Falcon Perez, Ricardo; Götz, Georg; Pulkki, Ville

Machine-learning-based estimation of reverberation time using room geometry for room effect rendering

Published in:
Proceedings of the 23rd International Congress on Acoustics : integrating 4th EAA Euroregio 2019 : 9-13 September 2019 in Aachen, Germany

Published: 13/09/2019

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Falcon Perez, R., Götz, G., & Pulkki, V. (2019). Machine-learning-based estimation of reverberation time using room geometry for room effect rendering. In *Proceedings of the 23rd International Congress on Acoustics : integrating 4th EAA Euroregio 2019 : 9-13 September 2019 in Aachen, Germany* (pp. 7258-7265). Deutsche Gesellschaft für Akustik. http://pub.dega-akustik.de/ICA2019/data/articles/000624.pdf

PROCEEDINGS of the
23rd International Congress on Acoustics
9 to 13 September 2019 in Aachen, Germany

Machine-learning-based estimation of reverberation time using room geometry for room effect rendering

Ricardo FALCÓN PÉREZ[1], Georg GÖTZ[2], Ville PULKKI [3]

[1]Aalto University, Finland, ricardo.falconperez@aalto.fi
[2]Aalto University, Finland, georg.gotz@aalto.fi
[3]Aalto University, Finland, Ville.Pulkki@aalto.fi

Abstract

This work presents a machine-learning-based method to estimate the reverberation time of a virtual room for auralization purposes. The models take as input geometric features of the room and output the estimated reverberation time values as function of frequency. The proposed model is trained and evaluated using a novel dataset composed of real-world acoustical measurements of a single room with 832 different configurations of furniture and absorptive materials, for multiple loudspeaker positions. The method achieves a prediction accuracy of approximately 90% for most frequency bands. Furthermore, when comparing against the Sabine and Eyring methods, the proposed approach exhibits a much higher accuracy, especially at low frequencies.

Keywords: Reverberation time, machine learning, room acoustics

1 INTRODUCTION

The reverberation of sound inside rooms has a significant impact on sound quality. In acoustic virtual reality, the response of a room should be reproduced with sufficiently high accuracy to produce plausible rendering of sound [11, 13, 9]. One of the most important parameters when rendering reverberation artificially is the reverberation time (RT60), and the value for it should be estimated from the geometry of the virtual room, for any given source/receiver position and orientation. While complete physical models are typically computationally too demanding, an approximation can be made using simplified mathematical models such as Sabine or Eyring formulas [6], where RT60 values are computed based on few attributes of the room, namely the total volume, total area of surfaces, and the total absorptive area of the surfaces of the room. The methods are based on the assumption of even distribution of absorptive material on the surfaces, which does not often hold in typical domestic environments.

In this project we aim to estimate the RT60 depending on frequency from the geometry of virtual acoustic scenario, where the estimated RT60 values should be accurate enough for plausible rendering of acoustic virtual reality. In other words, the current geometric situation of the environment, source and listener are entered in some form to the model, which estimates the reverberation time values depending on frequency, which are further utilised in a reverberator to create a believable, or plausible, perception of the room effect. For this application the Eyring and Sabine formulas are not accurate enough, and we decided to seek for alternative approaches.

A clear shortcoming in Sabine and Eyring formulas is the assumption of an even distribution of absorptive material on the surfaces. An intuitive method to augment the data is to present the distribution of absorptive area as a spherical image, where each pixel value would describe the amount of absorption in corresponding direction from source or receiver. In this work, the image is formed using the geometry of the icosahedron, where each value presents the value of absorptive area that is visible through a face of the icosahedron.

Furthermore, the RT60 values for each frequency band are then obtained by non-linear regression, performed by an artificial neural network that has been trained with data measured from a real room.

2 DATA

2.1 Public datasets

Machine learning algorithms are powered by data, and the performance of these algorithms on any given task is limited by the quality and quantity of this data. For room acoustics, there are very few large, open, standard datasets that researchers from different institutions use, share and compare results. Instead, most studies collect their own data according to their needs. And although it is common to share this data with the scientific community, most of the time it has limited use beyond the scope of the original study, due to small size, lack of variety, or features types.

A few examples of such datasets are provided in [12, 2, 1, 7]. However, most publicly available datasets lack detailed information about the geometry and absorption coefficients of environments and objects placed inside. Therefore we analyze a novel dataset.

2.2 Dataset

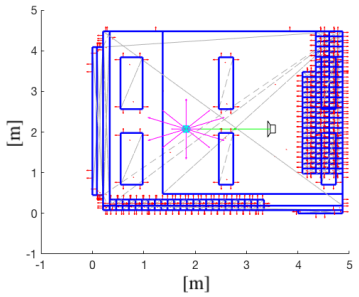
The data consists of real-life acoustical measurements and full 3D geometrical modelling of the room, furniture and other objects placed inside the room, in addition to absorption coefficients for most elements. In total, there are 3332 observations that include audio measurements and other properties. All measurements were made in a single room at the Otaniemi campus of Aalto University during the summer of 2017. The room is a small office, with hard floor and a few soft carpets, light acoustic treatment in the ceiling (in the form of perforated gypsum panels), a large window on one wall, and a single door. The dimensions of the room are $4.4 \times 2.9 \times 4.9$ meters. A total of 4 different fixed loudspeaker positions and 1 microphone position were used. For each observation, the room impulse response (RIR) was obtained by reproducing an exponential sine sweep with a loudspeaker inside the room, and then recording the sound with a microphone. The microphone used is an spherical microphone Eigenmic with 32 capsules. The sampling frequency used is 48 000 Hz.

Even though all the data was obtained from a single room with fixed loudspeaker and microphone positions, variance was introduced by manipulating the objects placed inside the room. The room itself is rectangular and mostly empty, with only a few surfaces that break the flatness of the walls: lamps in the ceiling, and an edge around the window. The objects added to the room included shelves, crates and most noticeably highly absorptive wedges, which are commonly used in anechoic chambers. In the end, the placement and position of these objects in the room creates a unique acoustical environment. In this work we refer to the combination of room, furniture and objects as room configuration. Each configuration has a unique amount, position, and orientation of the aforementioned items. In total, there are 833 different room configurations available.

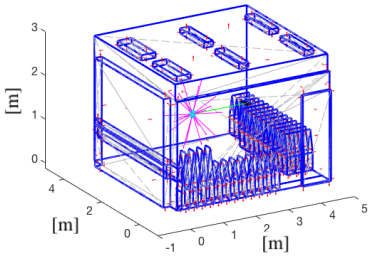
A data observation is collected from each room configuration and loudspeaker position (out of the 4 available), thus there are 3332 observations in total. For each observation, the full 3D model of the room configuration in available. This model contains geometrical information in the form of vertices, normal vectors and face connections for all surfaces of the room as well as the objects inside it. Additionally, for each surface in the model, an approximate absorption coefficient is provided for frequency bands 250, 500, 1000, 2000, 4000 and 8000 Hz. These coefficients where obtained from a table of 6 typical materials such as wood, glass, gypsum, or acoustic foam (for the absorptive wedges). Finally, the observation includes the position and orientation coordinates of the loudspeaker and microphone, as well as the 32 channel RIR. A summary of these raw features is shown in table 1.

Feature	Type	Dimensionality
3D model	complex data structure	N/A
Loudspeaker position and orientation	vector	n x 6
Microphone position and orientation	vector	n x 6
Room impulse response	multichannel sequence	n x 240000 x 32

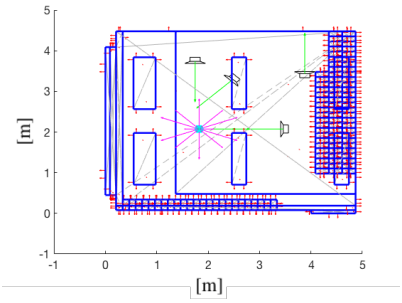
Table 1. Features available for each raw data observation.



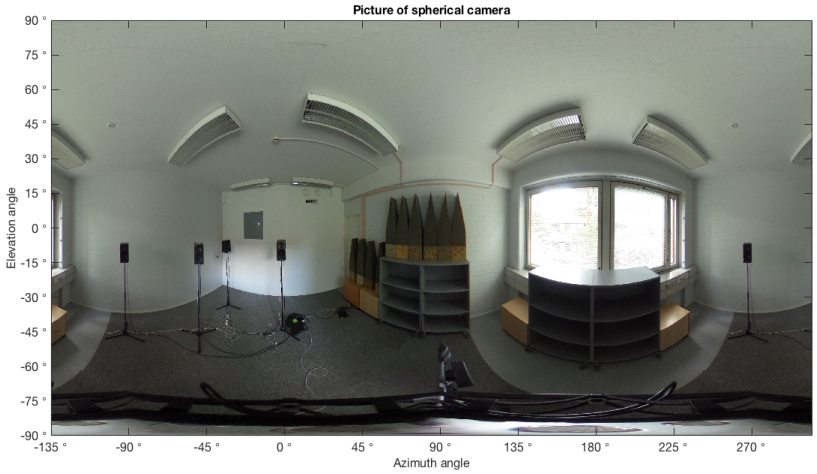
(a) Top view



(b) 3D View



(c) Speaker positions



(d) 2D Spherical photograph

Figure 1. Geometrical data for a typical room configuration, including the general layout of the room along with the placement of fixtures such as lamps, furniture, or absorptive wedges. In the 3D models (a) and (b), the speaker icon represents the loudspeaker position, and the green line marks its orientation. The cyan sphere shows the microphone position, where each magenta line marks the orientation of the 20 faces of the icosahedron volume used to analyze the room. (c) shows the location and orientation of the 4 fixed loudspeaker positions inside the room. The spherical camera (d) shows the photograph of the same room configuration at the microphone position.

A visualization of the raw data for a typical room configuration is shown in figure 1. The 3D model is rendered from two different perspectives, and it includes the room, objects inside, and the loudspeaker and microphone locations. For this configuration, most of the room in empty, except for some shelves and absorptive wedges placed on the floor, along the front and right walls. Subfigures a) and b) show only the position of the loudspeaker used in that particular observation. Subfigure c) shows all available loudspeaker positions. There are 3 positions where the loudspeaker is relatively close and pointing directly towards the microphone, and one where the loudspeaker is far away and pointing towards one of the walls of the room. Lastly, the spherical photograph shows the state of the room during the measurement procedure.

2.3 Data preprocessing

The raw data described earlier is prepared before being fed to the machine learning models. First, the 3D model of the observations have to be transformed into a simple form that can be used as input for the model. In this step, an imaginary icosahedral volume is placed at the microphone position. Each of the 20 faces of the icosahedron points to a unique direction in the room. For each of the 20 directions, ray casting is used to analyze the 3D model to find two sets of values: all surfaces available (ignoring obstacles) and all surfaces visible (taking obstacles into account). Afterwards, the area of each surface is multiplied by the corresponding approximate absorption coefficient to find the absorption area. This results in 2 new features sets for each observation, with dimensionality of 20 x 6 (20 icosahedron faces, and 6 frequency bands). We refer to each set as a spherical map.

Secondly, the RIR are processed and analyzed to extract the RT60 of each observation. To do this, the 32 channels of the RIR are loaded, and using microphone array processing we generate 20 different 3rd order hypercardiod signals, with them pointing to the same 20 directions as the center point of the faces of the icosahedron. The result is a spatially-constrained single channel RIR for each direction. These signals are further combined into a single omnidirectional RIR. Finally, the single channel RIR is used to calculate the value of the RT60. The actual computation is done with the toolbox provided by University of Surrey [4].

Aftwerwards, the preprocessed data is prepared for the machine learning algorithms with additional steps that include:

- **Outliers removal** - All observations where the RT60 is larger than 2 seconds for any frequency band are removed. These values are considered errors produced by the calculation of the reverberation and therefore not relevant to the experiments.
- **Normalization** - For all models, the spherical maps obtained are normalized so that each feature has zero mean, and unit variance.
- **Split** - The data was split into different subsets after removing outliers. The split used 60 % of the data for training, 40% for test. In addition, the data observations where shuffled before splitting, so that each run works on a different sample.
- **Augmentation** - Data augmentation was used to artificially increase the amount of training data and to prevent overfitting. The method used was to create copies of the training observations, and add a small amount of white noise to the input features of the copies. The data was replicated 4 times, so that the training set has 5 times the total number of available observations. The added noise was sampled with Gaussian distribution with mean 0 and standard deviation of 0.2. A cursory analysis of the effect of augmentation showed that increasing the amount of augmentation did not improved the performance significantly, while higher standard deviations for the noise escalated errors. Only training data was augmented, the test data were left unperturbed.

3 METHODOLOGY

Previous studies have used machine learning models to estimate RT60. [10] and [3] both estimate the RT60 using a noisy RIR as input, while [8] estimates the presence or absence of specific absorption materials in a room, using a RIR as input. In this case, we estimate the RT60 using geometric features as input. This is the opposite scenario of [8], where they estimate the basic room geometry (length, width, height) of a room from the RIR.

3.1 Machine Learning Model

The spherical maps described earlier, along with the coordinate vectors for the position and orientation of the microphone and loudspeaker, are used as input features for a machine learning model based on neural

networks. The problem is formulated as a regression task where the targets are the RT60 extracted from the omnidirectional RIR, for each frequency band. The task can be performed either for a single band or multiband case. The total dimensionality of the inputs is $n \times 52$ (40 for the spherical maps of absorption area, and 12 for the location vectors) for the former, and $n \times 252$ (240 for the spherical maps, and 12 for the location vectors) for the latter, where n is the batch size.

The machine learning model is a fully connected (FC) neural network with 2 layers and 30 and 15 hidden units for each layer respectively. The activation function for both layers is relu, and there is no dropout. The training (for this and all models) was done using Adam optimizer [5] and the cost function is the mean square error with L_2 regularization. The minibatch size was set to 256, using learning rate of 5 x 10^{-3} and L_2 coefficient λ of 1 x 10^{-4}. Weights are initialized using random values from Normal distribution with mean of 0 and standard deviation of 0.01, while initial biases are set to 0. The gradient threshold is set to 1, which means that all gradient values larger than 1 will be clipped to 1. Additionally, the validation patience is set to 25 epochs, so that whenever the validation error stops decrease for 25 continuous epochs, training is stopped early. This architecture and the corresponding hyperparameters are summarized in table 2.

Group	Hyperparameter	Value
Model	# layers	2
Model	# hidden units (per layer)	[30, 15]
Model	activation (per layer)	{ relu, relu}
Optimizer	learning rate	10^{-3}
Optimizer	gradient threshold	1
Regularization	L_2	L_2
Regularization	dropout	[0,0]

Table 2. Neural network architecture and Hyperparameters for the machine learning model.

3.2 Performance metrics

The performance metric used in this work is the regression accuracy, which measures how close the estimated value is to the true value. It is computed as

$$\text{accuracy} = (1 - \text{MAPE}), \tag{1}$$

where MAPE is the **mean absolute percentage error**, defined as

$$\text{MAPE} = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{\hat{y}_i - y_i}{y_i + \varepsilon} \right|, \tag{2}$$

where n is the total number of observations, ε is a small number to avoid division by 0, and y_i, $\hat{y}_i$ are the true value and predicted value of the ith observation respectively. The MAPE takes into account the actual value of each observation, giving a scaled measure of error where values close to zero represent good performance. On the other hand, for the regression accuracy, values close to 100 % represent good estimation.

4 RESULTS

In this work, only the results for the single band case are presented, however, there is little difference with the performance of the multiband case. Figure 2 shows the prediction accuracy of the test set for a single band, where the neural network model clearly outperforms the Sabine and Eyring methods, with a mean accuracy

more than 2 times higher. Both Sabine and Eyring methods generally underestimates the RT60, and has a wider range of predictions than the true values. For example, the Sabine method predicts a significant amount of values as low as 0.15 s, while the real data never goes under 0.4 s.

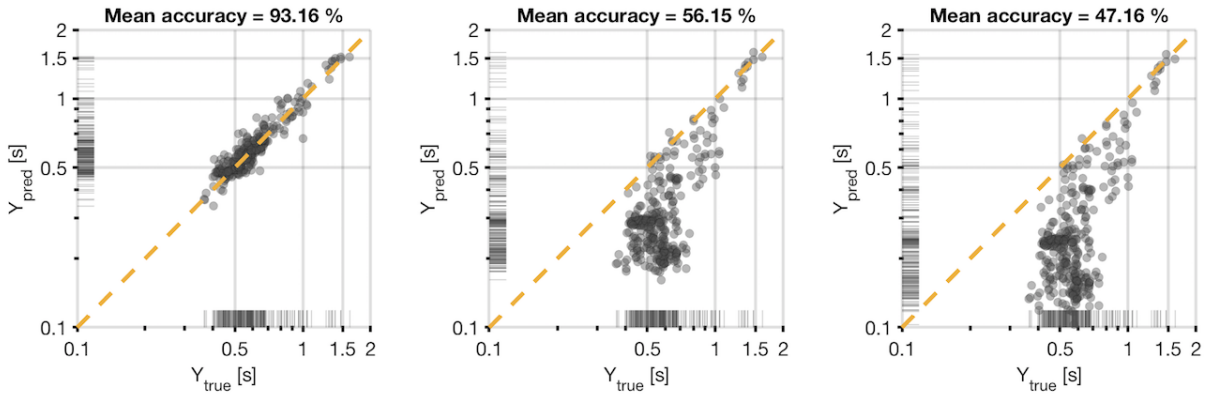


Figure 2. Regression plot of ground truth omnidirectional RT60 values Y_{true} and the estimates Y_{pred} using the baseline neural network model (left), the traditional Sabine (center), and Eyring (right) formulas, at 1000 Hz, with the speaker position in front of the microphone.

Even in cases where the performance of the neural network model is less accurate, there is a clear improvement over the Sabine and Eyring methods. For example, for the frequency band of 250 Hz, and when the speaker position is far away from the microphone, the Sabine and Eyring methods again have a large bias towards underestimating the values, whereas the prediction error for the neural network model is more evenly distributed (figure 3), with predictions that go over and under the true values on similar ratios.

The boxplots for the distribution of the prediction accuracy of all bands and speaker positions are shown in figure 4. For the neural network model, most bands and loudspeaker positions have similar performance, where only the 250 Hz band consistently shows both lower mean and higher variance.

The poor performance of the Sabine and Eyring methods can be explained in part by the known limitations of the methods [6], and also because the absorption coefficients used are rough approximations of the materials present in the room. That said, it is remarkable that the neural network models is able to predict the RT60 with high accuracy even if this absorption data is not exact.

5 CONCLUSIONS

This paper showed a proof of concept for a machine learning method to estimate the reverberation time of a small room using only geometric information and approximate absorption coefficients as input features. All models were trained and evaluated in a dataset that contains real world acoustical measurements and full 3D models of a room and most objects and surfaces inside it.

The estimation was performed using a machine learning model based on a neural network. A baseline model using a shallow fully connected network, achieved good results using geometrical features extracted from the 3d model of the room. The method was validated against the Sabine and Eyring methods, where the neural network approach consistently achieves better performance, even for challenging scenarios.

While these results prove that the proposed approach can work, future work is needed to explore the effectiveness under different conditions. The data needs to be more diverse, with rooms of different geometries and more positions for the microphone and loudspeakers.

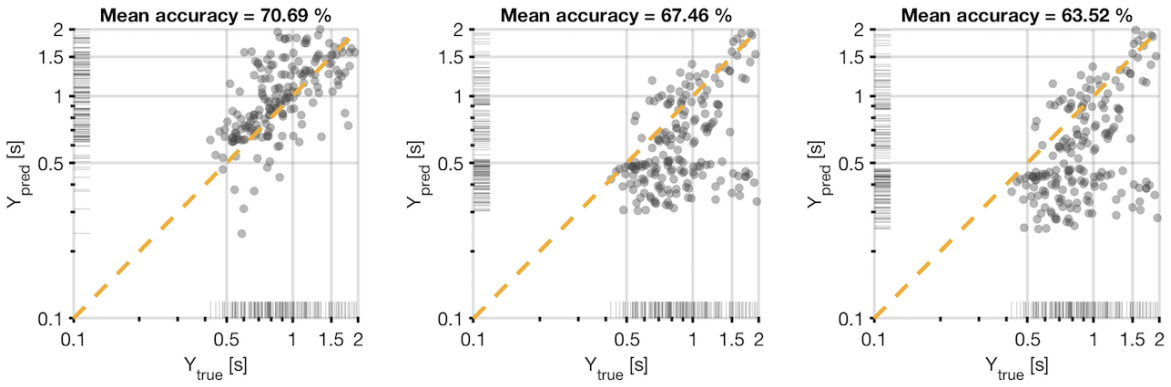


Figure 3. Regression plot for the estimation of omnidirectional T_{60} values using the baseline neural network model (left), the traditional Sabine (center), and Eyring (right) formulas, at 250 Hz, with the loudspeaker position away from the microphone and pointing towards the wall.

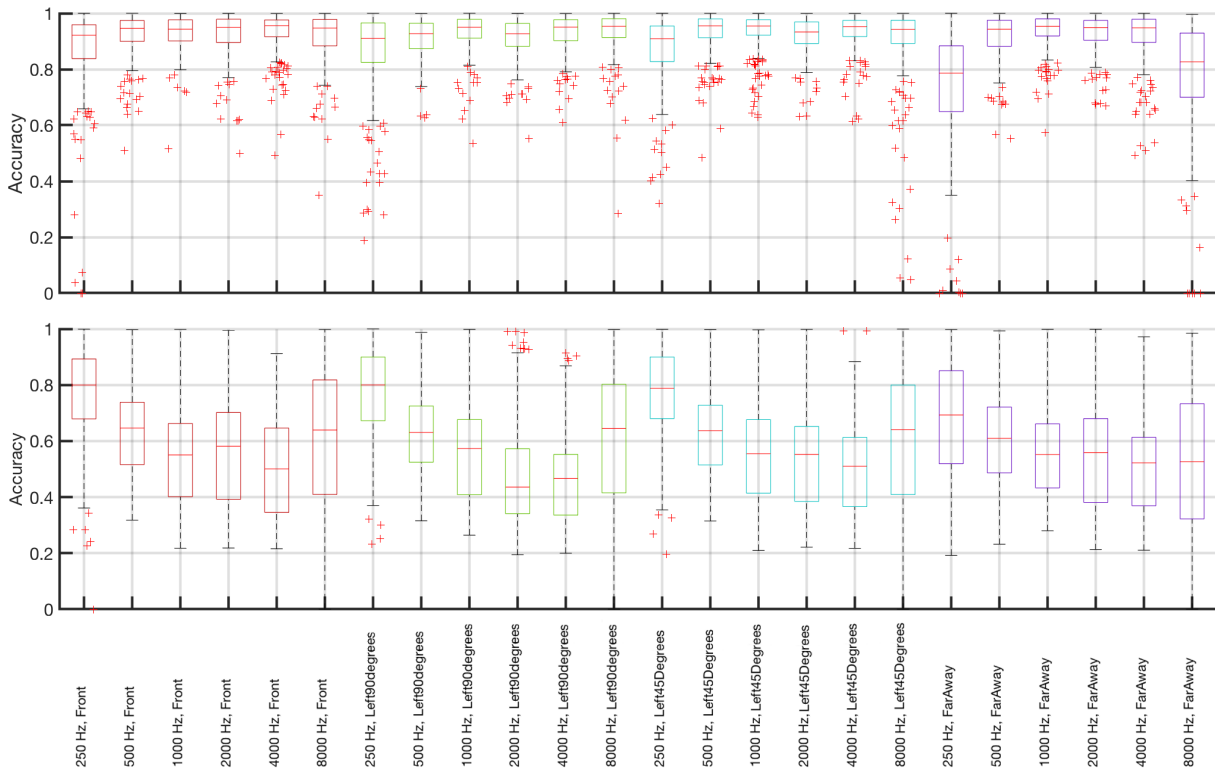


Figure 4. Distribution of the estimation accuracy for the omnidirectional RT60 values using the neural network method (top) and the Sabine method (bottom), at various frequency bands and speaker positions. For better visibility, the results are grouped by speaker position using colors.

ACKNOWLEDGEMENTS

The project has received funding from from the Academy of Finland project no 317341, and from Nordic Sound and Music Computing Network (NordicSMC), project no.86892.

REFERENCES

[1] S. Adavanne, J. Nikunen, A. Politis, and T. Virtanen. TUT Sound Events 2018 - Ambisonic, Reverberant and Real-life Impulse Response Dataset, Apr. 2018.

[2] S. Adavanne, A. Politis, and T. Virtanen. TUT Sound Events 2018 - Ambisonic, Reverberant and Synthetic Impulse Response Dataset, Apr. 2018.

[3] H. Gamper and I. J. Tashev. Blind Reverberation Time Estimation Using a Convolutional Neural Network. In *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 136–140, Tokyo, Sept. 2018. IEEE.

[4] Institute of Sound Recording. IoSR Matlab Toolbox. https://github.com/IoSR-Surrey/MatlabToolbox, accessed on May 2018.

[5] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[6] H. Kuttruff. *Room Acoustics*. Taylor & Francis, 2000.

[7] M. Lovedee-Turner and D. Murphy. Dataset for: Application of Machine Learning for the Spatial Analysis of Binaural Room Impulse Responses, Oct. 2017. Funding was provided by a UK Engineering and Physical Sciences Research Council (EPSRC) Doctoral Training Award, the Department of Electronic Engineering at the University of York.

[8] C. Papayiannis, C. Evers, and P. A. Naylor. Using DNNs to Detect Materials in a Room based on Sound Absorption. *arXiv:1901.05852 [cs, eess]*, Jan. 2019. arXiv: 1901.05852.

[9] V. Pulkki and M. Karjalainen. *Communication acoustics: an introduction to speech, audio and psychoacoustics*. John Wiley & Sons, 2015.

[10] R. Ratnam, D. L. Jones, B. C. Wheeler, W. D. O'Brien, C. R. Lansing, and A. S. Feng. Blind estimation of reverberation time. *The Journal of the Acoustical Society of America*, 114(5):2877–2892, Oct. 2003.

[11] L. Savioja, J. Huopaniemi, T. Lokki, and R. Väänänen. Creating interactive virtual acoustic environments. *J. Audio Eng. Soc.*, 47(9):675–705, 1999.

[12] R. Stewart and M. Sandler. Database of omnidirectional and b-format impulse responses. In *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2010)*, March 2010.

[13] M. Vorländer. *Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality*. Springer, 2007.

[14] W. Yu and W. B. Kleijn. Room Geometry Estimation from Room Impulse Responses using Convolutional Neural Networks. *arXiv:1904.00869 [eess]*, Apr. 2019. arXiv: 1904.00869.