An Overview of Methods for Generating, Augmenting and Evaluating Room Impulse Response Using Artificial Neural Networks

Mantas Tamulionis

Vilnius Gediminas Technical University, Saulėtekio al. 11, Vilnius, Lithuania
mantas.tamulionis@vilniustech.lt

Abstract. *Methods based on artificial neural networks (ANN) are widely used in various audio signal processing tasks. This provides opportunities to optimize processes and save resources required for calculations. One of the main objects we need to get to numerically capture the acoustics of a room is the room impulse response (RIR). Increasingly, research authors choose not to record these impulses in a real room but to generate them using ANN, as this gives them the freedom to prepare unlimited-sized training datasets. Neural networks are also used to augment the generated impulses to make them similar to the ones actually recorded. The widest use of ANN so far is observed in the evaluation of the generated results, for example, in automatic speech recognition (ASR) tasks. This review also describes datasets of recorded RIR impulses commonly found in various studies that are used as training data for neural networks.*

Keywords. Room impulse response, reverberation, acoustic simulation, data augmentation, artificial neural networks, speech recognition.

Reprint. This article was first published in Mokslas – Lietuvos Ateitis. Elektronika Ir Elektros Inžinerija, 13, 1–5. https://doi.org/10.3846/mla.2021.15152

1 Introduction

Room impulse response is a transfer function that describes the acoustics of a room, corresponding to one specific position between the sound source and the listener. Features of the RIR depend on the geometry of the room, the absorption and scattering coefficients of the surfaces, the distances of the source and receiver to the nearest reflecting surface and to each other. The RIR consists of silence at the beginning (its deuration determines how long it takes for the signal to travel to the receiver), as well as early and late reflections. We can convolve the RIR with an anechoic signal and thus place the signal virtually in the desired room. We can record the RIR in a real room, but if there is at least a slight change in the position of the source or receiver, we should repeat the recording. It can also be modeled using acoustic modeling algorithms, there are popular commercial applications such as ODEON or CATT-Acoustic, but using these applications it is very difficult to obtain authenticity due to the standard absorption coefficients assigned to the surfaces. These algorithms generate RIR using the Image source method (ISM). This method allows us to expect realistic results only if we model an almost empty room with standard geometric shapes. With the development of ANN technologies, in recent years, they have also been applied to the estimation of RIRs. ANN can be used not only to generate RIRs, but also to augment impulses generated by other methods to make them similar to recorded RIRs. ANN can also be used to perform evaluation tasks on proposed RIR generation methods. In this review, we will discuss methods for applying ANN to achieve all of these goals.

2 Estimation and generation methods

Tang et al. proposed a new geometric acoustic simulation method (GAS), which was compared with the ISM method. The article states that this method allows to model not only specular but also diffuse reflections, which makes it possible to simulate rooms with much more complex geometric shapes and more reflective surfaces (Tang et al., 2020a). GAS is based on Monte Carlo path tracing, which differs from the ISM method in that the reflections are generated in randomly selected directions. The authors report that their proposed method cannot model diffraction and low frequency reflections. This algorithm does not use neural networks to generate RIRs, but they are used in evaluation tasks and will be discussed in Section 4.

Ratnarajah and colleagues presented a method for generating RIRs using the Generative Adversarial Network (GAN) and named it IR-GAN. The authors used the WaveGAN structure for their work, which was originally designed to generate short audio files (Donahue et al., 2019). The structure of WaveGAN is one-dimensional deep convolutional generative adversarial networks (DCGANs) that first generate a spectrogram and then convert it into an audio signal. In this case, the GAN trained from a dataset of RIRs recorded in a real room, and could later change the acoustic parameters of the generated RIRs, such as reverberation time (RT60), direct to reverberant ratio (DRR) and others, to generate an unlimited number of new RIRs simulating new rooms (Ratnarajah et al., 2020). It should be noted that the authors converted the recorded RIRs to 16 kHz before sending them to the network for training, which means that high frequency energy is removed from the RIRs.

Yu and Kleijn presented a method for estimating room acoustic parameters. Separate algorithms estimate the geometry of the room and the absorption coefficients of its surfaces. Convolutional neural networks (CNNs) are used to estimate geometry, and feedforward multilayer perceptrons (MLPs) are used to estimate absorption coefficients (Yu & Kleijn, 2021). The authors state that satisfactory results can be achieved by training neural networks with only one RIR impulse, although increasing the learning dataset slightly improves the performance of the algorithms.

For room geometry estimation, the CNN consisting of eight one-dimensional convolutional layers and three fully connected layers was used. Each convolutional layer was followed by a one-dimensional batch normalization layer and a leaky rectified linear unit (Leaky ReLU) activation function. The CNN at its end has three output nodes that provide the length, width, and height of the room. CNN was first trained with simulated RIRs, later the model was adapted to work well with recorded RIRs. The simulated RIRs were generated by the ISM method.

The estimation of surface absorption coefficients was tested only on a set of simulated RIR data, as databases of recorded RIRs together with their absorption coefficients are not usually available. Both in the geometry estimation and at this stage, time domain RIRs were used. Surface absorption coefficients usually differ when analyzing individual frequencies, so the authors performed an additional processing step before sending impulses to neural networks - dividing RIRs into several frequency bands. In this way, the estimation can be performed for each frequency band separately. Chebyshev type I, 10th order filters were chosen for filtering as it allowed to achieve higher computational speed. The MLP used for this estimation had nine hidden layers, the number of neurons in each of them was halved from 2048 to 8 neurons each time. A rectified linear unit (ReLU) activation function was used after each hidden layer.

3 Datasets

In the ASpIRE (Automatic Speech Recognition In Reverberant Environments) challenge, participants worked with different datasets for training, development and evaluation (Harper, 2015). The Fisher conversational telephone corpus dataset (Cieri et al., 2004), which contains more than 10,000 telephone conversations in English was provided for training. The Mixer 6 corpus dataset (Brandschain et al., 2010), which contains 1.425 telephone conversations recorded in two different rooms using 15 differently arranged microphones, was designed for development. A new database for the evaluation of algorithms was created and named “Mixer 8 pilot corpus”. It differed from the Mixer 6 corpus in that recordings were made in seven different rooms using 8 microphones spaced at different distances.

Ko et al. in their study compared the simulated and recorded RIRs. They compiled a recorded RIRs database consisting of the RWCP (Nakamura et al., 2000), the REVERB challenge (Kinoshita et al., 2013), and the Aachen impulse response (AIR) datasets (Jeub et al., 2009). They were able to achieve satisfactory results in the study only after adding point-source noises to the simulated RIRs. These noises were taken from the MUSAN (music, speech, and noise corpus) database (Snyder et al., 2015).

The authors of IR-GAN compared the RIRs generated by their method with the recorded RIRs of BUT ReverbDB (Szoke et al., 2019) and the aforementioned AIR database. Additionally, in this study, anechoic signals from the LibriSpeech database (Panayotov et al., 2015) were used, which were convolved with both simulated and recorded RIRs. From the BUT ReverbDB database, the authors additionally used environmental noise files that were added to the convolved signals in an attempt to generate the far field signals required for ASR tests.

Yu and Kleijn also used BUT ReverbDB data in their experiments as a set of recorded RIR data. This decision was made due to the large number of impulses in the set from different types of rooms that were not empty during the measurement. The dataset consists of an average of 155 RIRs from each room (5 source and 31 receiver positions). The RT30 parameter of the rooms ranged between 0.59 and 1.85 s. The dataset also contains geometric information for all measured rooms. The simulated RIR dataset was generated using a Room Impulse Response Generator (Habets, 2010) with a sampling rate of 8kHz and a RIR length of 4096 samples, which allowed the generation of impulses lasting approximately 0.5 s. The recorded RIR dataset had a higher sampling rate, but before applying these impulses to neural networks, the authors converted the dataset to a 8 kHz sampling rate, truncated, and continued to use only 4096 samples.

4 Data augmentation methods

Ko and colleagues conducted a study to determine how the difference between the results of the ASR test could be eliminated when simulated and recorded RIRs are used in different tests (Ko et al., 2017). It was found that the ASR test results are significantly improved if we add point-source noises to the simulated RIRs. Table 1 shows how the word error rate (WER) values differ when the same acoustic model is trained on different data – using the point-source noise addition method and without using any augmentation method.

Table 1. Differences in WER values when one of the augmentation methods is applied to the training data of the algorithm

Algorithm / authors	**Augmentation method used**	**WER** [%]
Ko et al.	Without augmentation	40.9
	Addition of point-source noises	27.0
IR-GAN	Without augmentation	19.71
	Constraint method	14.99

The authors of the IR-GAN algorithm had to solve the problem arising from the ability of GAN network to generate an unlimited variety of RIR impulses. There was a high probability that the generated RIRs would be noisy and have an unrealistically large reverberation time. A constraint method was used, which allowed to limit the variety of generated RIRs (Ratnarajah et al., 2020). The limits of the changes in the main acoustic parameters were calculated from the training data, and when generating new RIRs, the GAN network was not allowed to exceed these limits. Table 1 also shows how the WER value improves in the implementation of the IR-GAN algorithm by additionally applying this constraint method.

To adapt the room geometry estimation algorithm proposed by Yu and Kleijn for use with recorded RIRs, the authors used the insights of the SpecAugment method (Park et al., 2019) and added 30–50 dB signal-to-noise ratio (SNR) additive noise to the simulated RIRs (Yu & Kleijn, 2021). Moreover, RIRs generated by the ISM method usually lack information about obstacles and additional objects that can interfere with the sound wave. In real rooms, these objects block the trajectories of reflections or create unusual new reflections. To solve this problem, it is possible to remove or add randomly selected reflections from simulated RIRs or to add blocked reflection structures taken from recorded RIRs. The authors used the adaptive rectangular decomposition (ARD) method (Raghuvanshi et al., 2009) and thus tried to simulate possible obstacles in the room for simulated RIRs.

Table 2. Differences in RMSE values when one of the augmentation methods is removed from the system

Bypassed data augmentation method	Average RMSE [m]
Addition of noise	0.0310
Adding / removing reflections	0.0570
Adding blocked reflection structures	0.0648
ARD method	0.1210

Table 2 shows the root mean square error (RMSE) differences, which represent the accuracy of the results generated by the algorithm compared to known room geometry data. The algorithm uses all augmentation methods listed above. To identify the importance of each of the methods, the experiments were repeated, each time a different method of augmentation was bypassed. From the results, we can see that after the deactivation of the ARD method, the RMSE value increased the most, which means that the use of this method ensures the highest accuracy of the results.

5 Evaluation methods

Intelligence Advanced Research Projects Activity (IARPA) organized a competition called the ASpIRE challenge (Harper, 2015). Those wishing to participate had to develop ASR systems without access to matched data for system training and development. This competition was for the evaluation of far-field recordings and differed from previous competitions in that the algorithms had to work with conversation-type voice data, the number of words in the vocabulary used was not limited, which means that the evaluation dataset could contain words that were not in the training dataset. Participants were not provided with any information about the audio files in the dataset. The algorithm proposed by the winner of this competition used the ROVER system (Fiscus, 1997), which allows to combine several different ASR models and thus obtain better results. In this case, the combination of Gaussian Mixture Models (GMMs) and Deep Neural Networks (DNNs) gave the best results (Hsiao et al., 2015).

In the Ko study, the results were evaluated by performing Large Vocabulary Continuous Speech Recognition (LVCSR) tasks. Tasks were performed using Time-delay neural network (TDNN) and bi-directional long-short-term memory (BLSTM) acoustic models. The authors state that using the RIRs generated by their method, with the added point-source noise, the BLSTM model can achieve a WER value of 24.6% (Ko et al., 2017).

The results of the GAS algorithm are evaluated by performing ASR and Keyword spotting (KWS) tasks. The ASR task was performed using an acoustic model that consists of two layers of two-dimensional convolutional neural network (2D CNN) and five layers of long short-term memory (LSTM). The model used for the KWS task consists of a single-layer 1D CNN and a two-layer LSTM. The ASR task on the BUT ReverbDB database achieved a WER value of 16.53%. The results of the KWS task are measured by equal error rates (EERs). The authors show that their proposed GAS method can reduce EER values by 21% compared to the ISM method (Tang et al., 2020a).

The authors of the IR-GAN algorithm also evaluate the generated RIRs by performing an ASR test, using the Kaldi LibriSpeech acoustic model, which is based on the TDNN network (Tang et al., 2020b). The paper states that the proposed algorithm can reduce WER by almost 9% compared to the GAS method. However, this can only be achieved when the AIR dataset is selected as the training set. The authors also show that by combining RIRs generated by IR-GAN and GAS algorithms, WER can be reduced by more than 14% (Ratnarajah et al., 2020).

Table 3. Comparison of different algorithms, databases used, ASR models and WER results obtained

Method/ authors	**Dataset**	**ASR model**	**WER** [%]
Hsiao et al.	Fisher	GMM & DNN	27.1
Ko et al.	RWCP + REVERB + AIR	BLSTM	24.6
GAS	BUT	2D CNN & LSTM	16.53
IR-GAN	BUT	TDNN	14.99
IR-GAN	AIR	TDNN	7.71

The results of Yu and Kleijn's room geometry estimation algorithm are compared with another algorithm using the graph-based echo labeling method (Jager et al., 2016). It should be noted that the recordings, selected in this study for comparison, had a sampling frequency of 96 kHz, while the authors used sampling frequency of 8 kHz. The authors showed that both methods achieve almost identical average error, but the algorithm proposed by Yu and Kleijn can offer significantly better computational efficiency – even 10^4 shorter working time due to the lower sampling rate used. Comparing the known room geometry parameters and those estimated by neural networks, the authors were able to achieve a minimum average error of 4 cm for the simulated RIR data and 6.5 cm for the recorded RIR data. The smallest error in estimating absorption coefficients was 0.09 (Yu & Kleijn, 2021).

6 Conclusions

In this review, we discussed the use of ANN to generate, augment and evaluate RIRs. From the reviewed studies, we see that GAN, CNN, and MLP networks are used to generate RIRs as well as to estimate room geometry and absorption coefficients. In most studies, the authors have shown that better results can be achieved by applying additional augmentation to the generated RIRs. We can also conclude that the choice of datasets for training, as well as the choice of an acoustic model consisting of certain neural networks, strongly determines the results obtained when performing evaluation tasks of the generated RIRs, such as ASR. The authors of the IR-GAN algorithm, whose acoustic model was based on the TDNN, achieved the best results. They were able to obtain the best WER value when the AIR dataset was selected as the training data.

References

Brandschain, L., Graff, D., Cieri, C., Walker, K., Caruso, C., & Neely, A. (2010). *The Mixer 6 corpus: Resources for cross-channel and text independent speaker recognition* [Conference presentation]. Proceedings of 7th International Conference on Language Resources and Evaluation (LREC), Valletta, Malta.

Cieri, C., Miller, D., & Walker, K. (2004). *The Fisher Corpus: a resource for the next generations of speech-to-text* [Conference presentation]. Proceedings of 4th International Conference on Language Resources and Evaluation (LREC), Lisbon, Portugal.

Donahue, C., McAuley, J., & Puckette, M. (2019). *Adversarial audio synthesis* [Conference presentation]. Proceedings of International Conference on Learning Representations.

Fiscus, J. G. (1997). *A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER)* [Conference presentation]. Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding, Santa Barbara, CA, USA. https://doi.org/10.1109/ASRU.1997.659110

Habets, E. (2010). *Room impulse response generator.* https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator

Harper, M. (2015). The Automatic Speech recogition in Reverberant Environments (ASpIRE) challenge. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)* (pp. 547–554). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/ASRU.2015.7404843

Hsiao, R., Ma, J., Hartmann, W., Karafiat, M., Grezl, F., Burget, L., Szoke, I., Cernocky, J., Watanabe, S., Chen, Z., Mallidi, S. H., Hermansky, H., Tsakalidis, S., & Schwartz, R. (2015). Robust speech recognition in unknown reverberant and noisy conditions. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)* (pp. 533–538). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/ASRU.2015.7404841

Jager, I., Heusdens, R., & Gaubitch, N. D. (2016). Room geometry estimation from acoustic echoes using graph-based echo labeling. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing* (pp. 1–5). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/ICASSP.2016.7471625

Jeub, M., Schäfer, M., & Vary, P. (2009). A binaural room impulse response database for the evaluation of dereverberation algorithms. In *Proceedings of 16th International Conference on Digital Signal Processing* (pp. 1–5). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/ICDSP.2009.5201259

Kinoshita, K., Delcroix, M., Yoshioka, T., Nakatani, T., Habets, E., Haeb-Umbach, R., Leutnant, V., Sehr, A., Kellermann, W., Mass, R., Gannot, S., & Raj, B. (2013). The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech. In *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (pp. 1–4). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/WASPAA.2013.6701894

Ko, T., Peddinti, V., Povey, D., Seltzer, M. L., & Khudanpur, S. (2017). A study on data augmentation of reverberant speech for robust speech recognition. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5220–5224). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/ICASSP.2017.7953152

Nakamura, S., Hiyane, K., Asano, F., Nishiura, T., & Yamada, T. (2000). *Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition* [Conference presentation]. Proceedings of 2nd International Conference on Language Resources and Evaluation (LREC 2000), Athens, Greece.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5206–5210). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/ICASSP.2015.7178964

Park, D. S., Chan, W., Zhang, Y., Chiu, C. C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). SpecAugment: A simple data augmentation method for automatic speech recognition. In *Proceedings of the Annual Conference of the International Speech Communication Association* (pp. 2613–2617). ISCA. https://doi.org/10.21437/Interspeech.2019-2680

Raghuvanshi, N., Narain, R., & Lin, M. C. (2009). Efficient and accurate sound propagation using adaptive rectangular decomposition. *IEEE Transactions on Visualization and Computer Graphics*, *15*(5), 789–801. https://doi.org/10.1109/TVCG.2009.28

Ratnarajah, A., Tang, Z., & Manocha, D. (2020). *IR-GAN: room impulse response generator for speech augmentation.* http://arxiv.org/abs/2010.13219

Snyder, D., Chen, G., & Povey, D. (2015). *MUSAN: a music, speech, and noise corpus.* http://arxiv.org/abs/1510.08484

Szoke, I., Skacel, M., Mosner, L., Paliesek, J., & Cernocky, J. H. (2019). Building and evaluation of a real room impulse response dataset. *IEEE Journal on Selected Topics in Signal Processing*, *13*(4), 863–876. https://doi.org/10.1109/JSTSP.2019.2917582

Tang, Z., Chen, L., Wu, B., Yu, D., & Manocha, D. (2020a). Improving Reverberant Speech Training Using Diffuse Acoustic Simulation. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing* (pp. 6969–6973). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/ICASSP40776.2020.9052932

Tang, Z., Meng, H. Y., & Manocha, D. (2020b). Low-frequency compensated synthetic impulse responses for improved far-field speech recognition. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6974–6978). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/ICASSP40776.2020.9054454

Yu, W., & Kleijn, W. B. (2021). Room acoustical parameter estimation from room impulse responses using deep neural networks. *IEEE/ACM Transactions on Audio Speech and Language Processing*, *29*, 436–447. https://doi.org/10.1109/TASLP.2020.3043115