Audio Engineering Society

Conference Paper

Presented at the 2022 International Conference on
Audio for Virtual and Augmented Reality
2022 August 15–17, Redmond, WA, USA

Towards the prediction of perceived room acoustical similarity

Hannes Helmholz[1,2], Ishwarya Ananthabhotla[1], Paul T. Calamia[1], and Sebastià V. Amengual Garí[1]

[1]*Reality Labs Research at Meta, 9845 Willows Road, 98052 Redmond, USA*
[2]*Chalmers University of Technology, Division of Applied Acoustics, Gothenburg, Sweden*

Correspondence should be addressed to Sebastià V. Amengual Garí (samengual@fb.com)

ABSTRACT

Understanding perceived room acoustical similarity is crucial to generating perceptually optimized audio rendering algorithms that maximize the perceived quality while minimizing the computational cost. In this paper we present a perceptual study in which listeners compare dynamic binaural renderings generated from spatial room impulse responses (SRIRs) obtained in several rooms and positions and are asked to identify whether they belong to the same space. The perceptual results, together with monaural room acoustical parameters, are used to generate a prediction model that estimates the perceived similarity of two SRIRs.

1 Introduction

Robust externalization of virtual sounds is crucial in *augmented reality* (AR) applications, as it allows a seamless blending between sounds from the real world with those binaurally reproduced by wearable devices. Multiple factors such as the quality of the direct sound cues, reverberation, and head movements can all contribute to the achievement or collapse of an externalized sound image [1]. In particular, in AR applications, it is important to ensure that the reverberation properties of the virtually generated sounds match the acoustics of the listening space, in order to prevent an externalization collapse due to acoustic room divergence [2]. While it is known that a pronounced mismatch between the reverberation time of real and virtual sounds results in externalization collapse [3], it is possible to significantly reduce the spatial resolution of salient reflections and reverberation without affecting the perceived plausibility of virtual sounds, even when compared against real sources [4]. Additionally, room divergence can be overcome partially by continued exposure [5]. Thus, a better understanding of the perceived acoustical similarity between rooms is necessary in order to guide perceptually optimized AR simulations with the goal of generating transfer-plausible acoustical percepts [6, 7].

The problem of perceived room acoustical similarity has been addressed in the past in various studies. For concert hall acoustics, the perceptual fingerprint of a room is determined by the early response [8], while using different musical passages renders such comparisons difficult [9]. With regard to small rooms, studies in variable acoustic rooms suggest that it is relatively easy for listeners to identify various room conditions [10, 11]. Additionally, in an attempt to quantify listener expertise, von Berg et al. [12] conducted a *binaural room impulse response* (BRIR) identification test with manipulated responses compared against a reference. However, to our knowledge, no studies have been conducted in which multiple rooms have been compared and it is thus unknown what the acousti-

cal phenomena are that govern the similarity of small rooms.

The goal of the present work is to lay the foundations towards a perceptually inspired prediction model of room acoustical similarity. While other works attempted to identify rooms based on audio features [13, 14], these approaches solve a classification problem, and thus assume that there is an exact match for the evaluated room within a dataset. Our goal is different, in the sense that we aim to predict the perceived acoustical similarity between two arbitrary rooms that are not necessarily part of a common dataset before nor the system has seen before. This would result in multiple applications in the broader *extended reality* (XR) domain, as the simulation and rendering process of new spaces could be greatly simplified by using the renderings of perceptually equivalent rooms.

In Section 2 of this paper, we present a user study in which we evaluate the perceived similarity between 11 different *spatial room impulse responses* (SRIRs) extracted from a large-scale database, presented binaurally over headphones with *two degrees-of-freedom* (2DoF) rendering. In Section 3, the perceptual results, together with monaural room acoustical parameters estimated from the entire dataset, are used to generate a small-scale non-linear model that is able to predict the probability that two SRIRs are perceived as belonging to the same room.

2 Methods

In this section we present a perceptual experiment that evaluates the perceived similarity between different measured rooms (cf. Section 2.4) presented with dynamic binaural rendering (cf. Section 2.3). The goal of the experiment is two-fold: first, to determine whether listeners are consistently able to discriminate between different rooms and / or form cohesive perceptual room mappings (cf. Section 3.1); and second, to utilize this data to implement data-driven perceptual similarity prediction models for room acoustics (cf. Section 3.4), applicable to XR scenarios.

In order to carry out the perceptual experiment we utilize a portion of a large-scale room acoustical dataset (cf. Section 2.1). The dataset's room acoustical parameters (cf. Section 2.2) are used to perform a dimensionality reduction of the parameter space (cf. Section 3.2), in order to identify meaningful relationships between standard parameters and overall perceived similarity (cf. Section 3.3).

Table 1: Room categories and distribution of the utilized internal measurement dataset.

Category	Rooms		SRIRs	
	#	%	#	%
Bathroom	1	1.1	6	0.3
Cafeteria	3	3.3	272	11.4
Classroom	1	1.1	16	0.7
Cozy room	13	14.3	108	4.5
Game room	1	1.1	54	2.3
Hall	2	2.2	76	3.2
Kitchen	1	1.1	14	0.6
Living room	13	14.3	196	8.2
Lounge	1	1.1	70	2.9
Meeting room	32	35.2	907	38.1
Office	13	14.3	128	5.4
Open office	8	8.8	355	14.9
Shop	2	2.2	176	7.4

2.1 Room measurement dataset

An internal room acoustical dataset composed of 2378 multichannel SRIRs, measured in 91 different rooms, is utilized in the study. Multiple source and / or receiver configurations are present in each room. A summary regarding the categories of the measured rooms and their distribution are provided in Table 1.

Measurements were done with a seven-channel open microphone array of 10 cm diameter, with a center omnidirectional microphone and six omnidirectional microphones arranged in orthogonal pairs, as in [4]. The central microphone is a *Earthworks M30*, while the rest are *DPA 4060*s. This array configuration is well suited for auralization using the *Spatial Decomposition Method* (SDM) [15].

The source is an omnidirectional *Brüel & Kjær Type 4295* which exhibits an increasing on-axis directivity above 1 kHz. Compared to a directional loudspeaker oriented towards the receiver, the used source presents slightly more pronounced excitation of ceiling reflections and attenuated direct sound at high frequencies.

2.2 Room acoustic parameters

For the entire dataset we computed a set of standard monaural parameters in octave bands up to 4 kHz and below as described in ISO 3382:2009 [16] from the central microphone of the utilized measurement array.

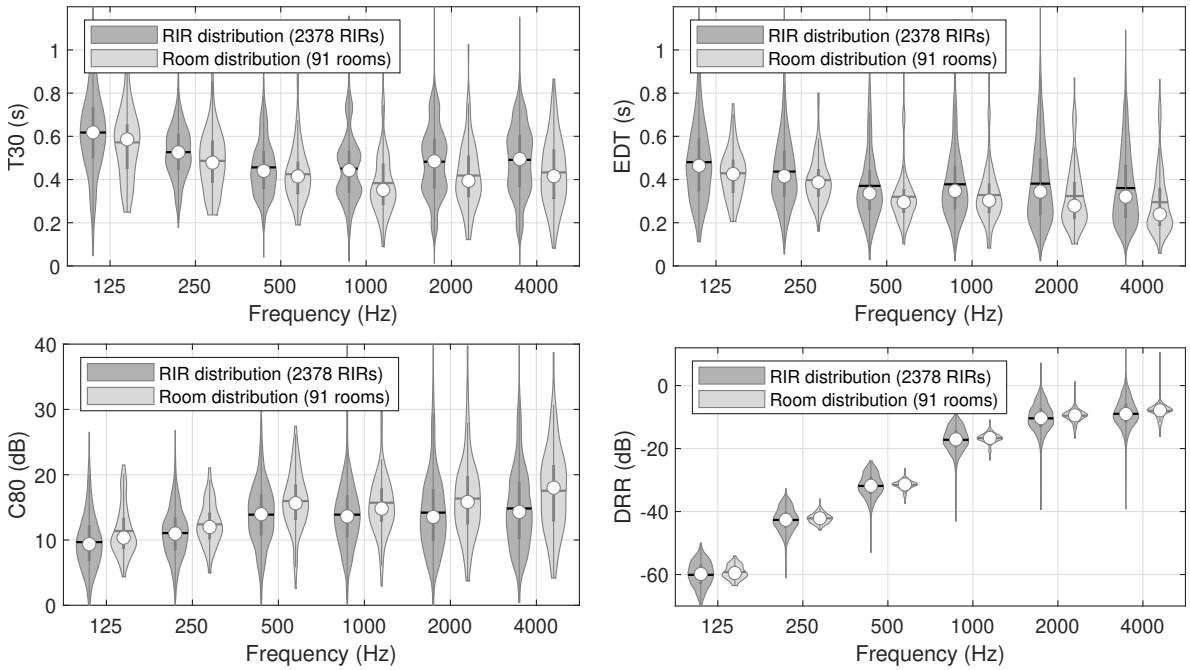

Fig. 1: Distribution of monaural room acoustic parameters T30 (top left), EDT (top right), C80 (bottom left) and DRR (bottom right) showing average (black lines) and median (white circles) values, evaluated in six octave bands of all individual measurement points (dark gray) and averaged within each room (light gray).

The parameters are *reverberation time* (T30), *early decay time* (EDT), *clarity* (C80), and *definition* (D50). Additionally, we computed *direct-to-reverberant ratio* (DRR), which, although not included in the ISO standard, is a highly relevant parameter for distance perception in reverberant environments [17].

The distributions of standard monaural room acoustical parameters for all the measurement points and averaged across rooms are presented in Figure 1. The vast majority of rooms in the dataset correspond to small rooms, with T30 generally between 0 and 1 s, and an average of 0.4 s at mid frequencies. Note as well that the distributions of the parameters differ slightly when grouping RIRs into rooms, due to the different number of measurements in each space and some parameters being highly position dependent, i.e., DRR.

In this study we decided to restrict our investigations to the relationship of standard monaural parameters and perceived similarity, although it is well known that ISO 3382 parameters are often insufficient to fully describe the acoustics of small rooms [17, 18]. In follow up investigations we plan on including other monaural, binaural, and spatial parameters.

2.3 Binaural rendering

The generation of the BRIRs used for auralization in the listening test was performed using the *BinauralSDM* method [4, 15]. The SRIR analysis is performed using the *Spatial Decomposition Method Toolbox*[1] [15], while the rendering portion is achieved using an optimized version for binaural reproduction in the *BinauralSDM* toolbox[2] [4]. In the present study we utilized RTMod+AP equalization to address the known problems of SDM rendering related to reverberation whitening and late reverberation artifacts[3]. Furthermore, the spatial information related to the early reflections and reverberation was quantized to 50 points corresponding to a Lebedev grid, as in [4].

[1] `https://www.mathworks.com/matlabcentral/fileexchange/56663-sdm-toolbox`

[2] `https://github.com/facebookresearch/BinauralSDM`

[3] Note that the quality of the RTMod+AP equalization is highly dependent on the quality of the reverberation time estimation, which tends to be less accurate in settings with very low reverberation. In some extreme cases, we decided to bypass the equalization.

The head-related impulse responses utilized for the rendering correspond to a *Neumann KU100*[4] [19] dummy head. The range and resolution of each BRIR set were $-90°$ to $+90°$ and $2°$ for the horizontal plane, and $-50°$ to $+50°$ and $5°$ for the vertical plane, rendered in a regular grid. The orientation $(0°, 0°)$ is thereby aligned to the estimated direction-of-arrival of the direct sound from the source. For brevity, we address a single BRIR in the remainder of the paper when referring to the whole set of related BRIRs generated from a SRIR.

The late reverberation was rendered statically, corresponding to the head orientation $(0°, 0°)$ and using a mixing time of 80 ms in all cases. Considering that all the rooms used in the experiment correspond to environments with relatively short reverberation times, the chosen mixing time is expected to be well above the perceived mixing time [20].

The real-time rendering for dynamic binaural playback was implemented in Max/MSP, using the same rendering engine described in [4]. Head-tracking was implemented using a *Supperware*[5] head tracker. The dynamic rendering was done in 2DoF, i.e., the tracking of yaw and pitch head movements. Informal evaluation revealed that the motion-to-sound latency was not perceivable. The headphones used for playback were *Beyerdynamic DT-990 Pro* equalized using FIR minimum phase filters[6] [19] matching the headphone's model.

Given that all of the BRIRs used in the study corresponded to different positions and/or rooms and thus presented different time-energy profiles, it was necessary to equalize for the perceived loudness of the stimuli. We equalized all the responses by matching the A-weighted RMS level of the direct sound and early reflections for the frontal head orientation.

2.4 Perceptual experiment

The main objective of the listening test was to investigate the perceptual similarity of various rooms in an attempt to understand what parameters contribute the most to creating a perceptual mapping or internal representation of an acoustic space. To this end, we evaluated 11 different measurements, for which the dataset SRIRs, the utilized rendering scripts as well as the generated BRIRs are made publicly available[7].

Table 2: Description of rooms / positions utilized in the perceptual evaluation. The selection was based on acoustical and room size variety, as well as similar source-receiver-distances. The decisive criteria for each room are marked in bold font.

Room	Room area in m^2	Source dist. in m	$T30_{500\,Hz-2\,kHz}$ in s
Bathroom1	7.6	**1.04**	**0.96**
Cafeteria1	**126.5**	**1.22**	0.72
Cozy1	6.6	**1.24**	**0.22**
Meeting1	**28.6**	**1.38**	**0.42**
Meeting2	36.9	**1.00**	0.72
Meeting3_a	24.6	4.31	0.39
Meeting3_b	24.6	1.13	0.40
Meeting3_c	24.6	1.38	0.39
Meeting3_d	24.6	1.60	0.39
Office1	**4.4**	**1.22**	0.31
Office2	20.9	**1.09**	0.51

As shown in Table 2, four BRIRs corresponded to the same room *Meeting3*, although measured at different distances, positions and orientations. This room represents an average-sized meeting / living space with representative sound absorption and diffusion. The other conditions were chosen at a similar source-receiver-distance in rooms based on their similarity (*Meeting1*) and small differences (*Meeting2*, *Office2*) towards *Meeting3*; as well as to represent extreme cases of the entire measurement dataset for the lowest / highest room size (*Office1*, *Cafeteria1*) and average reverberation time (*Cozy1*, *Bathroom1*), respectively.

A number of 22 subjects (mean age 35 years) participated in the experiment i.e., after the results of two subjects were excluded (cf. Section 3.1). The experimental setup was identical for all participants, whereby three subjects completed the test remotely, while the rest participated on site. The participants were allowed to adjust the reproduction level to a comfortable level during the training phase. Around 60% of the subjects reported to have at least some experience with similar listening experiments in the context of spatial audio.

[4] https://zenodo.org/record/3928297/files/HRIR_FULL2DEG.sofa

[5] https://supperware.co.uk/

[6] https://zenodo.org/record/3928297/files/HPCF.zip

[7] https://github.com/facebookresearch/AVAR_2022_RA_Similarity

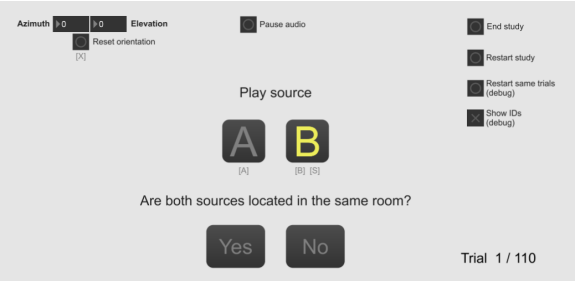

Fig. 2: Screen capture of the graphical user interface used in the perceptual evaluation.

Prior to the experiment, listeners were introduced to the test by completing a variable number of trials with the presence and assistance of the experimenter. Once listeners reported that they were comfortable with the setup and had an appropriate understanding of the task, the actual test would begin. Additionally, listeners were instructed to perform natural head movements during the test in order to explore the direction dependent aspects of the reproduced spaces and to enhance externalization.

The experimental task was implemented as a *two alternative forced choice* (2AFC), in which participants were presented with a source signal convolved dynamically with two different BRIRs and were asked to respond *"Yes"* or *"No"* to the question *"Are both sources located in the same room?"*. Further clarification was provided for the task, specifying that sounds did not need to be identical nor located at the same positions, but they should reply if they thought both sources had been recorded in the same room. The listeners could freely switch between the two presented room conditions with the source signal being continuously auralized and repeated. Both stimuli needed to be heard before the buttons to select a response would be enabled. Once a response had been entered, it could not be edited. A screenshot of the experiment interface is presented in Figure 2.

Two different source signals were investigated: a loop of *drums* and an excerpt of male *speech*. Combined with all the possible stimuli comparisons, this resulted in a total of 110 trials without repetitions (55 room comparisons $\times$ 2 source signals). The average completion time was 24 minutes, with the shortest and longest completion times being 10 and 37 minutes, respectively.

3 Results

3.1 Perceptual evaluation

For every unique condition (the combination of BRIR pairs and one of the two source signals), we average the binary responses from all included participants to obtain a similarity score $p_i \in [0,1]$, which we also refer to as mean perceptual ratings. The results of the perceptual experiment are displayed as similarity matrices in Figure 3. The general trends regarding the similarity scores seem largely independent from the kind of source signal, except in some specific cases, and suggest that listeners are consistently able to identify similar spaces. Upon a preliminary data analysis we discarded the data from two subjects, as their results showed no significant performance to distinguish BRIRs from the same vs. different rooms.

The four BRIRs corresponding to the same room (*Meeting3*) report high perceived similarity values (mean score 0.75, min score 0.50, max score 0.88), which suggests that, despite differences in source and receiver positions, listeners are consistently able to separate position and room dependent acoustical phenomena. The room *Meeting1* was included for its similar acoustic properties (cf. Table 2) and was consistently identified as the same as *Meeting3* (mean score 0.73, min score 0.39, max score 0.96). Rooms with minor deviations in terms of room dimensions and/or frequency-dependent T30 values, e.g. *Office1*, *Office2* and *Meeting2*, were still rated as similar in some cases but with much less consistency.

The room *Cafeteria1* was consistently perceived as highly similar to *Office2* (mean score 0.69) and *Meeting2* (mean score 0.73) due to their similarities in apparent reverberation at the listening position, seemingly rendering large differences in room size as perceptually irrelevant. Finally, the rooms *Cozy1* and *Bathroom1* were consistently perceived as being very different from each other and the rest of the rooms, due to their unique reverberation signatures (smallest and largest T30 values, respectively. A specific case that seems dependent on the played content is the comparison of *Cozy1* and *Office1*, with these rooms being rated as highly similar for the *drums* signal (mean score 0.83), but highly dissimilar for *speech* (mean score 0.25).

Although initial exploration of the relationships between the perceptual ratings and the room acoustical

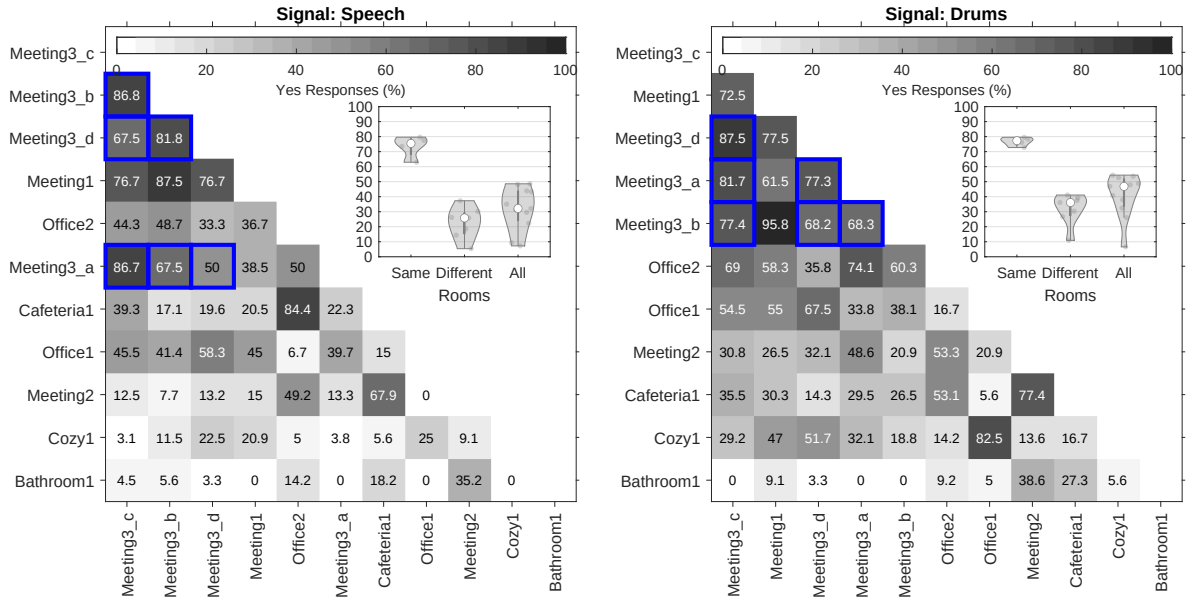

Fig. 3: Mean perceptual ratings (similarity scores) expressed as percentages for the source signals *speech* (left) and *drums* (right). Comparisons between different measured positions in the same room (marked with blue squares) and some selected combinations in different rooms present high similarity scores.

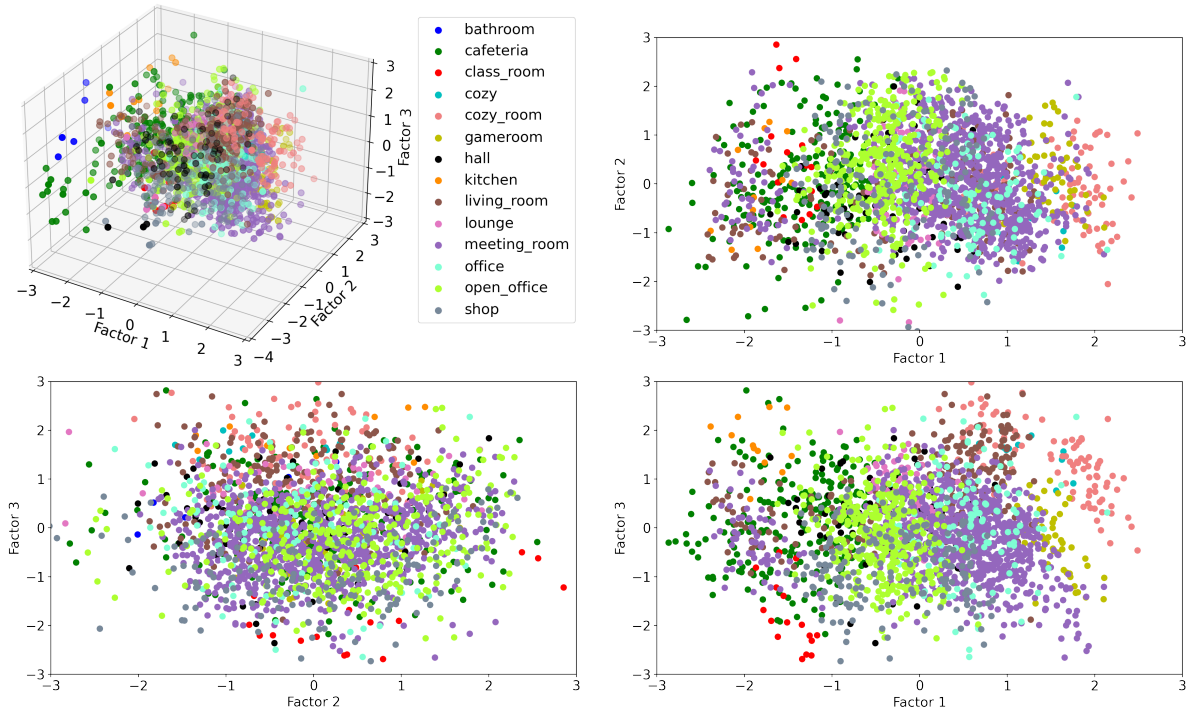

Fig. 4: All measurement points in our SRIR dataset plotted by their factor values, colored by the type of room. 2D projections of the 3D are added for demonstration of possible separability between room types.

parameters revealed clear relationships, it is not practical to explore all room and parameter relationships without reducing the dimensionality of the data.

3.2 Factor analysis

In order to reduce the dimensionality of the room acoustical data we performed a *factor analysis* (FA) with three factors and used a *VariMax* rotation to subsequently compute the factor loadings, as shown in Table 3. The total explained variance of the three factors is approximately 80%. Factor 1 (explained variance around 44%) is mostly composed of contributions from all parameters at mid frequencies except DRR. The contributions to Factor 2 (explained variance around 22%) are mostly from DRR at all frequencies. It is reasonable to assume that in the evaluated dataset the variations in DRR are quasi orthogonal to those of all the other parameters, since within the same room DRR can vastly vary depending on the position, while other energetic parameters are generally more uniform across the space. Finally, contributors to Factor 3 (explained variance around 14%) are C80, D50, and EDT at low frequencies. Note as well that in the case of Factor 1 and Factor 3, the contributions of C80 and D50 are positive, while EDT and T30 present negative loadings. This is explained due to the generally inverse correlation of these two parameter groups.

A similar analysis was conducted by Cerdá et al. [21]. Although in their study the room acoustical parameters are different, some important similarities can be observed. Factor 1 – the one reporting the highest explained variance – is mostly composed of monaural mid frequency parameters (including T30, EDT, and C80, among others), as in our case. The main difference are Factor 2 and Factor 3, which are composed of parameters that we did not include in our analysis i.e., spatial parameters as well as *bass ratio* (BR) and *strength* (G), respectively. Additionally, in our case we included parameters in individual frequency bands, while only averaged single values were used in [21].

Figure 4 shows all SRIRs in the dataset against the three resulting factors from our analysis, colored by their associated room type (cf. Table 1). Visual inspection suggests that categories of rooms are likely separable by non-linear methods that operate on the factor data and suggests scope for important future work involving automatic estimation of room properties based on selected impulse response measurements. We propose to investigate this idea in the future.

Table 3: Explained variance and factor loadings for each of the acoustical parameters. Factor loadings $\geq |0.7|$ are marked in bold font.

Parameter	Factor 1	Factor 2	Factor 3
Explained variance	0.44	0.22	0.14
$C80_{125\,Hz}$	0.22	0.23	**0.84**
$C80_{250\,Hz}$	0.41	0.36	0.65
$C80_{500\,Hz}$	**0.78**	0.29	0.38
$C80_{1\,kHz}$	**0.84**	0.27	0.27
$C80_{2\,kHz}$	**0.86**	0.23	0.21
$C80_{4\,kHz}$	**0.87**	0.23	0.21
$C80_{500\,Hz-2\,kHz}$	**0.88**	0.26	0.28
$D50_{125\,Hz}$	0.13	0.34	**0.72**
$D50_{250\,Hz}$	0.31	0.43	0.55
$D50_{500\,Hz}$	**0.71**	0.41	0.28
$D50_{1\,kHz}$	**0.79**	0.41	0.23
$D50_{2\,kHz}$	**0.82**	0.39	0.20
$D50_{4\,kHz}$	**0.84**	0.38	0.17
$D50_{500\,Hz-2\,kHz}$	**0.83**	0.41	0.23
$DRR_{125\,Hz}$	0.10	**0.78**	0.22
$DRR_{250\,Hz}$	0.16	**0.83**	0.22
$DRR_{500\,Hz}$	0.18	**0.90**	0.20
$DRR_{1\,kHz}$	0.26	**0.91**	0.18
$DRR_{2\,kHz}$	0.31	**0.90**	0.19
$DRR_{4\,kHz}$	0.30	**0.89**	0.18
$DRR_{500\,Hz-2\,kHz}$	0.27	**0.93**	0.19
$EDT_{125\,Hz}$	−0.17	−0.24	**−0.78**
$EDT_{250\,Hz}$	−0.36	−0.32	−0.59
$EDT_{500\,Hz}$	**−0.75**	−0.28	−0.29
$EDT_{1\,kHz}$	**−0.84**	−0.28	−0.24
$EDT_{2\,kHz}$	**−0.87**	−0.24	−0.19
$EDT_{4\,kHz}$	**−0.89**	−0.25	−0.18
$EDT_{500\,Hz-2\,kHz}$	**−0.90**	−0.25	−0.24
$T30_{125\,Hz}$	−0.26	0.02	−0.48
$T30_{250\,Hz}$	−0.52	−0.05	−0.44
$T30_{500\,Hz}$	**−0.80**	−0.06	−0.33
$T30_{1\,kHz}$	**−0.89**	−0.08	−0.22
$T30_{2\,kHz}$	**−0.89**	−0.06	−0.14
$T30_{4\,kHz}$	**−0.85**	−0.07	−0.14
$T30_{500\,Hz-2\,kHz}$	**−0.92**	−0.06	−0.17

3.3 Combination of FA and perceptual results

We next attempt to quantify the contribution of our individual factor loadings (cf. Table 3) to the perceptual ratings provided by the participants in our user study (cf. Figure 3). Figure 5 shows the absolute value of difference between the factors corresponding to each evaluated condition against the corresponding mean perceptual ratings p_i, and compute a linear

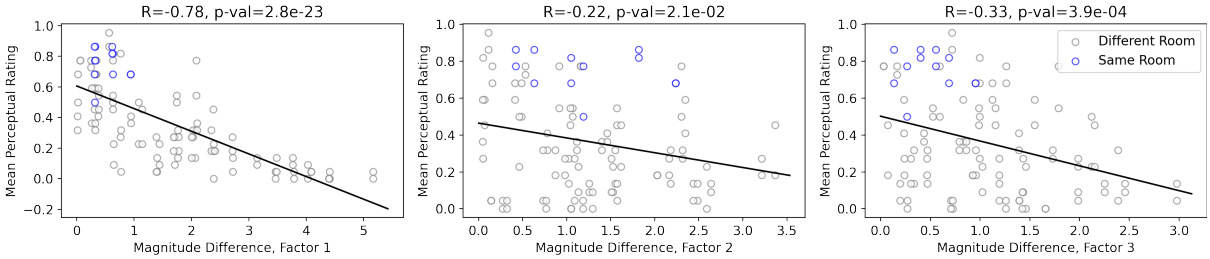

Fig. 5: Relationship between magnitude differences for individual SRIR factors and mean perceptual ratings per evaluated BRIR condition provided by participants. Conditions where the measurements belong to the same room are marked in blue. Observations are reported separately for each source signal.

best fit line and Spearman coefficient between the two. The blue / gray circles indicate perceptual ratings provided for comparisons belonging to the same / different room, respectively. From the figure, we see that magnitude differences in Factor 1 contribute significantly to changes in the mean perceptual rating, with additional contributions from Factor 2 and Factor 3. This suggests that a simple linear mixed model based on our factor loadings could be an effective predictor of the perceptual task. However, we use this information to go one step further and develop a simple probabilistic, non-linear model.

3.4 A model to predict perceptual distance

Using the data from our measurement dataset and the perceptual experiment, we attempt to build a model that suggests the probability that sources rendered from two given SRIRs are perceived as situated in a common room. While a range of non-linear modeling methods are available for use, we are interested in developing a model that is reflective of the sources of user response noise that is likely present within the data, expresses knowledge of this noise as distributions of uncertainty that can be sampled for new test cases in future experimental settings, and that reflects strong domain intuition which allows for generalizing to new measurements and testing scenarios that are significantly different from the conditions under which the data was collected for this work. To this end, we do not assume that our observations exhibit uniform noise, but rather as a function of the input space, considering that our chosen input representation may not be sufficiently expressive. In other words, we construct a model that incorporates epistemological uncertainty.

First, as input to our model, we compute a one-dimensional distance measure that is a function of our factors. Specifically:

$$d = \sqrt{\sum_{i=1}^{N=3} \lambda_i (f_i^a - f_i^b)^2}\,, \tag{1}$$

which represents a weighted ℓ_2 distance between the factors. Here, f_i^a is the i-th factor value for stimulus *"A"* of the two being compared, and λ_i is a weighting term that is the normalized value of the explained variance corresponding to the i-th factor. We compute this measure for all $m = 110$ observations in our perceptual dataset, called $X = \{d_i\}_{i=1}^m$. The mean perceptual ratings are labeled $Y = \{y_i\}_{i=1}^m$. We design a model such that:

$$y = f(x) + \varepsilon\,, \tag{2}$$

wherein we assume that f is the underlying function mapping X to Y and is drawn from a Gaussian distribution, and that ε models the non-uniform noise in the data. We introduce a *Gaussian Process Regression* (GPR) to model this mapping:

$$y \sim GP\left(\mu(x),\, K_\theta(x,x')\right)\,. \tag{3}$$

We assume the observations $y \in Y$ are generated by sampling from a Gaussian process with mean function $\mu(x)$, and the composite covariance matrix function:

$$K_\theta(x,x') = C + exp\left(-\frac{|x-x'|}{2l_{mean}^2}\right) + \delta(x)\,I\,. \tag{4}$$

The kernel matrix is a sum of a constant bias term, a radial basis function, and a heteroscedastic noise kernel which attempts to learn explicit noise levels for input points X. This kernel uses a set of inducing points $X_p \in X$ to learn a noise model, and then generalizes

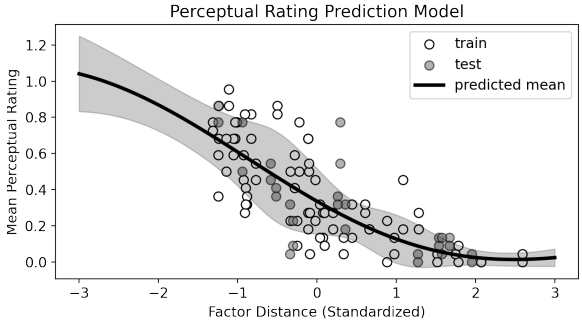

Fig. 6: Mean function and input-varying noise learned by our heteroscedastic GPR model, plotted against our randomly split train and test observations.

it to the rest of the input space via kernel regression, following [22, 23]. A single entry x_i in the final kernel term can be computed as:

$$\delta(x_i) = \sum_{X_{p_i} \in X_p} \left[exp\left(-\frac{|x_{p_i} - x_i|}{2l_{noise}^2} \right) \cdot \sigma_{x_{p_i}, x_i} \right] . \quad (5)$$

The set X_p is chosen via a 10-point *KMeans* clustering analysis on X [24]. The kernel parameters $\theta = \{C, l_{mean}, l_{noise}, \sigma_{x_{p_i}, x_i}\}$ are learned and optimized to maximize the log marginal likelihood $\log(p(Y|X,\theta))$ using the standard *L-BFGS-B* algorithm via the *SciPy* GPR implementation[8].

To evaluate our model, we randomly split our observations into a train and a test set, consisting of 75% and 25% of the data respectively, and standardize X to have a mean of zero and standard deviation of one. The kernel parameters are optimized over the train set, and our test set is compared against the values predicted by the mean function $\mu(x)$ of the model. Figure 6 shows the train and test observations against the mean function and the learned variance. As reported in Table 4, when the mean function of the GPR is used for prediction, the model achieves a *mean absolute error* (MAE) of 0.12, and a *root mean squared error* (RMSE) of 0.17. Examining the variance learned by the model, we note that the model learns to predict a greater spread in the mean perceptual rating probabilities for smaller factor distances than for larger ones; in other words, the model is less certain about its predictions for SRIRs with highly similar room acoustic parameter values, reflecting the nature of the experimental data.

[8] https://scikit-learn.org/stable/modules/gaussian_process.html

Table 4: Benchmark of our model with several other linear and non-linear models, which operate directly on the two sets of factor values from the SRIRs being compared. We report mean absolute error and root mean squared error using the models' predictions on the test set.

Model type	MAE	RMSE
Baseline	0.22	0.26
ElasticNet Regression	0.23	0.27
Linear Regression	0.16	0.21
Extra Trees	0.13	0.16
Random Forest	0.12	0.15
SVM	0.12	0.14
Gaussian Process Reg.	0.12	0.17

As a benchmark and a validation for our hand-tuned input function d, we train several other standard linear and non-linear machine learning methods on the same data partition. These models operate directly on the two sets of factor values as input, instead of using the input function d. We also include a comparison against an uninformed baseline predictor that outputs only the median values of Y. We provide the resulting MAE and RMSE values for comparison in Table 4. We find that our model performs comparably to the other benchmarks, with the added benefit that sampling the model at a particular value d_i provides a probabilistic estimate of the epistemological noise. In the future, we intend to use this model to design specific perceptual cases using selected measurement points that correspond to the regions of the greatest uncertainty, and update the model with additional perceptual data as it is acquired.

4 Summary and conclusions

In this paper we presented a perceptual study in which we investigated the perceived acoustical similarity of several rooms reproduced with dynamic head-tracked binaural audio (BRIRs available for reference[7]). Moreover, we conducted a *factor analysis* on the acoustical parameters computed from a large-scale room acoustical dataset to reduce the dimensionality of the data. Finally, we used the perceptual ratings together with the results of the FA to implement a *Gaussian Process Regression* model for the prediction of perceived room acoustical similarity.

While we are not releasing the full room acoustical dataset at the moment, we conclude from our investigations that such data may allow for the following:

- Listeners are consistently able to identify that played stimuli correspond to the same room, even if measured positions are not the same. Moreover, certain rooms present consistently high perceived similarity. The results are largely independent from the used source signal (*speech* or *drums*).
- The FA conducted on the room-acoustic dataset reveal that three factors account for 80% of the variance in the data. The main contributors to these factors are mid frequency parameters (T30, EDT, C80, D50), DRR, and low frequency parameters (C80, D50, EDT), respectively. This partially confirms previous results from the literature [21].
- Factor 1, accounting for 44% of the explained variance in the room parameters, is strongly correlated with the perceptual ratings.
- Our proposed GPR model is able to predict perceptual ratings with a mean absolute error of 0.12 (relative to our mean perceptual observations which lie in the range 0 to 1).

In future work we plan to continue with further data collection for room conditions with small differences in their acoustical parameters, in order to provide our model with more information in the more ambiguous region and to increase the potentially limited generalization arising from using a reduced room dataset in the tests. Additionally, we plan on incorporating other room-acoustical parameters, including monaural parameters such as *strength* (G), *bass ratio* (BR), *brilliance* (Br), *spectral centroid* (f_c) or *echo density profile* [25, 20]; as well as spatial parameters such as *lateral energy fraction* (J_{LF}), *late lateral sound level* (L_j); or binaural parameters such as *interaural cross-correlation* (IACC) or *predicted binaural coloration* (PBC) [26], for early and late parts of the SRIRs / BRIRs respectively. Additionally, since the dataset contains SRIRs, we will have the possibility of exploring more complex spatial parameters, such as the (an)isotropy of the late reverberation [27] or front / back / lateral relative energy [28].

Finally, we plan on implementing multiple versions of our model, suitable for different sets of measurement data and analysis, e.g. monaural RIRs, BRIRs or other forms of SRIRs.

5 Acknowledgments

We want to thank Zamir Ben-Hur for discussions along the project as well as the test participants, research assistants and other personnel supporting the data collection.

References

[1] Best, V., Baumgartner, R., Lavandier, M., Majdak, P., and Kopčo, N., "Sound Externalization: A Review of Recent Research," *Trends in Hearing*, 24, pp. 1–14, 2020, doi:10.1177/2331216520948390.

[2] Werner, S., Klein, F., Mayenfels, T., and Brandenburg, K., "A Summary on Acoustic Rpom Divergence and its Effect on Externalization of Auditory Events," in *Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6, IEEE, Lisbon, Portugal, 2016, doi:10.1109/QoMEX.2016.7498973.

[3] Amengual Garí, S. V., Hassager, H. G., Klein, F., Arend, J. M., and Robinson, P. W., "Towards Determining Thresholds for Room Divergence: A Pilot Study on Perceived Externalization," in *International Conference on Immersive and 3D Audio*, pp. 1–7, IEEE, Bologna, Italy, 2021, doi:10.1109/I3DA48870.2021.9610835.

[4] Amengual Garí, S. V., Arend, J. M., Calamia, P. T., and Robinson, P. W., "Optimizations of the Spatial Decomposition Method for Binaural Reproduction," *Journal of the Audio Engineering Society*, 68(12), pp. 959–976, 2020, doi:10.17743/JAES.2020.0063.

[5] Klein, F., Werner, S., and Mayenfels, T., "Influences of Training on Externalization of Binaural Synthesis in Situations of Room Divergence," *Journal of the Audio Engineering Society*, 65(3), pp. 178–187, 2017, doi:10.17743/jaes.2016.0072.

[6] Wirler, S. A., Meyer-Kahlen, N., and Schlecht, S. J., "Towards Transfer-Plausibility for Evaluating Mixed Reality Audio in Complex Scenes," in *International Conference on Audio for Virtual and Augmented Reality*, pp. 1–10, Audio Engineering Society, Virtual, 2020.

[7] Meyer-Kahlen, N., Amengual Garí, S. V., Mckenzie, T., Schlecht, S. J., and Lokki, T., "Transfer-Plausibility of Binaural Rendering with Different Real-World References," in *DAGA 2022: 48. Deutsche Jahrestagung für Akustik*, pp. 154–157, Deutsche Gesellschaft für Akustik, Stuttgart, Germany, 2022.

[8] Haapaniemi, A. and Lokki, T., "Identifying concert halls from source presence vs room presence," *Journal of the Acoustical Society of America*, 135(6), pp. EL311–EL317, 2014, doi:10.1121/1.4879671.

[9] Kuusinen, A. and Lokki, T., "Recognizing Individual Concert Halls is Difficult When Listening to the Acoustics with Different Musical Passages," *Journal of the Acoustical Society of America*, 148(3), pp. 1380–1390, 2020, doi:10.1121/10.0001915.

[10] Dlugosz, Z., *The Recognition of Room Acoustics with Different Speech Signals: An Experiment with Auralisations*, Master thesis, Aalto University, 2021.

[11] Klein, F., Amengual Garí, S. V., Arend, J. M., and Robinson, P. W., "Towards Determining Thresholds for Room Divergence: A Pilot Study on Detection Thresholds," in *International Conference on Immersive and 3D Audio*, pp. 1–7, IEEE, Bologna, Italy, 2021, doi:10.1109/I3DA48870.2021.9610876.

[12] von Berg, M., Steffens, J., Weinzierl, S., and Müllensiefen, D., "Assessing Room Acoustic Listening Expertise," *Journal of the Acoustical Society of America*, 150(2539), pp. 2539–2548, 2021, doi:10.1121/10.0006574.

[13] Peters, N., Lei, H., and Friedland, G., "Name That Room: Room Identification Using Acoustic Features in a Recording," in *International Conference on Multimedia*, p. 841, ACM Press, New York, USA, 2012, doi:10.1145/2393347.2396326.

[14] Moore, A. H., Naylor, P. A., and Brookes, M., "Room Identification Using Frequency Dependence of Spectral Decay Statistics," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6902–6906, IEEE, Calgary, Canada, 2018, doi:10.1109/ICASSP.2018.8462008.

[15] Tervo, S., Pätynen, J., Kuusinen, A., and Lokki, T., "Spatial Decomposition Method for Room Impulse Responses," *Journal of the Audio Engineering Society*, 61(1/2), pp. 17–28, 2013.

[16] ISO 3382-1, "Acoustics – Measurement of Room Acoustic Parameters – Part 1: Performance Spaces," 2009.

[17] Zahorik, P., "Perceptually Relevant Parameters for Virtual Listening Simulation of Small Room Acoustics," *Journal of the Acoustical Society of America*, 126(2), pp. 776–791, 2009, doi:10.1121/1.3167842.

[18] Kaplanis, N., Bech, S., Jensen, S. H., and Van Waterschoot, T., "Perception of Reverberation in Small Rooms: A Literature Study," in *55th International Conference: Spatial Audio*, pp. 1–14, Audio Engineering Society, Helsinki, Finland, 2014.

[19] Bernschütz, B., "A Spherical Far Field HRIR/HRTF Compilation of the Neumann KU 100," in *Fortschritte der Akustik – AIA/DAGA 2013*, pp. 592–595, Deutsche Gesellschaft für Akustik, Meran, Italy, 2013, doi:10.5281/zenodo.3928296.

[20] Lindau, A., Kosanke, L., and Weinzierl, S., "Perceptual Evaluation of Physical Predictors of the Mixing Time in Binaural Room Impulse Responses," *Journal of the Audio Engineering Society*, 60(11), pp. 887–898, 2012.

[21] Cerdá, S., Giménez, A., Romero, J., Cibrián, R., and Miralles, J. L., "Room Acoustical Parameters: A Factor Analysis Approach," *Applied Acoustics*, 70(1), pp. 97–109, 2009, doi:10.1016/j.apacoust.2008.01.001.

[22] Cawley, G. C., Talbot, N. L., Foxall, R. J., Dorling, S. R., and Mandic, D. P., "Heteroscedastic Kernel Ridge Regression," *Neurocomputing*, 57, pp. 105–124, 2004, doi:10.1016/j.neucom.2004.01.005.

[23] Rasmussen, C. E., "Gaussian Processes in Machine Learning," in O. Bousquet, U. von Luxburg, and G. Rätsch, editors, *Advanced Lectures on Machine Learning*, pp. 63–71, Springer Berlin Heidelberg, 2004, doi:10.1007/978-3-540-28650-9_4.

[24] Likas, A., Vlassis, N., and J. Verbeek, J., "The Global K-Means Clustering Algorithm," *Pattern Recognition*, 36(2), pp. 451–461, 2003, doi:10.1016/S0031-3203(02)00060-2.

[25] Abel, J. S. and Huang, P., "A Simple, Robust Measure of Reverberation Echo Density," in *Audio Engineering Society Convention 121*, pp. 1449–1458, Audio Engineering Society, San Francisco, USA, 2006.

[26] McKenzie, T., Armstrong, C., Ward, L., and Murphy, D. T., "Predicting the Colouration Between Binaural Signals," *Applied Sciences*, 12(5), pp. 1–15, 2022, doi:10.3390/app12052441.

[27] Alary, B., Massé, P., Schlecht, S. J., Noisternig, M., and Välimäki, V., "Perceptual Analysis of Directional Late Reverberation," *Journal of the Acoustical Society of America*, 149(5), pp. 3189–3199, 2021, doi:10.1121/10.0004770.

[28] Amengual Garí, S. V., Kob, M., and Lokki, T., "Investigations on Stage Acoustic Preferences of Solo Trumpet Players Using Virtual Acoustics," in *14th Sound and Music Computing Conference (SMC)*, pp. 167–174, Aalto University, Espoo, Finland, 2017.