A Study on More Realistic Room Simulation for Far-Field Keyword Spotting

Eric Bezzam[*‡], Robin Scheibler[†], Cyril Cadoux[‡], and Thibault Gisselbrecht[*]

[*] Sonos Inc., Paris, France
[†] Line Corporation, Tokyo, Japan
[‡] École Polytechnique Fédérale de Lausanne, Switzerland
E-mail: eric.bezzam@epfl.ch

Abstract—We investigate the impact of more realistic room simulation for training far-field keyword spotting systems without fine-tuning on in-domain data. To this end, we study the impact of incorporating the following factors in the room impulse response (RIR) generation: air absorption, surface- and frequency-dependent coefficients of real materials, and stochastic ray tracing. Through an ablation study, a wake word task is used to measure the impact of these factors in comparison with a ground-truth set of measured RIRs. On a hold-out set of re-recordings under clean and noisy far-field conditions, we demonstrate up to 35.8% relative improvement over the commonly-used (single absorption coefficient) image source method. Source code is made available in the *Pyroomacoustics* package, allowing others to incorporate these techniques in their work.

I. INTRODUCTION

Deep-learning approaches are the state-of-the-art when it comes to keyword spotting (KWS) and automatic speech recognition (ASR) [1]. Their recent success can be attributed to the availability of large datasets, improved computing resources, and a cocktail of deep-learning heuristics that have been tried-and-tested by an active research community. Each data point and its reproducibility contribute to a better understanding of the best practices for speech recognition training.

Given an appropriate dataset, deep neural networks (DNN) are able to learn internal representations that are relatively stable with respect to variables independent of the desired output. For example in KWS and ASR tasks, the speaker's pronunciation should not impact the understanding of what the speaker is saying. Robustness to speaker pronunciation, e.g. gender and accents, is typically handled by collecting a sufficiently varied dataset, such as Librispeech for ASR [2], Google's speech commands for KWS [3] and the "Hey Snips" dataset for wake word (WW) detection [4]. Additional variations, e.g. speed and pitch, can be simulated to further augment the original dataset and expose the model to more variety during training [5], [6].

With the recent proliferation of smart speaker devices, there has been a growing interest and need for robust, far-field recognition, namely speech in the presence of reverberation and various types of noise. As collecting and labelling such in-domain data for training can be difficult and time-consuming, multi-condition training (MCT), i.e. simulating various acoustic environments, offers an attractive alternative [7], [8], [9]. Far-field settings are simulated by convolving clean anechoic

recordings with room impulse responses (RIR), which can be either simulated or measured.

In [8], [9], the authors augment their dataset through MCT with RIRs that are generated with the image source method (ISM) [10]. Both [8] and [9] show gains on a hold-out set of real recordings. The authors of [11] investigate a room simulation technique that uses stochastic ray tracing (SRT), in order to incorporate factors not taken into account by ISM, i.e. diffuse reflections and late reverberation. Their results demonstrate an improved performance on KWS and ASR tasks when using SRT instead of ISM.

The use of measured RIRs for MCT is preferable as it is difficult for simulation to fully capture the complexity of real rooms, e.g. the exact shape, materials, and furniture inside, and how sound propagates within them, e.g. low frequency wave effects. Although not as time-consuming as collecting in-domain data, gathering a sufficiently varied dataset of measured RIRs (in terms of number of rooms) is equally difficult to obtain. Therefore, most of the past work has relied on simulated RIRs or a mix of simulated and measured RIRs [8]. Moreover, there have been numerous studies on the differences between using simulated and measured RIRs, with several attempts to close the performance gap: point-source noises [8], directional sources [12], and artificially mimicking low-frequency wave effects [13].

SpecAugment [14] is another data augmentation technique that has shown promising results for ASR. Within the context of far-field recognition, its effect is unclear, with worse results shown in [15] when used in conjunction with MCT. As our focus is on far-field recognition and the use of MCT to generalize to these conditions, we do not use SpecAugment in our study.

While only SRT with a single absorption coefficient is used in [11], we propose to employ a *hybrid* approach by complementing ISM with SRT [16]. We also incorporate air absorption and surface- and frequency-dependent absorption coefficients of real materials in an attempt to bring simulation closer to reality. The goal of this paper is two-fold:

- Study the impact of incorporating the above factors into ISM with an ablation study on a WW detection task. To the best of our knowledge, there is no such ablation study within the context of speech recognition applications.
- Demonstrate and make available such techniques through

the *Pyroomacoustics* package [17].[1] [2]

In our ablation study, we use a dataset of measured RIRs, namely the BUT Speech@FIT Reverb Database (ReverbDB) [18], to quantify the gap between currently-used room simulation techniques for KWS and ASR, i.e. ISM and SRT, and the ideal scenario of having a dataset consisting of only measured RIRs. Although ReverbDB does not contain the amount of rooms and RIRs to train a model robust to the large variety that would be encountered by a typical smart speaker device, it is of sufficient size to serve as an oracle dataset for our controlled setup and contains the metadata of each room to simulate the various proposed factors.

The paper is organized as such: Section II presents ISM and SRT; Section III describes how the proposed factors, i.e. the hybrid approach, air absorption, and surface- and frequency- dependent absorption coefficients of real materials, are incorporated; Section IV and V detail our experimental setup and results; and Section VI contains concluding remarks.

II. MODELING ROOM IMPULSE RESPONSES

For I sources, we can model far-field speech as:

$$x[n] = \sum_{i=1}^{I} g_i \left(p_i[n] * h_i[n] \right), \tag{1}$$

where $\{p_i[n]\}_{i=1}^{I}$ are the I source signals (target and noise), each convolved with an RIR $\{h_i[n]\}_{i=1}^{I}$ from the i-th source to the microphone and scaled with a constant $\{g_i\}_{i=1}^{I}$.

A. Image source method

The ISM technique is a popular approach for generating RIRs, mainly due to its simplicity [8], [9], [10]. It models reflections off walls as virtual sources outside of the room, but at a distance corresponding to the length of the reflection path within the room. With ISM, we can write the RIR for the i-th source as:

$$h_i[n] = \sum_{\boldsymbol{s} \in V(\boldsymbol{m}, \boldsymbol{s}_i)} \frac{R(\boldsymbol{m}, \boldsymbol{s})}{4\pi \|\boldsymbol{m} - \boldsymbol{s}\|} \delta_{LP} \left(n - F_s \frac{\|\boldsymbol{m} - \boldsymbol{s}\|}{c} \right), \tag{2}$$

where $\boldsymbol{m}$ is the position of the microphone, $\boldsymbol{s}_i$ is the position of the i-th source, $V(\boldsymbol{m}, \boldsymbol{s}_i)$ is the set of visible image sources for the reflections between $\boldsymbol{m}$ and $\boldsymbol{s}_i$, δ_{LP} is a windowed sinc function [17], F_s is the sample rate, c is the speed of sound, and $R(\boldsymbol{m}, \boldsymbol{s})$ is the accumulated reflection coefficient between $\boldsymbol{m}$ and $\boldsymbol{s}$:

$$R(\boldsymbol{m}, \boldsymbol{s}) = \prod_{w \in W(\boldsymbol{m}, \boldsymbol{s})} \sqrt{1 - \alpha_w^2}, \tag{3}$$

where $W(\boldsymbol{m}, \boldsymbol{s})$ is the list of surfaces in the reflection path between $\boldsymbol{m}$ and $\boldsymbol{s}$, and α_w is the absorption coefficient of surface w.

---

[1] github.com/LCAV/pyroomacoustics
[2] Room impulse response generation code for this paper:
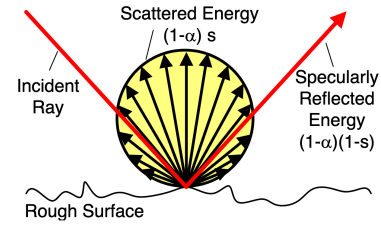github.com/ebezzam/room-simulation

Fig. 1: Illustration of the separation of reflected energy into scattered and specular components [16]. α and s are the surface's absorption and scattering coefficients.

A maximum reflection order M, i.e. how many wall reflections to simulate, is selected to build $V(\boldsymbol{m}, \boldsymbol{s}_i)$. For arbitrary-shaped rooms, the simulation complexity is exponential with regards to the reflection order: $\mathcal{O}(N^M)$ where N is the number of reflective surfaces. For cuboid-shaped (i.e. shoebox) rooms, the complexity simplifies to $\mathcal{O}(M^3)$ due to the symmetry of the room [10].

B. Stochastic ray tracing

Up to a desired order M, the ISM technique is able to perfectly model *specular* reflections, i.e. reflections that have the same outgoing angle as the incident angle to a surface. This modeling is accurate only as long as the wavelength of the sound is small relative to the size of the reflector. When this is not the case, we have *diffuse* reflections that scatter in all directions [16]. Fig. 1 illustrates this separation into specular and scattered energy.

Furthermore, the reflection order M limits the ability of ISM to capture the late reverberation of a room. Increasing the ISM order can capture more reflections in order to better model the late reverberation, but diffuse reflections are still not taken into account. With SRT, we can model these diffuse reflections and the late reverberation [11], [16]. From each point source, a large number of rays are emitted and traced until the energy of each ray falls below a certain threshold. Each microphone is modeled as a receiver volume, with specular rays that intersect with it contributing to the resulting RIR. The received rays are logged with an amplitude and timestamp in order to produce an energy histogram.

Unlike ISM, each reflection may emit scattered sound energy back to the receiver. In this work, we employ the *diffuse rain* technique [16], in which a secondary ray is emitted at each reflection point in the direction of the receiver. The energy of this secondary ray can be written as:

$$E_{scat} = E_{in} \cdot (1 - \alpha) \cdot s \cdot P_{Hit}, \tag{4}$$

where E_{in} is the ray's incoming energy, α and s are the surface's absorption and scattering coefficients respectively, and P_{Hit} is the probability that the scattered energy reaches the receiver [16]. The specular ray's energy is given by:

$$E_{spec} = E_{in} \cdot (1 - \alpha) \cdot (1 - s). \tag{5}$$

In order to produce the RIR, the envelope of the resulting energy histogram is used to shape the envelope of a randomly generated sequence of Dirac deltas [16].

III. PROPOSED SIMULATION FOR TRAINING

In Section III-A, we describe the hybrid room simulation technique; and in Section III-B, we discuss how to incorporate air absorption and surface- and frequency-dependent coefficients.

A. Hybrid approach

Due to its stochastic nature, we are not guaranteed to receive all specular reflections with SRT. A hybrid approach can capture these reflections up to a desired order M:

1) Apply ISM of order M for specular reflections: $h_i^{ISM}[n]$.
2) Apply SRT for diffuse reflections and late reverberation: $h_i^{SRT}[n]$.
3) Add the two simulations for the hybrid RIR:

$$h_i[n] = h_i^{ISM}[n] + h_i^{SRT}[n]. \qquad (6)$$

When performing SRT, specular reflections which are at or below M should be neglected, as to avoid counting the same reflections twice. Moreover, the energy levels between ISM and SRT must be balanced at the start of simulation [16].

B. Additional factors

Air absorption is incorporated by introducing a $e^{-\gamma r}$ factor to the RIRs, where r is the total distance travelled and γ is an air attenuation coefficient. This coefficient depends on temperature, humidity, and frequency [19].

Current approaches to room simulation for speech recognition employ a single absorption coefficient for the entire room [8], [9]. In reality, absorption properties depend on the material of the surface and on the frequency. This dependence can be taken into account by performing ISM and SRT with unique absorption coefficients for each surface with 3 and on separate octave bands. An appropriately designed filter bank $\{\phi_f[n]\}_{f=1}^F$ is used to combine the frequency-dependent RIRs at the end of simulation:

$$h_i[n] = \sum_{f=1}^{F} \phi_f[n] * (h_i^{ISM,f}[n] + h_i^{SRT,f}[n]). \qquad (7)$$

In this work, we use a third-octave filter bank, with $F = 7$ FIR bandpass filters centered at $[125, 250, 500, 1000, 2000, 4000, 8000]$ Hz, as these are frequencies at which absorption and scattering coefficients of materials are typically measured [20]. The filters are designed using the window method with an FFT of 512 and a raised cosine window. The frequency response of each filter can be seen in Fig. 2.
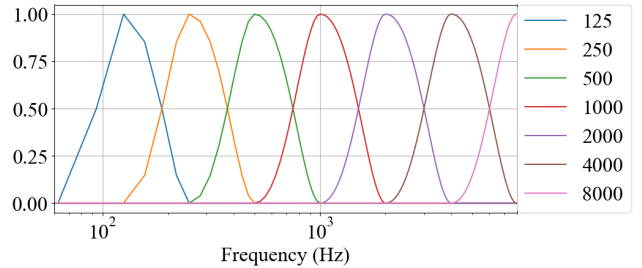

Fig. 2: Third-octave filter bank to simulate frequency-dependent room impulse responses.

IV. EXPERIMENTAL SETUP

A WW task is used for our ablation study as the smaller model size allows for quicker iterations and more trainings in order to perform a more in-depth analysis. Moreover, a binary task allows the use of DET (Detection Error Tradeoff) curves to analyze the performance over a wide range of operating points instead of a fixed point.

Three types of datasets are used in our study:

1) WW data (positive and negative utterances from the same speakers): "Hey Snips" dataset [4].
2) Measured RIRs: BUT Speech@FIT Reverb Database (ReverbDB) [18].
3) Noise datasets: MUSAN [21] for non-speech sounds and Librispeech [2] for interfering speech.

We use ReverbDB as the oracle RIRs that we wish to emulate via simulation. The dataset contains essential metadata in order to study the impact of the factors we wish to investigate. In Table I, the train / dev / test set distribution (before augmentation) is detailed.

In Section IV-A, we describe how ReverbDB is used to simulate RIRs with the different simulation factors. In Section IV-B and Section IV-C, we describe the augmented train, dev, and test sets for our study.

A. RIR simulation

For each room simulation variation, we emulate the RIRs from the ReverbDB dataset using the provided metadata. The split in Table I corresponds to 1032 RIRs for the train set and 280 RIRs for the dev set. For all simulations, the same room dimensions, microphone positions, and speaker positions are used, as specified in the ReverbDB metadata.

1) **Baseline**: a single absorption coefficient is used for all frequencies and all walls of a shoebox room as in [8],

TABLE I: Original dataset distribution.

	train	dev	test
# "Hey Snips"	5876	2588	2504
# negative	45344	20821	20321
# rooms	7	2	1
# noise samples	15504	2299	2087

[9], [11]. Using Eyring's equation as in [9], this coefficient is computed from the estimated RT60[3] provided in the ReverbDB metadata for each microphone-speaker pair.

2) **AIR**: the measured temperature for each room in ReverbDB is used to set the appropriate frequency-dependent air absorption coefficients as specified in [20]. This temperature is also used to set the speed of sound.[4]

3) **MAT**: using the name of the floor, ceiling, and wall materials provided for each room in ReverbDB, we identify the corresponding frequency-dependent coefficients in [20] for each wall. These values are used instead of the single absorption coefficient. It is important to note that these absorption coefficients are not provided in the ReverbDB metadata but are taken from a separate resource. Therefore, the values may not represent the actual absorption coefficients of the room walls. Moreover, we do not take into account furniture in the room.

4) **MAT AIR**: Finally, we use both AIR and MAT.

All four of the above variations are applied for ISM, SRT, and the hybrid approach (HYB) for our ablation study. An ISM order of 17 is used for both ISM and HYB. For the ray tracing approaches, we use the percentage of furniture covering in the ReverbDB metadata as a single scattering coefficient, even for the frequency-dependent simulation in MAT. All of the simulation variations are generated using *Pyroomacoustics*. As a direct comparison to [11], we also perform Baseline with their simulator;[5] incorporating the other factors is not possible with their current library.

B. Train and dev sets

The original train and dev sets are augmented as such, with the WW data (positive and negative utterances) as the target speaker:

1) A room is sampled at random, from which we sample a microphone, one speaker for the target speaker, and (75% of the time) 1 or 2 speaker(s) to serve as point-source noise(s) as in [8].

2) Pitch and speed variations are applied to the target speaker as in [5].

3) Using 1, the target and sampled noise sources are simulated in the corresponding room with an SNR sampled as in [9].

Each sample in the original train set is augmented 16 times with different samplings in order to produce the dataset that is used for training the WW detector. This 16x augmentation is performed for all 13 room simulation variations described in Section IV-A and the ReverbDB dataset, resulting in a total of 14 types of room simulation.

C. Test set

For evaluating the different room simulation approaches, we use re-recordings in order to observe how the simulation

approaches generalize to real far-field scenarios. The original test set in Table I is re-recorded in five conditions within a typical office setting: clean, 5 dB non-speech, 5 dB speech, 2 dB non-speech, and 2 dB speech. The target speaker is placed 3 m away. For the noisy conditions, a single speaker is placed 2 m away at a 45° angle with respect to the target.

D. Model architecture and training

The WW model employed in this study follows the architecture and training hyper-parameters from [4]. The model takes inspiration from WaveNet [22], as it uses dilated convolutions to capture long temporal patterns, gated activation units to control the propagation of information to the next layer, and skip connections to speed up convergence and avoid vanishing gradient issues. In total, there are 24 layers, corresponding to a receptive field of 1.83 s. The input acoustic features are 20-dimensional log-Mel filterbank energies (LFBEs), extracted every 10 ms over a window of 25 ms. Xavier initialization is used, and the Adam optimization method is used with a learning rate of 10^{-3}.

As the focus of this paper is on room simulation techniques, we refer the interested reader to [4] for a more detailed description on model architecture and training hyper-parameters, and an ablation study on the importance of gating and skip connections.

For each simulation approach, 10 models are trained with different seeds in order to prevent random effects due to initialization or sampling from affecting the analysis of our results (by averaging the performance over the 10 models).

V. RESULTS AND ANALYSIS

The top half of Table II shows the average relative change in false rejection rate for each simulation approach with respect to using measured RIRs (ReverbDB). This value is computed by averaging over the relative changes for the five re-recorded far-field conditions. Fig. 3 shows the DET curves for a select few simulation techniques.

A significant gap can be observed between the commonly-used ISM approach and using the measured RIRs of ReverbDB: 55.1% drop in average relative performance. This gap can be seen for each condition in Fig. 3.

From the Baseline column in Table II, we observe that the proposed HYB approach outperforms currently-used techniques: ISM [8], [9] and SRT [11]. As presented in [11], we observe an improvement of SRT over ISM. The further increase in performance from SRT to HYB seems to indicate the need of early specular reflections, which are not guaranteed with SRT alone. The difference between our implementation of SRT and [11] may arise from different simulation parameters and internal implementations.[6]

From Table II, we see that AIR benefits all baseline techniques; perhaps the induced energy decay results in more realistic RIRs. On the other hand, the impact of MAT is not clear. For ISM, we observe a significant improvement,

---

[3]Time for the sound to decay by 60dB.

[4]speed of sound $= 331.4 + 0.6 \times$ temperature

[5] `github.com/RoyJames/pygsound`

[6]See [23] for implementation details of the simulator in [11].

TABLE II: (Top half) average relative change in false rejection rate with respect to using measured RIRs (ReverbDB), for three false alarms per day. Higher is better, with values close to 0 indicating that the simulation technique is on par with using measured RIRs. (Bottom half) average relative change with respect to using ISM. Confidence intervals are for three standard deviations.

	Baseline	AIR	MAT	MAT AIR
ISM	$-55.1 \pm 1.61\%$	$-16.7 \pm 0.453\%$	$0.78 \pm 0.291\%$	$-14.3 \pm 0.709\%$
SRT [11]	$-32.2 \pm 1.02\%$	$-$	$-$	$-$
SRT	$-18.0 \pm 0.403\%$	$-14.6 \pm 0.219\%$	$-52.2 \pm 1.33\%$	$-25.1 \pm 0.670\%$
HYB	$-10.4 \pm 0.284\%$	$-5.02 \pm 0.321\%$	$-40.3 \pm 0.850\%$	$-34.9 \pm 0.859\%$
ISM	$-$	$24.3 \pm 0.830\%$	$35.8 \pm 1.60\%$	$25.3 \pm 1.14\%$
SRT [11]	$15.2 \pm 0.920\%$	$-$	$-$	$-$
SRT	$24.1 \pm 1.38\%$	$27.1 \pm 1.82\%$	$3.92 \pm 0.817\%$	$20.0 \pm 1.60\%$
HYB	$28.8 \pm 1.80\%$	$33.0 \pm 2.11\%$	$10.5 \pm 0.775\%$	$13.9 \pm 0.977\%$

(a) clean   (b) non-speech $5\,\mathrm{dB}$ (c) speech $5\,\mathrm{dB}$

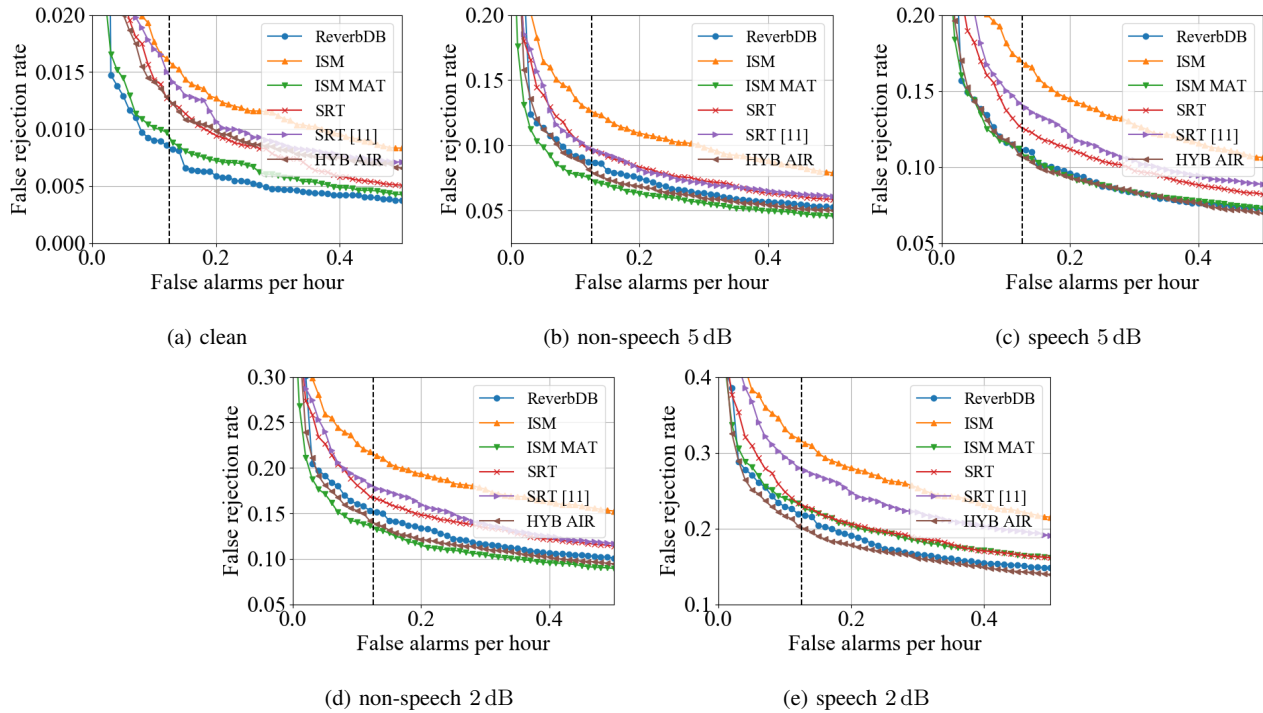(d) non-speech $2\,\mathrm{dB}$ (e) speech $2\,\mathrm{dB}$

Fig. 3: Detection error tradeoff curves comparing room simulation techniques (averaged over 10 random seeds) for a WW ("Hey Snips") detection task: false alarm per hour (x-axis) vs. false rejection rate (y-axis). Vertical, black dotted lines indicate the performance for three false alarms per day. For all conditions, the target is placed $3\,\mathrm{m}$ from the device and for noisy conditions, the interfering sound is placed $2\,\mathrm{m}$ at a $45°$ angle with respect to the target. Note that the y-axis is scaled differently for each condition to better visualize the difference between simulation approaches.

effectively closing the gap with ReverbDB. For t he ray tracing techniques, we see a large degradation in performance with respect to the baseline approaches. A potential reason for this is the scattering coefficient used in simulation. As there is no metadata on the scattering properties of the rooms in the ReverbDB, we use the furniture coverage as a scattering coefficient for all surfaces and frequencies, which could be very far from reality. This modeling choice would not impact ISM simulation as scattering is not taken into account. Furthermore, the MAT AIR column seems to indicate we do not get additive gains from incorporating both proposed factors.

The effect of MAT highlights a fundamental issue that arises when trying to match simulation to real world RIRs, namely the appropriate choice of simulation parameters and their distribution. It could be that the name of the wall materials, as provided in the ReverbDB metadata, is not enough information to perform a surface- and frequency- dependent emulation of a room. As the absorption coefficients used for MAT are deduced from a separate dataset [20], they could be very different from the coefficients of the wall (and furniture) materials in the corresponding room of ReverbDB.

Despite the unclear results on incorporating MAT, there
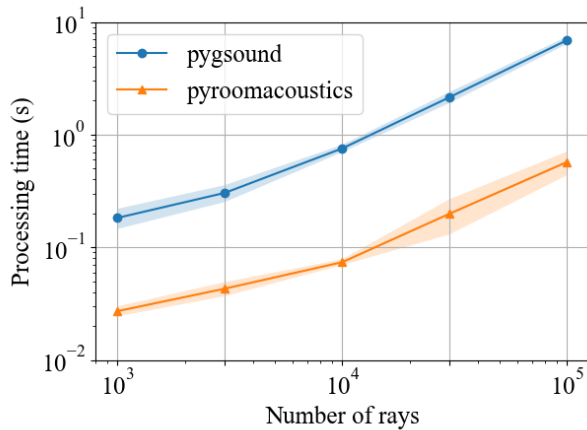
(a) Varying number of emitted rays (for diffuse reflections).
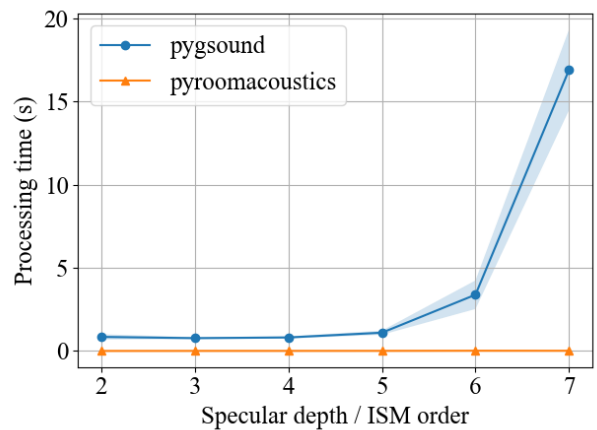
(b) Varying specular depth / ISM order.

Fig. 4: Average runtime (over 100 trials) for computing an RIR within a $8\text{m} \times 9\text{m} \times 3\text{m}$ room with an RT60 of 0.5s and a scattering coefficient of 0.5. Computations are performed on a MacBook Pro 2014 with a 2.6 GHz Dual-Core. In 4a, the number of emitted rays is varied with a fixed specular depth / ISM order of 3; in 4b the specular depth is increased for *pygsound* while the ISM order is increased for *Pyroomacoustics* for a fixed number of emitted rays ($1e4$). Confidence intervals are for one standard deviation.

is a clear indication that adding ray tracing and AIR is beneficial. Moreover, all proposed simulation factors display an improvement over ISM (bottom half of Table II).

A. Profiling RIR simulators

In Fig. 4, we compare the computational time between our implementation of SRT in *Pyroomacoustics* with that of *pygsound*, namely the simulator used in [11]. For an increasing number of emitted ray in Fig. 4a, we observe that *Pyroomacoustics* is roughly an order of magnitude faster than *pygsound*. For an increasing specular depth in Fig. 4b, we see an exponential growth in *pygsound*. Specular depth is roughly equivalent to the ISM order, namely how many specular wall reflections are included in the RIR modeling. This seems to suggest that *pygsound* does not take into account the symmetry of shoebox-shaped rooms.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we investigated the impact of more realistic room simulations on far-field wake word (WW) training without fine tuning on in-domain, recorded data. To this end, we quantified the gap between an oracle set of measured RIRs (ReverbDB) and the commonly-used ISM approach for room simulation. Through an ablation study, we studied the effect of incorporating additional simulation factors, i.e. stochastic ray tracing for diffuse and late reverberation, air attenuation (AIR), and surface- and frequency-dependent absorption coefficients (MAT).

On a hold-out set of re-recordings under clean and noisy far-field conditions, we observe a 28.8% average relative improvement to ISM when complementing it with ray tracing (HYB). Moreover, we find that incorporating AIR benefits all

simulation techniques (a further improvement to 33.0% when used with HYB). The impact of MAT is unclear. With ISM, we found the gap between simulated and measured RIRs to be effectively closed. With ray tracing techniques, we found a degradation in performance, which may be due to improper modeling of scattering properties, for which we did not have metadata about in ReverbDB.

As future work, we would like to investigate the impact of such factors on other far-field speech recognition tasks, e.g. command detection and ASR. Finally, the source code for generating the RIRs is made available in the *Pyroomacoustics* package. We hope this will aid other researchers and engineers in incorporating such techniques in their work.

The purpose of this work was to quantify the gap between various simulation techniques and the ideal case of using measured RIRs. However, in order to have a fair comparison with the measured RIRs, we were limited in terms of the number of rooms and RIRs we could simulate. However, to use the techniques presented in this paper and achieve the variety of rooms needed for far-field KWS and ASR tasks, one could envision a room generation scheme similar to one in [9]; where instead of sampling an RT60, variations in air absorption attenuation and different frequency-dependent materials for each wall could be sampled.[7]

REFERENCES

[1] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, and B. Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition," *IEEE Signal Processing Magazine*, vol. 29, 2012.

---

[7]In the repository for this paper, we have included all the frequency-dependent coefficients from [20]:`room-simulation/materials.py`

[2] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR Corpus Based on Public Domain Audio Books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[3] P. Warden, "Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition," *arXiv preprint arXiv:1804.03209*, 2018.

[4] A. Coucke, M. Chlieh, T. Gisselbrecht, D. Leroy, M. Poumeyrol, and T. Lavril, "Efficient Keyword Spotting Using Dilated Convolutions and Gating," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6351–6355.

[5] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio Augmentation for Speech Recognition," in *Interspeech*, vol. 100, 2013, pp. 144–152.

[6] N. Jaitly and G. Hinton, "Vocal Tract Length Perturbation (VTLP) Improves Speech Recognition," in *ICML Workshop on Deep Learning for Audio, Speech and Language*, vol. 117, 2013.

[7] Y. Gao, Y. Mishchenko, A. Shah, S. Matsoukas, and S. Vitaladevuni, "Towards Data-Efficient Modeling for Wake Word Spotting," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7479–7483.

[8] 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 *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5220–5224.

[9] C. Kim, A. Misra, K. Chin, T. Hughes, A. Narayanan, T. N. 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 *Interspeech*. ISCA: ISCA, 2017, pp. 379–383.

[10] 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.

[11] Z. Tang, L. Chen, B. Wu, D. Yu, and D. Manocha, "Improving Reverberant Speech Training Using Diffuse Acoustic Simulation," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, no. 1, 2020, pp. 6969–6973.

[12] M. Ravanelli, P. Svaizer, and M. Omologo, "Realistic Multi-Microphone Data Simulation for Distant Speech Recognition," in *Interspeech*, vol. 08-12-Sept, no. 1, 2016, pp. 2786–2790.

[13] Z. Tang, H.-Y. Meng, and D. Manocha, "Low-Frequency Compensated Synthetic Impulse Responses for Improved Far-Field Speech Recognition," *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6974–6978, 2020.

[14] D. S. Park, W. Chan, Y. Zhang, C. C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," *Interspeech*, vol. 2019-Septe, pp. 2613–2617, 2019.

[15] D. S. Park, Y. Zhang, C.-C. Chiu, Y. Chen, B. Li, W. Chan, Q. V. Le, and Y. Wu, "SpecAugment on Large Scale Datasets," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6879–6883.

[16] D. Schröder, "Physically Based Real-Time Auralization of Interactive Virtual Environments," Ph.D. dissertation, RWTH Aachen, 2011.

[17] R. Scheibler, E. Bezzam, and I. Dokmanic, "Pyroomacoustics: A Python Package for Audio Room Simulation and Array Processing Algorithms," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 351–355.

[18] I. Szoke, M. Skacel, L. Mosner, J. Paliesek, and J. H. Cernocky, "Building and Evaluation of a Real Room Impulse Response Dataset," *IEEE Journal on Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 863–876, 2019.

[19] H. Kuttruff, *Room Acoustics*, 4th ed. CRC Press, 2000.

[20] M. Vorländer, *Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality*, 1st ed., ser. RWTHedition. Springer-Verlag Berlin Heidelberg, 2008.

[21] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," *arXiv preprint arXiv:1510.08484*, pp. 2–5, 2015.

[22] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," in *9th ISCA Speech Synthesis Workshop*, 2016, pp. 1–15.

[23] C. Schissler and D. Manocha, "GSound: Interactive Sound Propagation for Games," *Proceedings of the AES International Conference*, 2011.