, J. Wu, X. Xiao, and J. Li, "Continuous speech separation: dataset and analysis," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Online conference, 2020.

[38] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, South Brisbane, QLD, Australia, 2015, pp. 5206–5210.

[39] G. Götz, S. J. Schlecht, and V. Pulkki, "A dataset of higher-order Ambisonic room impulse responses and 3D models measured in a room with varying furniture," in *Proc. Int. Conf. Immersive 3D Audio (I3DA)*, Online conference, 2021.

[40] T. McKenzie, L. McCormack, and C. Hold, "Dataset of spatial room impulse responses in a variable acoustics room for six degrees-of-freedom rendering and analysis," *arXiv preprint arXiv:2111.11882*, 2021.

[41] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas, S. Gannot, and B. Raj, "The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, USA, 2013.

[42] E. A. P. Habets, I. Cohen, and S. Gannot, "Generating nonstationary multisensor signals under a spatial coherence constraint," *J. Acoust. Soc. Am.*, vol. 124, no. 5, pp. 2911–2917, 2008.

[43] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The Kaldi speech recognition toolkit," in *Proc. IEEE Workshop Autom. Speech Recognit. Underst. (ASRU)*, Waikoloa Village, HI, USA, 2011.

[44] L. Lu, X. Xiao, Z. Chen, and Y. Gong, "PyKaldi2: Yet another speech toolkit based on Kaldi and PyTorch," *arXiv preprint arXiv:1907.05955*, 2019.

[45] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Vancouver, BC, Canada, 2013, pp. 6645–6649.

[46] L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer, "Maximum mutual information estimation of hidden markov model parameters for speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Tokyo, Japan, 1986, pp. 49–52.