This is an electronic reprint of the original article.
This reprint may differ from the original in pagination and typographic detail.

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

Spherical Maps of Acoustic Properties as Feature Vectors in Machine-Learning-Based Estimation of Acoustic Parameters

Published in:
Journal of the Audio Engineering Society

DOI:
[10.17743/jaes.2021.0011](https://doi.org/10.17743/jaes.2021.0011)

Published: 06/09/2021

Document Version
Peer reviewed version

Please cite the original version:
Falcon Perez, R., Götz, G., & Pulkki, V. (2021). Spherical Maps of Acoustic Properties as Feature Vectors in Machine-Learning-Based Estimation of Acoustic Parameters. *Journal of the Audio Engineering Society*, 69(9), 632-643. <https://doi.org/10.17743/jaes.2021.0011>

This material is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the repository collections is not permitted, except that material may be duplicated by you for your research use or educational purposes in electronic or print form. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone who is not an authorised user.

Spherical Maps of Acoustic Properties as Feature Vectors in Machine-Learning-Based Estimation of Acoustic Parameters *

Ricardo Falcón Pérez, Georg Götz, and Ville Pulkki

(ricardo.falconperez@aalto.fi)

(georg.gotz@aalto.fi)

(ville.pulkki@aalto.fi)

Aalto Acoustics Lab, Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland

This is the author's accepted manuscript of the paper. It is published under the AES "green" open access policy.

The published version of the paper can be accessed from the AES E-Library:

<https://www.aes.org/e-lib/browse.cfm?elib=21460>

This work suggests a method of presenting information about the acoustical and geometric properties of a room as spherical images to a machine-learning algorithm to estimate acoustical parameters of the room. The approach has the advantage that the spatial distribution of the properties can be presented in a generic and potentially compact way to machine learning methods. The estimation of reverberation time T_{60} is used as a proof-of-concept study here. The distribution of absorptive material is presented as a spherical map of feature values, in which each value is formed by calculating the equivalent absorption area visible through the corresponding facet of a polyhedron as seen from the polyhedron's center point. The pixel values are then used as feature vectors, and the real measured T_{60} values of corresponding rooms are used as target data. This work presents the method and trains a set of neural networks with different spherical map resolutions using a dataset composed of real-world acoustical measurements of a single room with 831 different configurations of furniture and absorptive materials. The estimation of reverberation time using the proposed approach exhibits a much higher accuracy compared to simple analytic methods, which proves the validity of the approach.

0 Introduction

The estimation of room acoustic parameters based on geometric data and acoustic properties of materials is needed in many applications, such as in the design of room spaces, the design of audio systems for public address, and in virtual reality audio engines. The parameters typically estimated are, for example, reverberation time T_{60} , direct-to-reverberant ratio, and sound pressure level SPL. The parameters can be estimated using simple mathematic expressions, such as the Sabine formula for T_{60} or the $1/r$ law for SPL generated by direct sound [1]. In most cases, such expressions are based on very few parameter values that describe the situation holistically, such as the volume and the amount of absorptive material in space, or the source-receiver distance. The holistic parameters are usually describing the scenario only coarsely. For example, the parameter quantifying the amount of absorption material in the Sabine equation is a single-valued scalar variable, and consequently, the approach implicitly assumes that the

absorption material is distributed evenly on the surfaces. This is often not the case, and violations to it produce errors in the estimation of T_{60} [2].

On the other hand, several methods exist for estimating room acoustic parameters based on simulations of sound propagation in enclosed spaces [3]. Such methods simulate the propagation of sound in a room from a source to a receiver, thus resulting in an impulse response, from which room parameters can be calculated. Simulations require detailed information of the room geometry and the acoustic properties of the materials. The methods can be computationally very expensive in complex rooms. Nevertheless, the estimated values for room acoustic parameters are more accurate than those obtained from simple equations.

Machine learning is a powerful method for various complex classification and parameter estimation tasks, in which a lot of data is available or can be collected [4]. The estimation of room acoustic parameters is in principle such a task. However, the presentation of the geometry of a room to a machine learning method is not straightforward. Geometric data is usually presented as a text-based list of geometric surface parameters

*To whom correspondence should be addressed: e-mail: ricardo.falconperez@aalto.fi

and their acoustic properties. However, such a representation is generally unintelligible by machine learning methods. The authors are not aware of earlier methods to extract features from such property lists, and a method is proposed for the task in this article.

We are proposing a method to represent the geometric information of the room and its acoustical properties as spherical images from a selected viewpoint in the room, where the image composes of a spherical grid of feature values. Each feature value corresponds to an integral of an acoustic property over the directions seen through a facet of a polyhedron viewed from the center location of the sphere. The values form the feature vector, which is then used in machine learning with neural networks. The acoustical property, of which the feature values are integrated, can be selected differently for different applications. For example, the feature values could be integrated from the distances to surfaces in the room, the amount of absorption visible through a facet, or the roughness of surfaces. The motivation for the approach is that the computation of such maps is an easy task with present graphic processing units, and running a pre-trained neural network with feature value vector is also a computationally light task. The method could thus potentially estimate acoustic parameters in real-time use, for example with a 20 Hz update rate in computer games.

In this work, we present the approach in general and apply it as a case study in the estimation of T_{60} . We use spherical images following platonic solids, in which the feature values correspond to the amount of absorptive material seen through each facet. The presentation of absorption area as a spherical map can be seen as a generalization of a single value absorption area to a 3D representation of the distribution of absorbent material in the room. The method is tested with responses measured from a single rectangular room with 831 different configurations, exhibiting a wide range of T_{60} . We show that, at least in this case, the proposed method using only 20 feature values contains the information which enables more accurate estimation of T_{60} than simple mathematical expressions.

1 Background and related work

In this work, the usability of spherical feature value maps is tested for the estimation of T_{60} of a single room with different furnishings. Therefore, the existing methods for T_{60} estimation are reviewed shortly in this section.

1.1 Simple methods to estimate T_{60}

If the geometry of a room and the absorption properties of its surfaces are known, Sabine's formula can be used to estimate the reverberation time T_{60} [5]

$$T_{60} = 0.161 \frac{V}{S}, \quad (1)$$

where V is the volume of the room [m^3] and S is the total absorption area of the room surfaces computed as

$$S = \sum_{i=1}^{N_{\text{surfs}}} S_i = \sum_{i=1}^{N_{\text{surfs}}} \alpha_i A_i \quad (2)$$

by summing the product of the absorption coefficient α_i and the surface area A_i [m^2] over each surface i .

Sabine's equation has been derived with the assumption that the absorptive material is distributed evenly among all surfaces. Quite often this is not the case, for example, the walls of a room may be made of a material with high absorption, while the floor and ceiling have low absorption. If the floor and ceiling are parallel to each other, a flutter echo may occur between them, and the measured T_{60} value will be much higher than the value estimated with Sabine's equation.

The inclusion of the distribution of absorptive material in such analytic formulations is not straightforward. In [2], several suggested formulations were tested over a large set of T_{60} measurements in a varechoic chamber. Overall, the formulations provided similarly biased results. The original Sabine's equation was performing with similar accuracy as the other systems. However, a formulation suggested by Fitzroy [6], which includes some geometrical aspects of the space, performed slightly better in the comparison. The empirically derived equation assumes a relation within three possible decay rates along the three basic axes in a rectangular room and is expressed by

$$T = \frac{0.161V}{A^2} \sum_j \frac{-A_j}{\ln(1 - \alpha_j)}, \quad (3)$$

where $j = x, y, z$ denotes the current axis, A is the total area of all surfaces, A_j is the total area of the opposite parallel surfaces along the axis, and α_j is the average absorption coefficients for each pair of opposite walls.

1.2 Estimation of T_{60} with acoustic models of rooms

In addition to simple analytic estimation methods, it is possible to use some more complete acoustic field simulation tools to estimate T_{60} . The room geometry is described as a list of surfaces and their acoustic properties, and the parameters of the source(s) and receiver(s) are set by the user. The room impulse response of sound propagating from a source through space to a receiver is first simulated, and the value of T_{60} is estimated from the decay of the response. The simulation is typically conducted using geometrical acoustics modeling methods, such as the ray-based simulation of direct and reflected sounds [7][8].

However, the computational complexity of such ray-tracing methods is still quite high, and they neglect the diffraction effect [3]. Consequently, the ray-based methods may not give correct results for T_{60} in complex geometries. For example, the seating area of an

auditorium is composed of a high number of non-planar surfaces and a large number of diffractive edges. The sound reflected from there can not be explained easily with reflections of sound rays.

A wave-based approach has also been proposed for T_{60} estimation with the promise of higher sound quality [9]. The method is computationally very expensive and is not fitted for real-time estimation of T_{60} in virtual reality applications. The issue of computational load has been solved by running a very large number of simulations of sources and receivers in advance and storing the T_{60} parameters in memory. However, this requires a relatively large allocation of memory, and also that the geometry of the world does not change prominently during run time.

The machine-learning-based approach of estimating T_{60} from the geometry of a room presented in this article can in principle be applied also in virtual reality. It has the promise of very low computational complexity while producing high plausibility or high perceptual quality in synthesized reverberation.

1.3 Presentation of room geometrical features for neural networks in computer graphics

The task of simulating sound propagation inside a room is comparable to the process of scene rendering in the computer graphics domain, where the former models sound waves while the latter models light rays. Although there exists a large number of classical computer vision algorithms based on physical models that operate on geometrical properties, recent developments utilize neural networks for applications such as ray tracing and image manipulations [10]. Depending on the specific task, these solutions require input data presented in an appropriate form, so that it can be processed by the neural networks. In general, the input data can include images, geometrical models with the shape of surfaces, light reflection properties of surfaces, or camera information, among others.

The choice of input data and its form is not trivial, as the data needs to (1) contain the information that is relevant to the task, and (2) be presented in a format that the neural network architecture can exploit. For computer graphics solutions, this information is usually presented in two forms: as rendered images of a scene, or as *meshes*, which are 3D models of objects or scenes [10, 11]. These images and meshes can include data beyond pure geometry, for example about reflectivity, occlusion, or albedo. Other approaches employ a set of layered 2D images together with a graph detailing hierarchical information about sections of the images [12]. The main disadvantage of these features is that they can have very high dimensionality in complex scenes, therefore requiring large neural networks models which can be computationally very expensive.

In this work, we use a method that is similar to the approach of utilizing high-resolution images of the

rooms. We also present the geometrical information to the networks as images, more specifically as 3D spherical images. However, in our application the required number of pixels, or feature values, is relatively low as explained in Section 2

2 Spherical map representation of absorption area distribution in a room

2.1 Fundamental idea

Acoustic parameters of a room are strongly influenced by the room geometry and its surface materials. Consequently, previous approaches for calculating T_{60} attempt to parameterize the room geometry and surface materials. Whereas Sabine's formula given by Eq. (1) solely relies on the volume of the room and its total absorption area, Fitzroy's formula given by Eq. (3) takes the distribution of absorption area along the three basic axes in a rectangular room into account. The spherical maps proposed in this section take this idea one step further by generalizing it to an arbitrarily fine sampling of the absorption area with respect to its directional distribution.

A spherical map is a spherical image that represents a certain parameter of the room with respect to multiple directions as seen from the spherical map origin. Every pixel of that spherical map represents the parameter's value in the corresponding direction. Mathematically, this can be understood as representing the parameter by continuous spherical coordinate angles (ϑ, φ) .

2.2 Calculation of spherical maps for T_{60} estimation

This section will illustrate the calculation and usage of spherical maps for the concrete use case of T_{60} estimation. We propose the usage of limited-resolution spherical maps, which describe the distribution of the absorption area for only a limited number of discrete directions (ϑ_k, φ_k) .

When estimating T_{60} , it is common to have access to a 3D model of the room, in which the room geometry is described as triangulated surfaces. Additionally, at least qualitative information about the materials of the room surfaces is usually available. This information can be used to determine the absorption coefficients α_i and calculate the respective absorption areas S_i for every surface i of the room (c.f. Eq. (2)). The triangulated surfaces and their associated absorption areas are the only required inputs when calculating the proposed spherical maps for T_{60} estimation.

Limiting the resolution to a set of directions implies that the resulting absorption area values $S(\vartheta_k, \varphi_k)$ should be understood as the integral of all absorption areas S_i in the proximity around the directions (ϑ_k, φ_k) . Such a proximity can for example be described by the volume inside a cone or pyramid oriented towards the investigated direction.

We propose to use platonic solids, such as tetrahedron, cube, or icosahedron, to divide space into proximities that cover the entire room as evenly as possible. The directions (ϑ_k, φ_k) are then given by the center points of the platonic solid faces and the absorption area proximities can be determined with the following steps. First, the platonic solid is placed in the center of the room. Then, a pyramid is constructed that has its apex at the center of the platonic solid and its base given by the platonic solid faces. After that, the lateral pyramid edges are elongated away from the apex towards the outside of the room. Finally, the absorption area proximity is defined as the inside of the pyramid.

Figure 1 illustrates the idea of absorption area proximities with an exemplary room configuration consisting of two shelves, a carpet, a door, six ceiling lights, and a window with a windowsill. The shelves, the door, and two of the ceiling lights are entirely located inside the shown pyramid. Therefore, their entire absorption area is added to the absorption area described by the corresponding cube face. In contrast, only approximately half of the carpet is located inside the pyramid. For the carpet, it is therefore necessary to determine which portions of its absorption area should contribute to the investigated proximity.

The following steps can be executed to determine which portion of a surface is located inside a proximity. A room surface is located entirely inside a proximity if all of its vertices are located inside of it. Otherwise, intersections between the pyramid's lateral faces and the investigated surface have to be calculated. For triangulated room surfaces, this can efficiently be done by computing triangle-triangle-intersections. The triangle-triangle-intersections will return a set of new vertices that lie both on the investigated room surface and either one of the pyramid's lateral faces. By taking the resulting vertices and the vertices inside the pyramid together, a polygon can be spanned that effectively

Figure 1: A cube is placed approximately in the middle of a room to calculate spherical maps of absorption area. A pyramid that is used to determine the absorption area proximities is depicted in orange.

delimits the portion of the room surface inside the pyramid. The area of this polygon is then taken as the area of the room surface located inside the investigated proximity. Figure 2 clarifies this procedure for the proximity corresponding to the blue icosahedron face. Whilst the small room surface, which is placed closer to the icosahedron, is located entirely inside the pyramid, the other room surface has one vertex sticking out. Triangle-triangle-intersections between the pyramid's lateral faces and this room surface result in the two new vertices depicted in orange. The portion of the room surface that contributes to the absorption area value of the proximity is therefore described by the green polygon connecting the green and orange dots.

2.3 Exemplary spherical maps with different resolutions

Spherical maps of acoustic features can be calculated by utilizing platonic solids. Figure 3 depicts exemplary spherical maps of absorption area of the room configuration illustrated in Figure 1. The maps were calculated by using either a cube or an icosahedron. The orientation of the platonic solid was defined, such that the direction $(\vartheta = 0^\circ, \varphi = 0^\circ)$ corresponds to the direction of the positive x -axis, with the azimuth angle φ increasing in the counter-clockwise direction and the elevation angle ϑ increasing upwards.

Both spherical maps indicate increased absorption areas towards the front of the platonic solid due to the two shelves. However, only the icosahedral map reveals that the absorption area is located closer to the floor and not at the center of the wall or the ceiling. Both

Figure 2: The absorption area in the direction corresponding to the blue icosahedron surface is calculated by determining which room surface portions are located inside the direction's proximity. Inside portions are depicted in green, whilst outside portions are depicted in red. The vertices obtained from the triangle-triangle-intersection between the lateral faces of the pyramid and the room surface are indicated by orange dots.

maps show absorption on the floor due to the carpet. Although the cube map has a distinct face that is supposed to include absorption below the cube, some portions of the carpet seem to be covered by the lateral cube faces as well. For instance, the left cube face ($\varphi = 90^\circ$) shows some absorption, although the room consists only of a concrete wall in that direction. In contrast, the icosahedron map captures the placement of the carpet better and displays less absorption for the previously discussed concrete wall.

Figure 3: Spherical maps of absorption area for the exemplary use case of T_{60} estimation. The maps are calculated for the room configuration depicted in Figure 1. The basis for calculating the maps was a cube and icosahedron respectively.

3 Data

In the following sections, we will use a machine-learning model to estimate T_{60} values from an arbitrary absorption area distribution, represented by spherical maps of different resolutions. In general, the quantity

and quality of the data used for training constrain the performance of machine-learning algorithms. For our purpose, we require a data set consisting of room-impulse response (RIR) measurements and 3D models of the corresponding room configurations. Furthermore, the 3D models should include information about surface materials. Although there exists already a variety of related room acoustic datasets, such as [13] [14] [15] [16], none of them provided all of the information required for our purpose. Therefore, a new dataset was collected.

3.1 Dataset description

The dataset consists of RIR measurements conducted with an em32 Eigenmike microphone array. Full 3D models of the room and its furniture are included for all measurements. Furthermore, the 3D models contain additional meta-information, such as surface materials and their corresponding absorption coefficients according to the tabulated values from [8].

All measurements were conducted in a single room – namely, a small shoe-box-shaped office of $4.9\text{m} \times 4.4\text{m} \times 2.9\text{m}$, having a vinyl-coated concrete floor and concrete walls, a large window on one wall, and a single door. Variance was introduced by adding or removing furniture, and by moving the furniture objects to various positions around the room. These positions were mostly random, although the majority of them roughly resemble typical scenarios of real spaces such as offices or living rooms. Perhaps the most interesting situations include furniture positions where the source is occluded from the receiver by highly absorptive material. Table 1 summarizes the 4 different types of furniture objects used in the measurements together with their available quantities. The position and orientation of these objects within the room create a unique acoustical environment. In this work, we refer to the combination of room and furniture objects as room configuration. Each room configuration is unique concerning the quantity and positioning of the aforementioned objects. A distinct 3D model is available for every room configuration. In total, the dataset consists of 831 different room configurations.

For every room configuration, RIRs were measured with exponential sine sweeps, utilizing a setup that consisted of an Eigenmike microphone array and four Genelec loudspeakers. Their positions remained static throughout the entire dataset acquisition. In total, the dataset contains 3324 RIR measurements (831 room

Table 1: Furniture objects used during the measurements of the new dataset.

Furniture object type	Description	Available quantity
Shelf	Rollable bookshelf	4
Drawer	Rollable drawer, the type which is used next to desk	4
Absorption wedge	Absorption wedges, the type which is used in anechoic chambers	50
Carpet tiles	approx. $0.5\text{m} \times 0.5\text{m}$	56

Figure 4: Geometrical data for a typical room configuration of the dataset. In the 3D model depicted in (a), the blue speaker icons represent the 4 fixed loudspeaker positions, and the red arrows indicate their orientation. The orange icosahedron shows the microphone position. The spherical photograph (b) shows the room at the time of the measurement from the microphone position and includes an overlay of the resulting cubical map (dark overlay = high absorption). The spherical photograph of (b) is not required for the calculation of the spherical map. It is only shown in the Figure to demonstrate the nature of absorption area spherical maps.

configurations $\times$ 4 loudspeaker positions), where each RIR should be understood as a 32 channel Eigenmike recording. Figure 4 shows a spherical photograph of the room in a typical room configuration, as well as the positions of the loudspeakers and the microphone.

4 Methodology

Figure 5: Block diagram of the proposed approach for estimation of room acoustic parameter T_{60} using spherical images representing the distribution of absorption on the surfaces of the room. This approach includes a feature extraction process that generates spherical maps, and a neural network that performs the estimation.

As shown in Figure 5 our approach consists of two components: (1) a feature extraction process that uses the raw 3D model of the room geometry and the approximate absorption coefficients to generate spherical maps of absorption area, and (2) a machine-learning model based on a shallow fully connected neural network. The network estimates the T_{60} for all frequency bands, using the spherical maps in addition to the position and orientation of the loudspeaker and microphone.

4.1 Feature extraction

The dataset outlined in Section 3 consists of RIR measurements and 3D models that describe the various room configurations. To make this data usable for a machine-learning model, it is first required to adequately pre-process it. Two different steps are required. Firstly, the absorption area distribution within the room has to be represented in a consistent format, where we utilize the spherical maps as introduced in Section 2. Secondly, the ground-truth T_{60} values have to be extracted from the RIR measurements, so that they can be used as target values to train the machine-learning model.

In our implementation, we represent the absorption area distribution of a room configuration by spherical maps of different resolutions. As elaborated in Section 2.2 platonic solids can be used to calculate spherical maps of absorption areas. In this paper, we calculated spherical maps based on the tetrahedron, the cube, and the icosahedron. Furthermore, we investigated different orientations of those spherical maps. This means, that

Table 2: Different spherical maps used in our implementation.

Description	Name	Orientation
room division based on tetrahedron	tetra_x	no tilt
	tetra_rot	45° tilt in azimuth
	tetra_dbrot	45° tilt in azimuth and elevation
room division based on cube	cube_x	no tilt
	cube_rot	45° tilt in azimuth
	cube_dbrot	45° tilt in azimuth and elevation
room division based on icosahedron	ico_x	no tilt
	ico_rot	45° tilt in azimuth
	ico_dbrot	45° tilt in azimuth and elevation
mean value of the cubical map	mean_x	no tilt
	mean_rot	45° tilt in azimuth
	mean_dbrot	45° tilt in azimuth and elevation
sum of each half of cubical map	half-half_x	no tilt
	half-half_rot	45° tilt in azimuth
	half-half_dbrot	45° tilt in azimuth and elevation

additional spherical maps were calculated, for which the underlying platonic solids were rotated. Such a rotation effectively changes the room surfaces that are „seen“ through the different platonic solid faces. This was motivated by the hypothesis that a spherical map, whose underlying cube is perfectly aligned with the walls of the room, will perform considerably better than a spherical map with a cube that is tilted by a certain angle.

Moreover, we included two simplified versions of the maps to use as references. For the first one, we take the mean value of the 20 faces of the icosahedral map, effectively reducing the absorption area information about the room to a single number. The second one splits the available absorption area into two values, computing the sum of each half of the room from a cubical spherical map. We refer to these simplified maps as *mean*, and *half-half*. The different spherical maps that were used in our investigation are summarized in Table 2.

To extract the ground-truth T_{60} values, the measured RIRs are processed and analyzed. To do this, we use microphone array processing to convert the 32-channel RIR measurements to 3rd order Ambisonic streams. Finally, we use the omnidirectional channel of the Ambisonic stream to calculate the T_{60} values for the six different frequency bands with center frequencies of 250 Hz, 500 Hz, 1000 Hz, 2000 Hz, 4000 Hz and 8000 Hz. We use the IoSR toolbox provided by the University of Surrey [17] to compute the T_{60} values from the omnidirectional RIRs by fitting a linear slope to the Schroeder decay curve.

Figure 6 shows the distribution of the ground-truth T_{60} of the dataset and the relationship between the ground-truth T_{60} and the extracted total absorption area values. Although only a single room is used, the variety of room configurations generates significant diversity in the parameters. In particular, the T_{60} values range from 0.3 to about 2 seconds. Additionally, there is a large range of absorption values that correspond to a relatively small range of T_{60} , meaning that the mapping is non-linear.

4.2 Data pre-processing

Once the features and target T_{60} values have been extracted, the data is pre-processed to prepare it for the machine-learning models. This process consists mainly of the following steps:

- **Outlier removal** - All observations, for which the T_{60} values presented errors produced by the calculation of the reverberation time were removed. These errors include cases, in which the noise floor was too high to successfully fit a linear slope to the decay curve.
- **Dataset split** - The data is split into three subsets after removing outliers. In typical machine-learning applications, it is important to ensure that the training and testing sets do not include observations that are closely related, as this will misrepresent the generalization error. To avoid this, we split by room configurations, such that each room configuration (with its 4 loudspeaker positions) belong to a single subset. The split proportions are set as follows: 60 % of the data are used for training, 10% for validation, and 30% for testing. The validation subset is used only for early stopping of the neural network models. Consequently, it is ignored for the baselines. In addition, the data observations were shuffled randomly before splitting. Lastly, to get a better estimate of the performance, this process of shuffling and splitting was repeated 100 times. For each random split, we train and evaluate all the models and the performance as statistics over the splits.
- **Normalization** - The input data is normalized to ensure that each feature has zero mean and unit variance.
- **Augmentation** - Given the limited amount of available data, we used data augmentation techniques to artificially increase the amount of training data and to prevent overfitting. For this purpose, we replicated the training data and added a small amount of white noise to the input features of the copies. The added noise was sampled from a Gaussian distribution with a mean of 0 and a standard deviation of 0.2. The target T_{60} values were not perturbed. The data was replicated 4 times, such that the training set has 5 times the total number of available observations. A cursory analysis of the effect of augmentation showed that increasing the amount of augmentation beyond this point did not improve the performance significantly, while higher standard deviations for the noise escalated errors.

4.3 Baseline Models

We use 3 different models as baseline methods to compare our approach. Firstly, we use a simple model that computes the mean of the T_{60} values for each band of the training subset and uses these values as predictions, regardless of the input features. This model

Figure 6: (Left) Distribution of the average T_{60} across all frequency bands. (Right) Joint distribution of the total absorption area and average T_{60} for each room configuration.

effectively ignores any absorption data and estimates the same T_{60} for every single data observation. For the other two models, we use the Sabine and Fitzroy formula, as described in Equation (1) and (3), respectively. The required S , A , and A_j values are obtained from cubical maps.

4.4 Machine-Learning Model

For the machine-learning model, we formulate the T_{60} estimation problem as a regression task. The target values for the regression task are the T_{60} values at six frequency bands that were obtained from the RIR measurements as described in Section 4.1. The inputs for the model are the previously described spherical maps, as well as the coordinate vectors describing the position and orientation of the source and microphone. For the estimator, we use a shallow neural network with mostly fully connected (FC) layers. The specific architecture used in the experiments is shown in Table 3. The model consists of 3 hidden layers with 50 units in each layer. The activation function for both layers is rectified linear units (ReLU) [18], and there are no dropout or normalization layers.

The training procedure uses Adam optimizer, where the cost function is the means square error with L_2 regularization. The minibatch size was set to 256, using a learning rate of $5 \cdot 10^{-3}$ and L_2 coefficient λ of $1 \cdot 10^{-4}$. Weights are initialized using random values from a Normal distribution with a mean of 0 and a 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 decreasing for 25 consecutive epochs, training is stopped early. In practice, we found that a vast range of neural network architectures and configurations perform similarly, therefore the exact training strategy and

Table 3: Architecture for the neural network model. In this example, m corresponds to the number of surfaces in the spherical map. The number of parameters is shown for a typical model using the icosahedral spherical map.

Neural Network Model			
Input	<i>spherical map</i>	$m \times 6$	
	Mic Loc + Ori	6×1	
	Spkr Loc + Ori	6×1	
Layer	Width	Output	# Params
FC input + ReLU	50	1×50	6,350
FC + ReLU	50	1×50	2,550
FC + ReLU	50	1×50	2,550
FC + ReLU	50	1×50	2,550
FC output	6	1×6	306
Total params			14,306

hyperparameter selection are included here only as a reference.

4.5 Performance metrics

As there is no single, standard metric to evaluate the performance of T_{60} estimation methods, we use a set of multiple evaluation metrics. Each metric presents a different perspective of the performance and provides additional insights on the methods. The metrics include:

- **Root Mean Square Error (MSE)**

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2} \quad (4)$$

- **Mean Absolute Error (MAE)**

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |\hat{y}_i - y_i| \quad (5)$$

- **Mean Absolute Relative Error (MARE)** In practice, the absolute estimation errors can be misleading, as they ignore the true magnitude of the estimated values. Therefore, the MARE computes the error relative to the true values, it is computed as:

$$\text{MARE} = \frac{100\%}{N} \sum_{i=1}^N \left| \frac{\hat{y}_i - y_i}{y_i + \epsilon} \right|, \quad (6)$$

where y_i and $\hat{y}_i$ are the true value and the predicted value of the i th observation, respectively, N is the total number of observations, and ϵ is a small number to avoid division by 0. MARE values close to zero represent good performance, as they indicate no error.

- **Empirical Cumulative Distribution Function (ECDF)** The aforementioned metrics compute the mean value across all the examples in the dataset, thus obtaining a single number to summarize the performance. However, this process fails to capture the full distribution of errors, which might be important for some applications. For example, in cases in which it is important to guarantee that the estimation error never exceeds a specific threshold. To counter this issue, we use the empirical cumulative distribution function to approximate the distribution of the absolute relative errors.
- **MARE@90p** As a summary of the ECDFs, we look at the relative error at the 90th percentile of the data. This means that 90 % of the data will have at most this particular MARE.

5 Results

5.1 Reverberation time estimation

In general, the neural network models trained with the spherical maps have much better performance than the baselines. Figure 7 shows the prediction accuracy of the test set of a single random split of the dataset, comparing the baselines using the Sabine and Fitzroy formulas and a neural network model trained with an icosahedral spherical map. The neural network model outperforms both Sabine and Fitzroy, having a much lower estimation error on average and a more even bias. Both Sabine and Fitzroy generally underestimate T_{60} , while also having a wider range of predictions than the true values. For example, both formulas predict a significant amount of values lower than 0.15 s, while the ground truth T_{60} values never drop below 0.4 s.

This behavior is present along most frequency bands. However, the neural network model is less reliable at low frequencies than at high frequencies. It is possible that, for both the neural network models and the baselines, the higher error rate at low frequencies is affected

by errors in the data, such as those caused by a faulty curve fitting during the T_{60} extraction from the measurements. That said, even for this challenging scenario, our approach performs much better than the baselines. Finally, the errors are more evenly distributed for the neural network model, where prediction errors both exceed and underestimate the true values with equal proportions.

The poor performance of the Sabine and Fitzroy baselines can be explained by two factors. One is the known limitations of the methods [1], and the other is the absorption coefficients used as textbook approximations of the materials present in the room. In the present study, given that the absorption data is not exact, it is remarkable that the proposed approach can predict the T_{60} with low relative error for the most part. This finding is most likely because the model learns to compensate for any bias that the approximate coefficients introduce, as long as the same coefficient values are assigned to similar surfaces consistently on the full dataset. In practice, this is a strong advantage, because collecting accurate absorption information is challenging and often not possible.

5.2 Spherical map resolution

An increased resolution of the spherical maps shows a difference in performance. In Figure 8 the ECDF curves show that there is a tendency that maps with higher resolution have lower errors, although not in all cases. The differences of accuracy obtained with different spherical maps are not very large. Although the icosahedral map performs indeed the best, it is surprising that the tetrahedral map exhibits slightly better performance than the cubical map. Nonetheless, this outcome does not occur when the maps are rotated. Therefore, this can be considered a coincidence due to some conditions in the room and not a general trend of the spherical maps.

Furthermore, all models based on spherical maps perform much better than the baselines. Interestingly, the baseline with the mean of the training values performs better than both Sabine and Fitzroy. This does not lead to conclusions about the performance of Sabine and Fitzroy in general, because the limited dataset, which only considers a single room, might not show the full picture.

The neural network models show the same significant performance across all metrics. Table 4 shows the performance summary of 100 different random splits for all models and metrics. The icosahedral maps perform consistently the best, having lower errors and also lower variance. The results shown include only the spherical maps without rotations, however, the other rotations perform similarly and there are no meaningful differences between them.

Figure 7: Regression plots for the estimation of T_{60} values of the test subset of a single random split, for a neural network model trained with the icosahedral spherical map (left), the Sabine baseline (middle), and the Fitzroy baseline (right). Best viewed in color.

Figure 8: (top) Empirical cumulative distribution functions (ECDFs) of relative absolute errors for the prediction of T_{60} for baseline methods and neural network models trained with spherical maps of different resolutions. (bottom) ECDFs of the same data as (top), but zoomed in to highlight the differences between the spherical map resolutions. The distributions show the test set errors computed over 100 different random splits.

6 Discussion

The results suggest that spherical maps of absorption area convey the information of acoustical condi-

tions in a room to a neural network, at least for the given dataset of a single room with a large number of different furnishings. We see this already as a good result, however, we acknowledge that for more general applications, the approach has to be tested with larger amounts of data gathered from various room conditions. In the tested cases, the performance of the system did not vary much with different resolutions of spherical images when the number of faces was four or larger. It is possible that in some geometries higher spherical-image resolutions may be needed.

Note also, that in the investigated dataset, the volume of the room did not change. When the network is trained with different rooms, also that information has to be conveyed to the network, either as a single scalar value or as a spherical map carrying the value of the room volumes inside each pyramid defined by the map.

In the current study, the information of source and receiver locations were also provided to the network. This may seem unnecessary, since ideally, T_{60} does not depend on the locations. However, it was noticed that the T_{60} values, which were used in the training of the networks, varied slightly. This variation was possibly caused by the presence of room modes and reflections. When the source location was provided as training data, somewhat better training accuracy was achieved. Clearly, the networks learned to estimate the effect of source-receiver positions on the values of T_{60} . It is an open question if the network will be able to estimate such effects with arbitrary source-receiver positions.

7 Conclusions

This article proposes a feature extraction method that can be used for machine learning applications, in which the goal is to estimate an acoustic parameter of a room from the geometry and acoustical properties

Table 4: Performance metrics for baseline and neural network models trained with different spherical maps. The values show the mean and standard deviation of each metric evaluated on 100 random splits, averaging across the 6 frequency bands. RMSE and MAE are in seconds, Relative MAE and Relative MAE@90% are in percent. Lower errors indicate better performance.

Model	Map Type	Dimensions	RMSE	MAE	Rel. MAE	Rel. MAE@90%
Mean value (Baseline)	-	1	0.3910 ± 0.0087	0.1529 ± 0.0068	22.60 ± 0.63	44.72 ± 1.25
Sabine (Baseline)	-	1	0.4938 ± 0.0049	0.2460 ± 0.0048	40.09 ± 0.73	67.17 ± 0.93
Fitzroy (Baseline)	-	1	0.4176 ± 0.0070	0.1744 ± 0.0059	35.49 ± 0.67	64.84 ± 1.03
Neural Network	Mean	1	0.2678 ± 0.0095	0.0718 ± 0.0052	10.70 ± 0.62	22.11 ± 1.45
Neural Network	Half-Half	2	0.2529 ± 0.0082	0.0640 ± 0.0042	09.52 ± 0.61	19.78 ± 1.39
Neural Network	Tetra	4	0.2384 ± 0.0074	0.0569 ± 0.0036	08.29 ± 0.46	17.51 ± 1.03
Neural Network	Cube	6	0.2419 ± 0.0078	0.0586 ± 0.0038	08.65 ± 0.55	18.31 ± 1.35
Neural Network	Icosahedron	20	0.2381 ± 0.0094	0.0568 ± 0.0045	08.15 ± 0.55	16.95 ± 1.35

of the room. The method presents the distribution of the geometric and/or acoustic-parameter properties of room as spherical image(s). A systematic approach is to use well-defined polyhedra, such as platonic solids, that lead to parameter distributions with even divisions of the space. The resulting spherical images consist of feature values, where each value is computed by integrating the acoustic property, such as absorption area, inside the volume of a pyramid defined by the center of the polyhedron and the directions of the vertices of the corresponding facet.

The feature extraction method is tested using a dataset measured from a real room with 831 different distributions of absorptive material. The tests show that the representation of absorption area distribution as a relatively simple spherical image contains information on the acoustic conditions of a room in such a form that machine-learning algorithms can estimate the reverberation time with better accuracy than with simple analytic expressions.

The resolutions of the spherical maps were half-hemispheric, tetrahedral, cubic, and icosahedral maps. It is shown, that increasing the resolution improves the learning result, although the improvement was small with maps with more than 4-6 faces, at least with the limited dataset used in the study. Furthermore, the model was validated against the simple analytic formulas proposed by Sabine and Fitzroy, where the machine-learning approach consistently achieves better performance, even for challenging scenarios.

ACKNOWLEDGEMENTS

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

References

- [1] H. Kuttruff, *Room Acoustics* (Taylor & Francis) (2000).
- [2] K. Prawda, S. J. Schlecht, V. Välimäki, “Evaluation of Reverberation Time Models with Variable Acoustics,” presented at the *Sound and Music Computing Conference* (2020).
- [3] L. Savioja, U. P. Svensson, “Overview of geometrical room acoustic modeling techniques,” *The Journal of the Acoustical Society of America*, vol. 138, no. 2, pp. 708–730 (2015).
- [4] E. Alpaydin, *Introduction to machine learning* (MIT press) (2020).
- [5] W. C. Sabine, *Collected papers on acoustics* (Harvard University Press Cambridge, MA) (1927).
- [6] D. Fitzroy, “Reverberation formula which seems to be more accurate with nonuniform distribution of absorption,” *The Journal of the Acoustical Society of America*, vol. 31, no. 7, pp. 893–897 (1959).
- [7] L. Savioja, J. Huopaniemi, T. Lokki, R. Väänänen, “Creating interactive virtual acoustic environments,” *J. Audio Eng. Soc.*, vol. 47, no. 9, pp. 675–705 (1999).
- [8] M. Vorländer, *Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality* (Springer) (2007).
- [9] N. Raghuvanshi, J. Snyder, “Parametric wave field coding for precomputed sound propagation,” *ACM Transactions on Graphics (TOG)*, vol. 33, no. 4, pp. 1–11 (2014).
- [10] A. Tewari, O. Fried, J. Thies, V. Sitzmann, S. Lombardi, K. Sunkavalli, R. Martin-Brualla, T. Simon, J. Saragih, M. Nießner, R. Pandey, S. Fanello, G. Wetzstein, J.-Y. Zhu, C. Theobalt, M. Agrawala, E. Shechtman, D. Goldman, M. Zollhofer, “State of the Art on Neural Rendering,” *Computer Graphics Forum*, vol. 39 (2020).
- [11] S. Lombardi, J. Saragih, T. Simon, Y. Sheikh, “Deep Appearance Models for Face Rendering,” *ACM Trans. Graph.*, vol. 37, no. 4 (2018 Jul.), doi:10.1145/3197517.3201401, URL <https://doi.org/10.1145/3197517.3201401>
- [12] S. Chaillou, *ArchiGAN: Artificial Intelligence x Architecture*, pp. 117–127 (Springer Singapore, Singapore) (2020), doi:10.1007/978-981-15-6568-7_8.

- [13] R. Stewart, M. Sandler, "Database of Omnidirectional and B-Format Impulse Responses," presented at the *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2010)* (2010 March).
- [14] S. Adavanne, A. Politis, T. Virtanen, "TUT Sound Events 2018 - Ambisonic, Reverberant and Synthetic Impulse Response Dataset," (2018 Apr.), doi:10.5281/zenodo.1237707, URL <https://doi.org/10.5281/zenodo.1237707>.
- [15] S. Adavanne, J. Nikunen, A. Politis, T. Virtanen, "TUT Sound Events 2018 - Ambisonic, Reverberant and Real-life Impulse Response Dataset," (2018 Apr.), doi:10.5281/zenodo.1237793, URL <https://doi.org/10.5281/zenodo.1237793>.
- [16] M. Lovedee-Turner, D. Murphy, "Dataset for: Application of Machine Learning for the Spatial Analysis of Binaural Room Impulse Responses," (2017 Oct.), doi:10.5281/zenodo.1038021, URL <https://doi.org/10.5281/zenodo.1038021>, 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.
- [17] Institute of Sound Recording, "IoSR Matlab Toolbox," <https://github.com/IoSR-Surrey/MatlabToolbox> (accessed on May 2018), URL <https://github.com/IoSR-Surrey/MatlabToolbox>.
- [18] V. Nair, G. E. Hinton, "Rectified Linear Units Improve Restricted Boltzmann Machines." presented at the J. Fürnkranz, T. Joachims (Eds.), *ICML*, pp. 807–814 (2010), URL <http://dblp.uni-trier.de/db/conf/icml/icml2010.html#NairH10>

THE AUTHORS

Ricardo Falcón-Pérez

Georg Götz

Ville Pulkki

Ricardo Falcón Pérez is a doctoral candidate at the Acoustics Lab, Department of Signal Processing and Acoustics, of Aalto University, Finland, where he also received his M. Sc (2018). His areas of interest include deep learning, music information retrieval, and spatial audio. His research focuses on representation learning for room acoustics, as well as methods for sound event detection and localization that are robust to acoustically challenging environments. On his free time he enjoys producing electronic music, and cityscape photography.

Georg Götz is a doctoral candidate at the Acoustics Lab, Department of Signal Processing and Acoustics, of Aalto University, Finland. He received his B. Sc. (2016) and M. Sc. (2019) degrees in Media Technology from Ilmenau University of Technology, specializing in spatial audio and psychoacoustics. He was the deputy chair of the AES Student Section Ilmenau from 2013

to 2016. The research of his doctoral thesis is about machine-learning-based virtual acoustics rendering. His other research interests include room acoustics, psychoacoustics, and VR/AR technology.

Ville Pulkki is a professor in the Department of Signal Processing and Acoustics at Aalto University, Helsinki, Finland. He has been working in the field of spatial audio for over 20 years. He developed the vector-base amplitude panning (VBAP) method in his Ph.D. (2001) and directional audio coding after the Ph.D. with his research group. He also has contributions in perception of spatial sound, laser-based measurement of room responses, and binaural auditory models. He has received the Samuel L. Warner Memorial Medal Award from SMPTE and the AES Silver Medal Award. He enjoys being with his family, building his summer house, and performing in musical ensembles.