OPEN ACCESS Freely available online

L. Treybig, F. Klein, S. Werner, S. V. Amengual Garí, and S. Göring, "Predicting the Perceptibility of Room Acoustic Variations Using Generalized Linear Mixed Models" *J. Audio Eng. Soc.*, vol. 73, no. 11, pp. 692–706 (2025 Nov.). https://doi.org/10.17743/jaes.2022.0232.

Predicting the Perceptibility of Room Acoustic Variations Using Generalized Linear Mixed Models

LUKAS TREYBIG,[1,*] *AES Associate Member*, **FLORIAN KLEIN,**[1,2]
(lukas.treybig@tu-ilmenau.de) (florian.klein@tu-ilmenau.de)

STEPHAN WERNER,[1] **SEBASTIÀ V. AMENGUAL GARÍ,**[4] *AES Associate Member* **AND STEVE GÖRING**[3]
(stephan.werner@tu-ilmenau.de) (samengual@meta.com) (steve.goering@tu-ilmenau.de)

[1]*Technische Universität Ilmenau, Electronic Media Technology Group, Ilmenau, Germany*
[2]*Fraunhofer Institute for Digital Media Technology IDMT, Electroacoustics Group, 98693 Ilmenau, Germany*
[3]*Technische Universität Ilmenau, Audiovisual Technology Group, Ilmenau, Germany*
[4]*Reality Labs Research, Meta, Redmond, WA*

Acoustic augmented realities aim to incorporate virtual elements on top of real acoustic environments. A key question is how accurately virtual acoustics must match real acoustics to enable perceptually seamless integration. In six-degrees-of-freedom scenarios, where listeners can explore spaces, intraroom acoustic variation is of particular interest. This paper presents a method for modeling the perceptibility of acoustic changes in an acoustic augmented realities context using a generalized linear mixed model based on differences in room acoustic parameters computed from pairs of room impulse responses. An online listening test based on extensive acoustic measurements investigated perceptible differences between three acoustic configurations of the same room using identical source-receiver arrangements. A total of 4,753 ratings from 120 participants in an ABX-like format (a method where participants identify whether a presented stimulus X matches stimulus A or B) were used to train the generalized linear mixed model. The resulting model shows good predictive performance (Brier score: 0.19; AUC: 0.67) based on a fivefold cross-validation with 100 repetitions, although improvements are possible. However, the perceptibility of acoustic changes can vary significantly between dynamic and static listening scenarios. The paper discusses evaluation strategies for these aspects, examines the model's generalizability, highlights current limitations, and outlines directions for future improvement.

0 INTRODUCTION

Room acoustic perception is an important area of research within the field of audio engineering and psychoacoustics. It becomes even more relevant when experiences involve acoustically augmented environments where virtual audio objects must be seamlessly integrated into the real room acoustics via headphones. This requires a position-dynamic binaural synthesis that continuously adapts to the real acoustic environment, creating a plausible auditory augmented reality (AAR) [1, 2].

Acoustic environments generally vary in their local room acoustic properties. Especially in small rooms or rooms with complex geometries, significant position-dependent acoustic variations can be expected, such as the amplitude of room modes, patterns of first reflections, or early/late energy ratios. Although small but perceivable acoustical differences could influence localization and distance perception, larger differences could also lead to externalization issues or the collapse of externalization [3, 4]. It is therefore desirable to quantify these differences in terms of their perceptual relevance to optimize the computational cost of position-dynamic binaural synthesis to the local acoustic environment. The prediction of perceptible acoustic variations should ideally be based on acoustic parameters that can be estimated in real time.

Perception is highly individual; therefore, methods that enable a discrimination of listener-specific variations are required. In the current field of research, the use of such methods has not been established yet. Additionally, related research often lacks a substantial database to generate generalizable prediction models.

This paper presents an approach for modeling the perceptibility of acoustic changes in an AAR scenario using a generalized linear mixed model (GLMM). GLMMs are well suited for analyzing also nonnormally distributed out-

*To whom correspondence should be addressed, e-mail: lukas.treybig@tu-ilmenau.de

come variables, such as binary or ordinal data, and allow for the inclusion of individual variability across groups, for example, differences between participants. Given these properties, many application fields can benefit from their use, such as medicine, biology, social sciences, or education [5]. The prediction model developed in this study is based on room acoustic parameters as predictors and on ratings from a listening test that evaluated position-dependent perceptual differences in binaural renderings derived from acoustic measurements.

A large-scale listening test was conducted to obtain the ground truth for the development of the model. In this paper, the authors explain their decisions regarding the measurements and the design of a large-scale listening test. Furthermore, it is shown how GLMMs can be used to analyze and create predictive models of position-dependent room acoustical differences. Additionally, a comparison to classical just noticeable differences (JNDs) is presented.

1 MIXED EFFECT MODELING

In this paper, GLMMs are used to analyze the listening test and to create prediction models. This section first explains the relevant basics of GLMMs.

A GLMM consists of a combination of different assumptions or extensions of linear models (LMs). With LMs, it is assumed that the unexplained part (residuals) of the model is normally distributed. By extending it to a generalized linear model, other distributions, such as binomial or Poisson distributions, can also be assumed. A generalized linear model is only created on the basis of fixed effects. In a GLMM, a distinction is made between fixed and random effects. Fixed effects describe the effects of a parameter that is fixed for different observations. These are usually test parameters that can be varied in a controlled manner, [e.g., room acoustic parameters from the measured room impulse responses (RIRs)]. Random effects, on the other hand, describe the unexplained variability in the data. When different subjects evaluate a stimulus, the subjects can be considered a random effect. The general form of a GLMM in matrix notation is

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \varepsilon, \tag{1}$$

where $\mathbf{y}$ is an $N \times 1$ column vector (with N denoting the number of observations), the outcome or response variable; $\mathbf{X}$ is an $N \times p$ matrix of p predictor variables; β is a $p \times 1$ vector of the fixed effects coefficients (that is, the learned model weights); $\mathbf{Z}$ is the $N \times q$ design matrix for q random effects; $\mathbf{u}$ is a $q \times 1$ vector of the random effects; and ε is an $N \times 1$ column vector of the residuals, containing the rest of $\mathbf{y}$, which is not explained by $\mathbf{X}\beta + \mathbf{Z}\mathbf{u}$ [6, 7].

For better clarity, as an example, the data are applied to this formula. Assume that a listening test was performed in which differences between two stimuli were evaluated. The rating indicated whether a difference was perceived or not. The vector $\mathbf{y}$ then contains all the ratings from the listening test. The differences between the stimuli can be described, for example, by the differences in room acoustic parameters. When these are used as fixed effects, they are represented in $\mathbf{X}$. The coefficients determined for these fixed effects are contained in β. If participants are considered as random effects, $\mathbf{Z}$ has the information on which rating was given by which participant. Similar to β for fixed effects, $\mathbf{u}$ contains the coefficients for the random effects, that is, for each individual participant.

To apply GLMMs, models are typically specified using formulae, as is commonly done in statistical software such as MATLAB or R. The Wilkinson notation, as described by [8], is commonly used for this purpose. This notation allows for a compact and symbolic description of the model without the need to explicitly state the coefficients of individual terms. The notation consists of the response variable y and a model term that includes fixed and random effects, connected by the symbol $\sim$. Below is a simple example of such a formula relevant to this work:

$$\mathrm{y} \sim 1 + \mathrm{x}_1 + \mathrm{x}_2 + (1|\mathrm{Participant}). \tag{2}$$

Here, y, as previously explained, denotes the response variable. The expression $1 + \mathrm{x}_1 + \mathrm{x}_2$ represents the fixed effects: 1 indicates the inclusion of a global (fixed) intercept, which corresponds to the expected value of y when all predictors are set to zero (or their reference category). The variables x_1 and x_2 are additional explanatory variables whose effects are estimated as fixed effects. The term (1|Participant) specifies a random intercept for each participant. This means that each participant has their own baseline level of the response variable. By modeling each participant's deviation from the overall (global) intercept, the model accounts for both the correlation between participants and participant-specific variability. Similar to fixed effects, random effects can include additional terms beyond the intercept, such as dependence structures [e.g., $(1 + x_1|\mathrm{Participant})$]. However, because this approach is not employed in the present work, it will not be discussed in detail.

By including -1 (or equivalently 0) instead of 1 for the fixed effects in the model formula, the global intercept is omitted, and the model is fitted without an intercept term. In [8], further notations are listed, but these are not used in this paper.

Once the models have been created, the goodness of fit can be examined. The objective is to examine how well the various fixed effects describe the observations. There are a variety of criteria and methods for this. Deviance, Akaike information criterion (AIC), Bayesian information criterion (BIC), and adjusted R^2 ($\mathrm{R}^2_{\mathrm{adj}}$) are considered below.

The log-likelihood (LL) function is the basis for deviance, AIC, and BIC. The deviance is the difference between the maximized LL of a model under test and the LL for a so-called saturated model [9].

In contrast to deviance, AIC and BIC also account for model complexity by including the total number of estimated parameters (k) as a penalty term. The AIC is calculated as $\mathrm{AIC} = -2\mathrm{LL} + 2k$. For the BIC, the sample size (n) is additionally considered in the penalty term, resulting in the formula $\mathrm{BIC} = -2\mathrm{LL} + \ln(n)k$. Because this penalty increases logarithmically with the sample size, BIC penalizes complex models more heavily as the data set grows

and thus tends to favor simpler models when large amounts of data are available. For deviance, AIC, and BIC, smaller values indicate a better fitting. It should be noted that these values are only relative and cannot in themselves provide any information about the quality of the model. They can only be used as a measure in comparison with the values of other models that were generated on the basis of the same observations [9].

The last quality measure considered, the R^2, indicates the proportion of variation explained by the dependent variable. In the adjusted R^2 (R^2_{adj}) used here, the number of observations is taken into account. It is defined as $R^2_{adj} = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE}{SST}$, where *SSE* is the sum of squared errors and *SST* is the total sum of squares. Although R^2_{adj} typically ranges from 0 to 1, it can also take values below zero for poorly fitting models where a value of 1 indicates a perfect model fit [9]. For normally distributed observations, values of about 0.35 for R^2_{adj} are expected to indicate a good model. A value that is too high could indicate overfitting [10]. In the following analyses, binary data are used; accordingly, low adjusted R^2_{adj} values are considered acceptable because binary outcome variables permit less explained variance compared with continuous variables. A value greater than 0.1 is assumed to be acceptable. The crucial criterion is that the model identifies statistically significant predictors and demonstrates a superior fit compared with alternative models [10].

2 RELATED WORK

The International Standard Organization standard ISO 3382-1:2009 defines several room acoustical parameters and links them to corresponding perceptual features with associated JNDs [11, 12]. However, several studies have shown that such JNDs are not absolute but can in fact depend on several factors, such as signal content, signal level, or room conditions, among others [13–15]. Furthermore, even JNDs for the most common acoustical parameter, reverberation time (RT), have been found to differ in various studies using different experimental setups [16–19]. Recent work implementing a 2D threshold test explored whether the JNDs of seemingly perceptually orthogonal acoustic parameters [i.e., direct-to-reverberant energy ratio (DRR)] and RT are truly independent or whether their absolute values actually influence each other's JNDs [20].

Although a granular characterization of the JND of each acoustic parameter could contribute to a deeper understanding of room acoustical perception, the advent of AAR technologies requires a practical approach to produce transfer-plausible acoustics in which virtual sources can be seamlessly inserted into real soundscapes [21, 22, 2, 23, 24]. For instance, discrimination between stimuli with small acoustical differences becomes more complicated when signals with different content are used [25]. Whereas listeners are consistently able to identify whether recordings are made in the same room [26], they are generally not able to correctly identify the actual receiver position from which the recordings were taken, even when clearly audible acoustical differences between different positions in a room are present [20]. This suggests that AAR systems can indeed significantly relax the rendering constraints in certain situations.

To apply this learning systematically, AAR systems can leverage the use of perceptual models, which could predict whether the acoustics of two positions or two rooms are perceptually indistinguishable and thus lead to the reduction of computational demands of a real-time renderer. In [26], a room acoustical similarity model was developed from experimental tests based on paired comparisons of binaurally resynthesized measurements from several rooms. The model was based on a Gaussian process regression that mapped the values of a factor analysis (FA) of several monaural room acoustical parameters with the results of the perceptual test.

Another approach, which is also pursued in this paper, is the use of mixed models. In [27–29], a linear mixed model (LMM) is used for the analysis, whereby the subjects or participants who gave ratings are considered as random effects. In [30–32], room acoustic parameters are also considered as fixed effects, and a link to perception is established.

In [30], LMMs were used to examine the relationship between acoustic comfort and room acoustic parameters. Here, acoustic comfort refers to the subjective and perceptual assessment of how comfortable a person feels in response to the acoustic environment. The results showed that parameters such as clarity index C_{80} and C_{50}, early decay time (EDT), and the Speech Transmission Index predicted acoustic comfort more accurately than baseline models, which only included the type of sound source as a fixed effect. It was emphasized that collinearity must be avoided and that including more than two parameters as fixed effects does not significantly improve model predictions. The models were evaluated and selected using AIC and BIC.

In [32], LMMs were used to predict Room Acoustical Quality Inventory factors, with room acoustic parameters first grouped into three factors using FA. These factors, along with additional parameters, were then included as fixed effects. Model quality was compared using, among other measures, the marginal R^2 (based only on fixed effects) and the conditional R^2 (based on fixed and random effects).

In [31], LMMs were applied to analyze the relationship between room acoustics and subjective impressions. The study identified significant impressions related to room supportiveness, room noiselessness, and room timbre, using room acoustic parameters as fixed effects. Model selection was based on AIC.

In [33] and [34], the authors presented preliminary experiments and results based on LMMs. The listening test described in the present study is an adapted online version of the laboratory study from [34]. That study demonstrated high discriminability between almost all tested positions within a single room, with an overall hit ratio of 98%, where the hit ratio is defined as the ratio of correct identifications of the odd stimuli to the total number of trials. As a result, modeling based on the hit ratio alone was not feasible. As a fallback solution, [34] used attribute ratings for

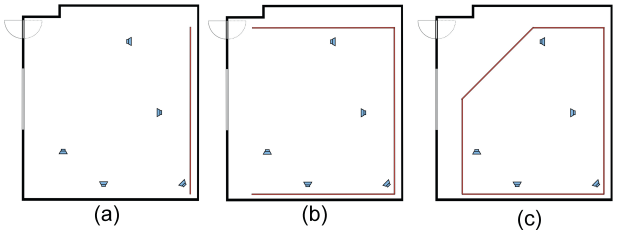


Fig. 1. Room configurations used in the listening test. HL01W (a), HL03W (b), and HL05W (c). The red lines mark the positions of the partition walls. The loudspeaker positions are marked in blue. Figure from [35].

modeling. Although this approach allowed a relative judgment (acoustic differences are more or less pronounced), the authors aimed for an absolute judgment, which is only possible when tested at the level of JNDs.

From the previous study, it was concluded that the high hit ratio was a combination of the specific room and the exclusive participation of expert listeners. To enable modeling based on hit ratios, a broader data set encompassing a variety of room configurations was required. The current study was expanded to include additional room configurations requiring a greater number of ratings. Consequently, an online study was chosen and conducted, featuring an extended set of stimuli and listeners, both naive and experts. This aims to improve the generalizability of the prediction model.

3 DATA ACQUISITION

To create a model that describes the perceptibility of acoustic changes, measurement data and the corresponding perceptual ratings are needed. The following sections briefly describe the utilized room acoustic measurements, how they were collected, and the design of the listening test.

3.1 ACOUSTIC DATA

Acoustic measurements from an open-source database recorded by the authors were used [35, 36, 37]. The measurements were conducted by an autonomous mobile robot in an acoustically dry room that meets the acoustic requirements of the ITU-R BS.1116-3 [38]. The room was modified by sequentially adding highly reflective walls to change the patterns of the early reflections and the RT. The full data set includes nine wall configurations and five loudspeakers of type Geithain RL906 as sound sources at several positions and orientations within the room. A 2D measurement grid with a uniform spacing of 0.25 m was chosen to achieve a high spatial density with a manageable measurement effort. For the measurements of seven-channel spatial room impulse responses (SRIRs), a custom-built, circular and open microphone array was chosen, which is suitable for the application of the spatial decomposition method [39].

For this publication, a subset of three room configurations was selected from the data set. Fig. 1 shows the selected configurations. The mean broadband RT and standard deviation over all measured source-receiver positions of the room configurations are 0.25 s ± 0.01 s (one wall), 0.60 s ± 0.11 s (three walls), respectively, and 1.05 s ± 0.04 s (five walls).

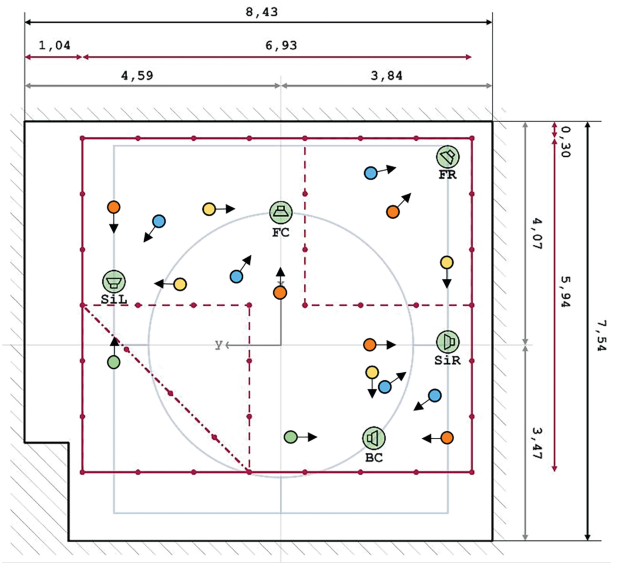


Fig. 2. Floor plan, room configurations, speaker positions and listening positions used in the listening test. For each speaker (SiL, FC, FR, SiR, BC), listening positions in the same relative direction and distance are defined. Frontal positions are in orange, side positions are in blue, far-side positions are in yellow, and back positions are in green. Figure modified from [35].

To create audio samples, listening positions have to be defined, which in turn influence the input features that the model will utilize. All positions have a distance of 1.5 m to the source to avoid the potential influence of direct sound level. Because the recordings were conducted with directional speakers, it was decided to choose listening positions with different relative angles. By changing relative position, reflected early energy will arrive from different spatial angles, which could be a relevant cue in small rooms. All chosen positions are shown in Fig. 2. For each speaker, a frontal position (at 0°, colored in orange), side position (at 30°, colored in blue), far-side position (at 90°, colored in yellow), and back position (180°, colored in green) were selected. In the listening test, only positions of the same color (i.e., the same source-receiver relative angle) are compared against each other. This is because changes in the frequency spectrum resulting from the speaker directivity could potentially mask room acoustic properties.

For the prediction, the differences between positions are described by the absolute differences in the acoustic parameters. However, this means that information about the absolute parameter values are not taken into account. The authors are aware that this simplification may be problematic for parameters such as DRR because the JND varies depending on the absolute DRR value. This issue will become apparent later when considering the prediction results. The following parameters are considered, which are calculated from the measurements of the central measurement microphone of the array: EDT, C_{50}, C_{80}, definition D_{50}, DRR, center time T_S, relative strength G_{rel} [22], and decay time DT_{20} [40].

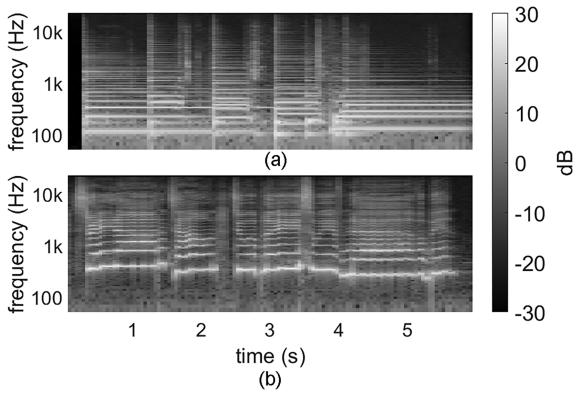


Fig. 3. The spectrograms of the anechoic recordings of a picked acoustic guitar (a) and a female voice (b).

The DT_{20} value corresponds to RT; however, unlike T_{20} or T_{30}, it is not calculated according to DIN EN ISO 3382 [11] because this standard is not applicable to small rooms with short source-receiver distances. To minimize the influence of direct sound, the direct sound is detected, and the estimation of DT_{20} begins 2 ms after its arrival. The approach then considers the decay behavior of the energy decay curve and averages it over multiple ranges [40]. For more details and the calculation of these parameters, refer to [11, 41].

3.2 LISTENING TEST

The publicly available binaural spatial decomposition method toolbox [42] and the preprocessed KEMAR head-related transfer functions from [43] are used to generate binaural RIRs (BRIRs) from the seven-channel SRIR recordings. For each position, a BRIR was chosen such that it places the virtual sound source in front of the listener. That means that, independent of the room configuration, speaker orientation, or listening position, the direct sound always arrives from the same direction. This was done to neutralize the influence of the direction of arrival of the direct sound, ensuring that only room acoustic differences of interest were relevant to listeners and their judgments. BRIRs were convolved with anechoic recordings of a female voice as a natural speech signal and a signal from a picked acoustic guitar as a transient music signal. Both signals have a length of 6 s . The spectrograms of the two signals are shown in Fig. 3.

For data collection, a discrimination test was conducted using an ABX interface. This test procedure features an explicit reference and two samples, where one of these samples is equivalent to the reference. In a typical ABX test, the task is to determine which sample matches the explicit reference. ABX tests are commonly used to assess the ability to detect subtle differences between two stimuli [44]. Because the perceived differences could be evaluated using specific attributes for each trial, participants were explicitly instructed to identify the sample that differed from the reference.

Refer to Fig. 2 for a description of how to construct an example ABX trial in the current listening test. The open reference is the yellow position-facing speaker SiL, and the available choices A and B are playing the yellow position-facing speaker BC and again the yellow position-facing speaker SiL. Participants had the task to identify the stimulus (A or B) that sounds *different* from the open reference. The ratio between correct identifications of the odd stimuli and the overall number of trials is called the hit ratio.

All positions within a listening test session are selected from the same room configuration. The one-wall and three-wall configurations each yield 46 possible combinations, whereas the five-wall configuration yields 44 possible combinations.

For each trial, there was an additional task: participants were asked to select attributes explaining the perceived differences (if applicable) and rate them on a six-point Likert scale. The attributes were: *coloration*, *loudness*, *distance*, *direction*, *reverberance*, *externalization*, *apparent source width*, and *ease of localization*. All attributes, a group, or just one attribute could be selected according to their relevance. For example, if only a small difference was perceivable and this could only be accounted to *coloration*, participants should only rate this attribute with a 1. If they, for example, perceive a strong difference in *direction* and *reverberance*, they could rate both attributes with a 4.

The listening test presented here was performed online, whereby the publicly available framework AVRateVoyager [45] was used. It was required to adjust some of the included templates to match as well as possible the laboratory test from [34]. The system was hosted on a university server with a high bandwidth connection. In contrast to the laboratory test from [34], however, the auralization was static due to the absence of a tracking system at the participant's place. Furthermore, the listening room is not controlled and is individual for each participant. This leads to a lower sensitivity of the test with regard to the detection of differences. However, considerably more individual tests can be carried out in the online test. By random assignment, each participant had to rate 40 ABX-like trials from one of the available room configurations (see Fig. 1) with either the voice or the guitar signal. Every test run only featured one audio signal and only positions within one room configuration.

Before the test, the participants were asked to complete a survey with general questions about age, gender, experience with audio perception tests, existing hearing impairments, their environment, and used hardware. This was followed by introducing the test, setting the task, and describing the attributes. The participants were also asked to adjust the playback volume using a voice signal. The signal was played at three different volumes. The first volume was set to be almost too loud, the second to be comfortable, and the third to be barely audible. The calibration signal had to be played back at least once.

4 RESULTS

A total of 120 participants took part in the online test. The responses of the participants to the initial survey are summarized in Table 1.

Table 1. Results of the participant survey conducted before the listening test, based on responses from 120 participants.

Condition	Percentage
Gender	
Male	62.5
Female	35.0
Diverse	0.83
Not specified	1.67
Age	
<18	0.0
18 to 24	10.83
25 to 29	13.33
30 to 39	33.33
40 to 49	20.83
50 to 59	20.0
60 to 69	1.67
70+	0.0
Experience	
Previous experience	45.83
No previous experience	54.17
Environment	
Alone in a quiet room	92.5
Some noise and distractions	5.83
Significant noise and distractions	1.67
Device	
Laptop/PC	88.33
Tablet	0.83
Phone	1.67
Data not captured	9.17
Headphones type	
Over-ear open/half-open	11.67
Over-ear closed	38.33
In-ear	39.17
Don't know	1.67
Data not captured	9.17
Headphones connection	
Wired headphones	65.83
Wireless headphones	31.67
Don't know	2.5

The listening test was planned to last between 30 and 40 minutes. The median completion time among participants was 28 minutes. The lower quartile (25%) was 22 minutes, and the upper quartile (75%) 45 minutes. Some participants took longer, but they also took breaks during the test. Although all other participants completed the entire test, two dropped out after 19 and 16 comparisons, respectively. Nevertheless, they were included in the analysis. Of the 120 participants, 32 rated the combination in HL01W, whereby the participants were evenly distributed between the two signals. 47 participants rated the comparison in HL03W, of which 24 rated the speech signal and 23 the guitar signal. In the room configuration HL05W, 41 participants gave ratings, 21 of them for the speech signal and 20 for the guitar signal.

The following evaluation focuses exclusively on the ABX test and the modeling of the hit ratio. Consequently, the analysis of the attribute ratings is not included in this publication.

4.1 Evaluation of the Hit Ratio

This section analyzes the hit ratio. A prediction model is created based on the subjects' results and the acoustic parameter differences. The overall hit ratio of the online test across all comparisons is 71.9%. For an ABX-test with 40 trials, the threshold for above-chance performance at a 95% confidence level is approximately 25 correct responses, corresponding to a hit ratio of about 63%.

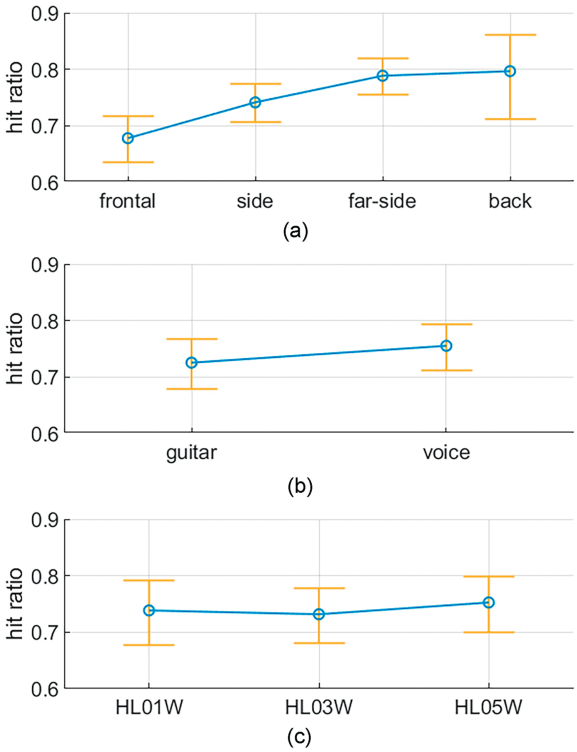


Fig. 4. Fixed effects plots. Estimated hit ratio (blue) with 95% confidence intervals (yellow) as a function of listening position relative to the loudspeaker (a), signal type (b), and room configuration (c).

4.1.1 Analysis

First, the GLMM is used to analyze whether there are any dependencies between signal type, room configuration, or listener position in relation to the loudspeaker orientation ($\mathrm{sr}_{\mathrm{orientation}}$). For this purpose, as shown in the model formula in Eq. (3), these three characteristics are considered to be fixed effects, and the participants are considered to random effect. Because the ratings from the listening test (hit/misses) are binomially distributed scores, the GLMM assumes a binomial distribution:

$$\text{hit ratio} \sim 1 + \text{signal} + \text{room} + \mathrm{sr}_{\mathrm{orientation}} + (1|\text{Participant}). \quad (3)$$

The GLMM is based on 4,753 observations, with seven fixed effect coefficients and 120 random effect coefficients. The covariance parameter is 1 and reflects the variance structure of the random intercept (1|Participant).

Fig. 4 shows the fixed effects plot. With regard to the listening position in relation to the loudspeaker [Fig. 4(a)], the probability of being able to hear a difference is lower directly in front of the loudspeaker than at the side and far-side positions. Behind the loudspeaker, the probability of hearing a difference is also higher on average, but the

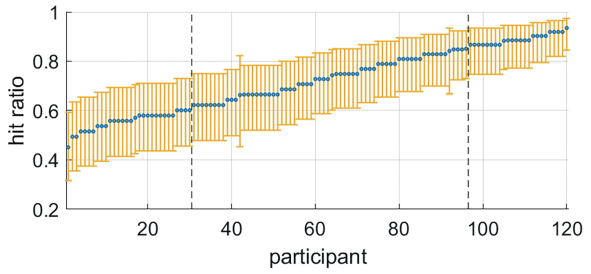


Fig. 5. Random effects plot. Estimated hit ratio (blue) with 95% confidence intervals (yellow) as a function of participants, sorted by their individual estimates. Range of participants whose random effects do not significantly differ from the intercept ($p > 0.05$) is between the dashed lines.

confidence interval is larger here. A possible reason for this is the fact that directional loudspeakers were used; thus, the direct sound could mask small room acoustic variations at frontal directions. This would support the results from [46].

This can also be seen in the t test values of the fixed effects. The intercept, corresponding to the *far-side* position, yielded a significant effect ($t = 1.18$; $p < 0.001$). Both the *front* ($t = -6.404$; $p < 0.001$) and *side* ($t = -3.126$; $p = 0.002$) positions differed significantly from the intercept, indicating substantial effects for these positions. In contrast, the *back* position ($t = 0.242$; $p = 0.808$) did not differ significantly from the intercept. In Fig. 4(b), a slight tendency can be seen that differences with the speech signal are more perceptible, but this is not significant ($t = 0.939$; $p = 0.348$). As can be seen in Fig. 4(c), there is also no significant difference between the room configurations.

This is also reflected in the analysis of variance for linear mixed effects. The F statistic for listening position relative to the loudspeaker is $F = 14.49$, $p < 0.001$; for the signal type $F = 0.88$, $p = 0.348$; and for the room configurations $F = 0.22$, $p = 0.80$. The significant F statistic for the intercept ($F = 54.42$; $p < 0.001$) indicates that the overall mean response when all predictors are at their reference levels is significantly different from zero.

For analyzing the random effects in more detail, the following model formula is used:

$$\text{hit ratio} \sim 1 + (1|\text{Participant}). \tag{4}$$

The model formula allows for the examination of differences between participants, independent of other fixed effects. Fig. 5 depicts the predicted hit ratio, along with its respective confidence intervals, as a function of the participants. The probabilities are presented in ascending order. The dashed vertical lines serve to divide the participants into three groups.

The p values from the t tests for the random effects, indicated by the dashed lines, range from 0.05 to 1. Within this interval, none of the participants significantly deviate from the intercept. In contrast, p values outside this range are less than or equal to 0.05, indicating that these participants differ significantly from the intercept at the 0.05 significance level.

The group of participants enclosed by the dashed lines represents the general or average listeners, comprising 55% of all participants. The group situated to the left of the aforementioned group includes participants who perceive fewer or no differences, corresponding to 25% of the total number. It is also probable that this group includes participants who guessed at the task or did not understand it. The group of participants with the highest hit probabilities is located in the third group, on the far right, that includes the remaining 20% of all participants.

At this point, it should be noted that the authors are critical of such a hard distinction and that it serves only as a guide. It is, rather, assumed that the transitions are smooth. Nor can any statement be made about the quality of the test subjects on the basis of these values. An interesting observation is the formation of distinct groups of participants with similar predicted probabilities and confidence intervals. Two participants stand out within their groups because they fall within the mean but have wider confidence intervals. This is because these two participants did not fully complete the test, resulting in fewer ratings.

Fig. 5 clearly shows the differences between the participants, which motivates the use of GLMM. This approach captures individual variability via random effects and ensures that predictions are based only on fixed effects, thus reflecting general trends rather than participant-specific differences.

4.1.2 Prediction Models

A prediction model for the hit ratio in relation to the acoustic parameters is now to be developed. For this purpose, the absolute parameter differences between the stimuli to be compared are used as fixed effects. Using the absolute difference offers the advantage of being independent of whether a stimulus is assigned as reference or comparison. The parameters used at this point are EDT, DT_{20}, C_{50}, DRR, and T_{S}. The acoustic parameters were chosen to capture different aspects of room acoustics, such as room-dependent (geometries, materials) and room-independent properties (source-receiver relations, source directivity). Broadband parameters were used instead of individual octave bands to minimize collinearities, maintain low model complexity, and enhance result interpretability.

Fig. 6 provides an overview of these parametric data, illustrating the broadband parameters employed in the initial model for all listening test positions in relation to the room configuration. The model formula used, which is the basis for the following considerations, can be seen in Eq. (5):

$$\begin{aligned}\text{hit ratio} \sim{}& -1 + \mathrm{EDT}_{|a-b|} + \mathrm{DT}_{20_{|a-b|}} + \mathrm{C}_{50_{|a-b|}} \\ &+ \mathrm{DRR}_{|a-b|} + \mathrm{T}_{S_{|a-b|}} + (1|\text{Participant}),\end{aligned} \tag{5}$$

where -1 indicates that no intercept is used, and $|a - b|$ reflects that the absolute parameter differences between the compared stimuli are used in the model. The contributions of the individual parameters will now be examined using the results of the t test. For the absolute differences of EDT, a t value of $t = 0.73887$ with a p value of $p = 0.46003$ was obtained. This indicates that EDT does not

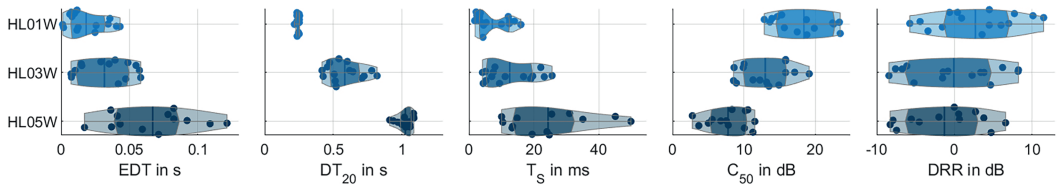


Fig. 6. Distribution of the broadband room acoustic parameters for the positions used in the listening test as a function of the room configuration. The parameters used in the model according to Eq. (5) are shown.

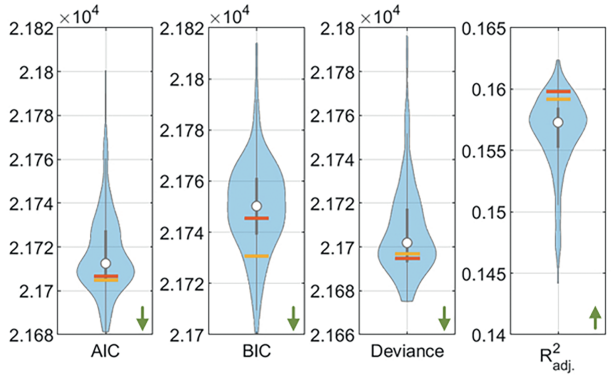


Fig. 7. Distributions of the quality measures AIC, BIC, deviation and R^2_{adj} for 255 models based on different parameter combinations as fixed effect. Red line is the model according to Eq. (5); yellow line is the model according to Eq. (6); green arrows indicate direction of model improvement for each criteria.

make a significant contribution to the prediction. The same applies to DT_{20}, with values of $t = 1.1803$ and $p = 0.23793$. The other absolute parameter differences T_S ($t = 4.055$; $p < 0.00001$), C_{50} ($t = 2.4641$; $p < 0.02$), and DRR ($t = 4.8059$; $p < 0.00001$) have a strong (T_S, DRR) to slight (C_{50}) contribution.

The values of the estimated fixed effects and confidence interval limits do not make sense when comparing the parameters because the parameters differ in their units and therefore also in their value ranges. This can be remedied by standardizing the acoustic parameter data. This does not change the results of the *t* test. Therefore, the authors refrain from standardizing the acoustic parameters. It would also mean that the data would have to be repeatedly converted for prediction and further analysis.

It can be seen that not all of the parameters used contribute to the prediction. For this reason, which parameter configuration produced the "best" prediction model was investigated. Therefore, 255 parameter combinations were compared on the basis of the room acoustic parameters EDT, DT_{20}, D_{50}, C_{50}, C_{80}, DRR, G_{rel}, and T_S.

As noted in SEC. 1, there are several measures available to assess the goodness of fit. Only four criteria are considered below: the deviance, the AIC, the BIC, and the R^2_{adj}.

Fig. 7 shows the distributions of the four quality measures of the 255 models with median and quartile range. The red line marks the values of the model according to Eq. (5). It can be seen that for deviance, AIC, and BIC, the selected model is below the median but above the lower quartile. For R^2_{adj}, on the other hand, the model is above the upper quartile. This shows that the chosen parameter combination describes the hit ratio relatively well with regard to R^2_{adj}. However, the model complexity could be reduced by limiting the number of parameters used, particularly with regard to the BIC.

The lowest values for deviance, AIC, and BIC are achieved with the parameter combination DRR and C_{80}. In contrast, this combination is ranked only 231 out of 255 in terms of R^2_{adj}. Conversely, the highest R^2_{adj} value is obtained with the parameter combination EDT and T_S. However, this configuration ranks between 244 and 246 for the other three quality measures. This clearly shows that a model performing best in one quality measure can perform poorly in others.

For this reason, an attempt was made to identify a parameter combination that achieves a compromise between the different quality measures. To this end, the following methodology was applied in this study: initially, the analysis focused on the quality measures AIC, BIC, and R^2_{adj}. Deviance was excluded because it mostly delivers redundant information compared with AIC and BIC and would make the parameter search more complex. Based on these three quality measures and a total of 255 models, an optimal balance between minimizing AIC and BIC and maximizing R^2_{adj} was sought using a threshold-based analysis in which the metrics were incorporated with equal weighting. This analysis involved two steps:

(1) *Studentizing* the data to standardize metrics (mean = 0, variance = 1), ensuring scale-invariant comparisons.
(2) *Filtering out models* where AIC and BIC > 0 and $R^2_{adj} < 0$, thereby eliminating poorly performing models.

Iterative studentization and filtering of the reduced data set yielded a final model representing a compromise among the selected criteria. This results in the parameter combination DRR and C_{80} and T_S, and thus the following model formula is obtained:

$$\text{hit ratio} \sim -1 + C80_{|a-b|} + DRR_{|a-b|} + TS_{|a-b|} + (1|\text{Participant}). \quad (6)$$

The quality measures for this model are marked with a yellow line in the Fig. 7. Whereas deviance and R^2_{adj} only deteriorate slightly and the AIC barely changes, the BIC improves significantly, as the number of fixed effects is

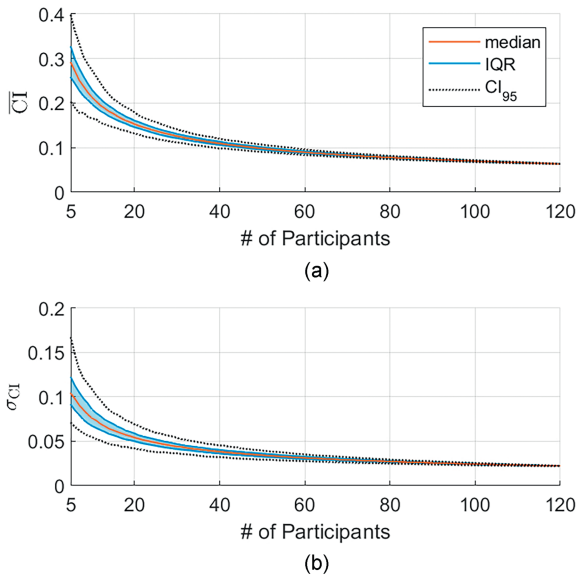


Fig. 8. Mean value (a) and standard deviation (b) of the confidence interval of the prediction as dependent on the number of participants for 500 random repetitions. Model used is according to Eq. (6). CI_{95} = 95% confidence interval; IQR = interquartile range.

reduced in comparison. The t test statistic shows that T_S ($t = 4.6712$; $p < 0.00001$), C_{80} ($t = 3.8889$; $p = 0.0001$), and DRR ($t = 6.0234$; $p < 0.00001$) contribute to the modeling.

4.1.3 Prediction Results

First, the models based on Eqs. (5) and (6) are compared. An examination of the confidence intervals reveals that the model based on Eq. (6) ($\overline{CI} = 0.0637$, $\sigma_{CI} = 0.0222$) exhibits narrower and less variable intervals than the model based on Eq. (5) ($\overline{CI} = 0.0781$, $\sigma_{CI} = 0.0294$). It should be noted that the differences between the models are relatively small. The predictions are based solely on fixed effects and do not account for individual participants. This means that they refer to a hypothetical participant who was not involved in the experiment. Moreover, the width of the confidence intervals also depends on the number of observations and is therefore inherently linked to the number of participants.

Fig. 8 shows how the confidence interval depends on the number of participants. The procedure to create this plot was as follows:

1. Selection of five random participants as a starting point.
   a. Model created according to model formula by Eq. (6).
   b. Prediction for a general participant based on all stimuli used.
   c. Calculate the mean $\overline{CI}$ and the standard deviation σ_{CI} of the confidence intervals of the prediction.
   d. Randomly add the ratings of a further participant.
2. Repeat points a to d until the ratings of all participants are included.

This was repeated a total of 500 times, so that 500 values are included in Fig. 8 for each count of participants.

Table 2. Mean values of the evaluation measures of the stratified fivefold cross-validation with 100 repetitions.

Accuracy	F1 score	Brier score	AUC
0.71	0.82	0.19	0.68

Fig. 8 shows the median, the interquartile range and the 95% confidence interval for the mean value of the confidence intervals $\overline{CI}$ [Fig. 8(a)] and the standard deviation of the confidence intervals σ_{CI} [Fig. 8(b)]. It can be seen that as the number of participants increases, $\overline{CI}$ and σ_{CI} become smaller and approach a minimum of 0.06 and 0.02, respectively.

To evaluate the predictive performance of the model according to Eq. (6), a stratified fivefold cross-validation with 100 repetitions is performed. The accuracy, the F1 score [47], the Brier score [48], and the area under the receiver operating characteristic curve (AUC) [49] are considered as evaluation measures. Because the initial data is binary, the predicted values used to calculate accuracy and the F1 score must also be binary. A threshold value of 0.5 is assumed, meaning that all predicted values >0.5 are interpreted as 1 and values ≤0.5 are interpreted as 0. Table 2 shows the mean values of the measures.

The accuracy indicates the proportion of correctly classified cases. Here, the model achieves an accuracy of 0.71, exceeding the baseline of 0.50 from random guessing. The F1 score is the harmonic mean of precision and recall. A value of 0.82 can be interpreted as very good and shows that precision and recall are balanced. The Brier score is a measure of the squared deviation of the prediction from the actual value and can assume values between 0 and 1, whereby small values are aimed for. A value of 0.19 can be interpreted as good. The AUC as a quality measure for the prediction accuracy can assume a maximum of 1, whereby values >0.7 are desired. The AUC of 0.68 is still acceptable, but shows potential for improvement.

Next, the prediction of perceptible differences within room configurations will be considered to find possible similar perceptual areas. For this purpose, the Fig. 9 is considered. Two different possibilities of prediction based on the model are presented. For both, the color scale ranges from 0.5 to 1.0 with a separation into four sections. This serves visualization purposes only.

The upper part of the figure will be discussed first. The grids of the three room configurations are shown there. The reference point is the green triangle at the position P(0, 0). P(0, 0) is compared with every other position on the grid relative to the loudspeaker BC. This can be equated to a virtual jump from P(0, 0) to any other position in the room. Regions become visible as the model predicts different probabilities for perceptible differences with respect to P(0, 0). The confidence values are not considered at this point. It is easy to see how the areas differ depending on the room configurations.

In the white area, the probability of perceiving a difference from the reference point is less than 65%. The area is

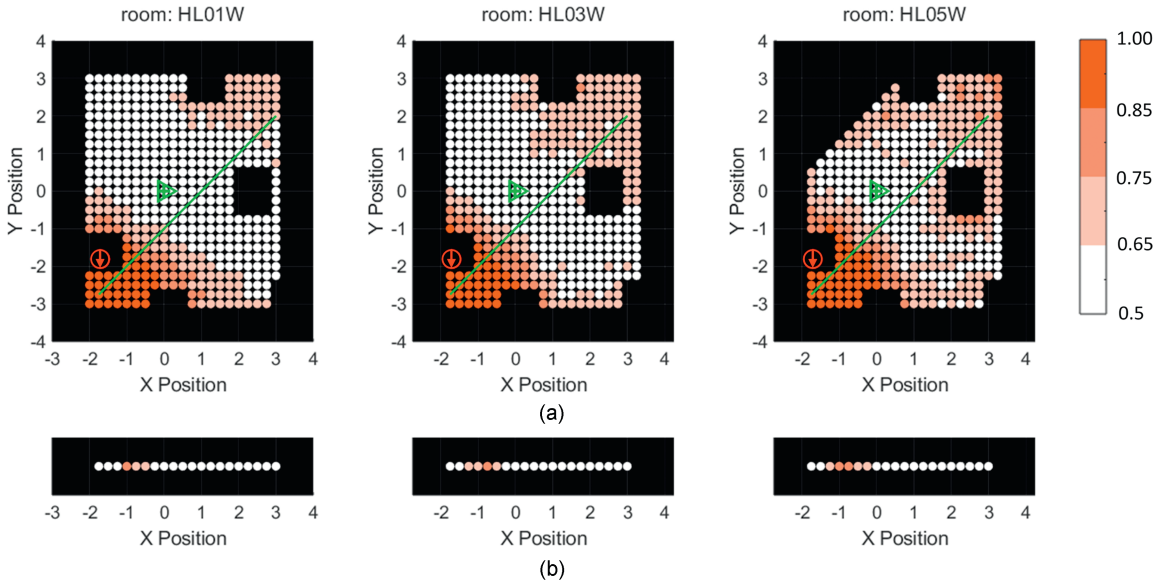


Fig. 9. (a) Prediction of similar perceptual areas in relation to the reference listening position $P(0, 0)$ (green triangle). (b) Prediction of perceived differences along the green line in the grid [from $P(-1.75, -2.75)$ to $P(3.00, 2.00)$]. For a more compact presentation, the prediction values are projected onto the corresponding x-coordinates. Considerations for loudspeaker BC: red circle with red arrow, indicating the orientation of the speaker. Model used is according to Eq. (6).

therefore very similar in terms of perception. In contrast, it can be assumed that the red areas differ from the reference point in terms of perception. There is a clear dependence on the orientation of the loudspeaker BC, which is described by the acoustic parameters.

Note that in addition to room configuration and source orientation, the results also depend on the position of the source and the reference position. If a different reference position is selected, the predictions of the grid will change accordingly. Similarly, the predictions change if a different source is considered while the listener's position remains constant. This covers, in particular, the effects that would also occur in a real listening environment.

The lower part of the Fig. 9 refers to the green lines in each grid. The model describes the difference to the next point along the line from point P(−1.75, −2.75) to P(3.00, 2.00). This gradual change describes a movement along the line rather than a jump between points. The predictions of the three lines are very similar. In the range P_X from −1.75 to 0.75, the model predicts a noticeable change. This range describes the departure of the directional direct sound from the loudspeaker. In the other areas along the line, no noticeable change is predicted.

The last part of this section examines the extent to which the predicted differences match the corresponding JND of the individual parameters. The parameters DRR, T_S, and C_{80} are considered here because these parameters were used in the model according to Eq. (6). In the following observations, only the individual absolute parameter differences are considered as fixed effects to examine their effects on the prediction. As before, the participants are included as a random effect, and the intercept is excluded. The results for each of the three parameters are shown in the Fig. 10. Because the JNDs depend in part on the relative values of the parameters, the parameter values are presented on the x-axis, and the changes in the parameters are shown on the y-axis. The individual points represent the stimuli tested in the test. They are arranged according to their mean parameter value and their absolute difference. The hit ratio predicted by the model is marked in color. Additionally, the areas with a 55%, 65%, and 75% hit ratio and their confidence intervals are shown.

Finally, the JND boundaries according to the corresponding reference are marked with a red line. For the JND of the DRR, reference is made to Larsen et. al. [13]. As mentioned above, the model cannot capture the nonlinearity of the JND as a function of the DRR. However, it can be seen that the predicted guess probability is below the minimum JND. The prediction of the model for this point is below 0.6. Additionally, the stimuli used in the test are relatively uniformly distributed in terms of DRR and change in DRR.

The same applies to C_{80}. For JND, reference is made to Martellotta [14]. The predicted guess probability is below the JND. The model's prediction for the JND is approximately 0.6. The data for C_{80} and the absolute change in C_{80} are also uniformly distributed.

For the T_S, the JND also refers to [14]. The stimuli are not evenly distributed with regard to the T_S and their absolute difference. The majority of the stimuli are in the range of the guessing probability. Nevertheless, the predicted probabilities agree with the JND. However, it is important to note that these JND values were derived from a listening test [14] in acoustic environments (churches) with 2 to 6-s RTs. In the current study, the room is much smaller and the RTs are in the range of 0.22 to 1.2 s. Nevertheless, the derivation of the JND according to [14] for lower T_S values seems plausible. It can be concluded that changes of T_S less than 2.5 ms cannot be perceived for $T_S \leq 20$ ms.

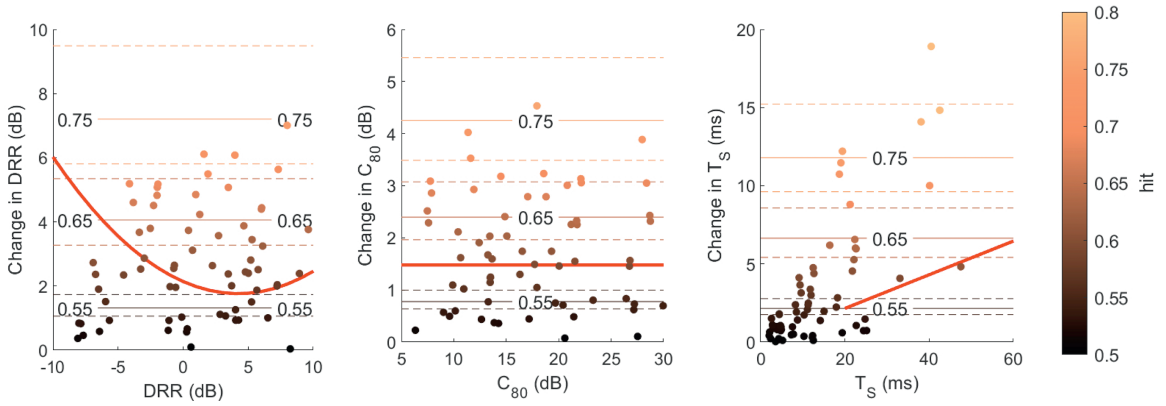


Fig. 10. Comparison of model prediction and JNDs for DRR, C_{80}, and T_S. Red line: JND according to Larsen et. al. [13] (DRR) and Martellotta [14] (C_{80}, T_S). Each dot corresponds to a comparison between two positions.

5 DISCUSSION

On the basis of the presented listening test, a perceptually motivated prediction model was developed. This model predicts the probability of perceiving a difference between two stimuli based on differences in room acoustic parameters.

The collection of perceptual differences by means of an online study can be viewed critically. Especially, the uncontrolled conditions and the static auralization are potentially limiting factors in the online test. As the authors intended, it was possible to involve many participants who had not previously participated in such listening tests, resulting in a model that is as widely applicable as possible to the general population. However, there is also an increased risk that participants will not perform the test correctly or honestly. This is compensated for by the number of participants and by the representation of these characteristics in the random effects of the model.

The prediction model was trained on the basis of the results of the listening test, whereby identical source-receiver arrangements were compared within a room configuration. The perceived differences therefore result from the differences between the arrangements in the room. Because the model only considers whether the choice was correct or incorrect, it can only predict the hit ratio. No statement can be made about the root cause of those perceived differences. Accordingly, the prediction of the similar perceptual areas as shown in Fig. 9 can only predict the probability that a difference will be heard.

Additionally, "jumping" between two points in the room with respect to the same source was not part of the actual experiment. Furthermore, the predicted probabilities of perceiving a difference only refer to the absolute change in the parameters used as fixed effects in the model, in this case, DRR, C_{80}, and T_S. This means that changes in directional perception or loudness differences are not taken into account, which was intended by the authors through the test design. However, due to the parameters used, the predicted differences are strongly dependent on the orientation of the source.

The prediction along a line, as shown in Fig. 9, is intended to check whether differences between neighboring positions are perceptible. The probability of perceiving an acoustical difference is highest when crossing in front of the loudspeaker, where on-axis radiation occurs in directional sources. This is mainly due to the attenuation of the direct sound and, thus, mainly a characteristic of the sound source. The remaining differences along the lines are so small that the model predicts no perceptible difference.

About the consideration of the model predictions and the JNDs of DRR, C_{80} and T_S, it can be seen that the predicted rate probabilities are below the JNDs, and thus the model does not contradict the JNDs. It can be recognized that with increasing differences, the probability of perceiving a difference increases. Because the JNDs of DRR and T_S depend on the value range, a direct comparison of the model prediction remains difficult inasmuch as this relationship is not captured in the model.

Based on the analyses conducted, it is assumed that the model can be generalized for rooms of similar size and geometry as well as reverberation times. For larger and acoustically more complex environments, generalizability cannot currently be assumed and should be confirmed in further studies.

6 CONCLUSION

This paper presents an approach for modeling the perceptibility of acoustic changes in audio-augmented reality using a GLMM. Acoustic measurements were conducted in three different room configurations, and an online listening test was carried out in which participants compared identical source-receiver combinations across varying acoustic environments using an ABX-like test design. Additionally, up to eight attributes could be rated on a 6-point Likert scale if participants were able to explain the perceived difference.

The focus of the study lies in the development and evaluation of a prediction model based on the ABX-like listening test results. In total, 4,753 individual ratings were collected from 120 participants. The GLMM was used to analyze the influence of signal type, room configuration, and source-receiver relationship on the perceptual detection rate, with only the source-receiver relationship showing a statistically

significant effect. Participants were modeled as a random effect to account for individual differences in perception.

The paper further outlines how GLMMs can be constructed and compared based on the test results and room acoustic parameters. One aspect of the analysis includes estimating the number of participants required to achieve reliable predictive accuracy. To assess the model's performance, a fivefold cross-validation with 100 repetitions was conducted. The resulting Brier score of 0.19 indicates good predictive accuracy, whereas the AUC value of 0.67 suggests there is still room for improvement.

The model enables the identification of perceptually similar acoustic conditions depending on source location and room configuration. Finally, it is shown that predictions based on differences in specific acoustic parameters (namely DRR, C_{80}, and T_S) are consistent with known JNDs reported in the literature.

The results of this work and in [34, 33] show that the use of GLMM for analysis and modeling can be useful. In particular, the explicit exclusion of random effects from forecasting models offers added value compared with general linear models. The methods utilized in this paper represent only a small part of the possibilities offered by GLMMs. For example, they offer further possibilities for integrating both fixed and random effects into a model as a function of each other via a model formula. Additionally, depending on the application, other data, such as data from auditory models, signal type, signal characteristics, or other metadata, can also be used as fixed effects in a model. More specifically, spatial information based on the SRIR measurements could be included. It is also possible to use acoustic parameters across different octave bands but only in conjunction with dimension reduction techniques, such as principal component analysis or FA, to avoid linear dependencies in the fixed effects.

The analysis of the quality attributes could not be included in this paper, but more detailed predictions are expected by including such information. A comparison of these attributes between the laboratory test and the online test will help to understand the differences between expert and naive listeners.

Predicting the perceived difference when moving in a six-degrees-of-freedom scenario seems to be a future challenge because it likely involves the consideration of movement speed and patterns and the availability of references as well as the mechanisms of auditory memory [50]. The first logical step in this direction will be the perceptual validation of the predictions along the walking line.

7 ACKNOWLEDGMENT

The presented work is funded within the projects: "ISOP-ERARE," funded by Reality Labs Research at Meta and "Multiparties" in the framework "KMU-innovativ: Interactive Technologies for Health and Quality of Life" program of the German Federal Ministry of Education and Research (BMBF). KMU stands for "Kleine und mittlere Unternehmen" (small and medium-sized enterprises).

8 REFERENCES

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

[2] A. Neidhardt, C. Schneiderwind, and F. Klein, "Perceptual Matching of Room Acoustics for Auditory Augmented Reality in Small Rooms - Literature Review and Theoretical Framework," *Trends Hear.*, vol. 26 , pp. 1–22 (2022 Jan.). https://doi.org/10.1177/23312165221092919.

[3] M. Hodgson, L. Scannell, and J. Higgins, "Acoustics of Small Rooms," *J. Acoust. Soc. Am.*, vol. 137, no. 3, pp. 1609–1610 (2015 Mar.). https://doi.org/10.1121/1.4913275.

[4] K. Brandenburg, F. Klein, A. Neidhardt, U. Sloma, and S. Werner, "Creating Auditory Illusions With Binaural Technology," in J. Blauert and J. Braasch eds., *The Technology of Binaural Understanding*, Modern Acoustics in Signal Processing, pp. 623–663 (Springer International Publishing, Cham, Switzerland, 2020). https://doi.org/10.1007/978-3-030-00386-9_21.

[5] C. McCulloch, S. Searle, and J. Neuhaus, *Generalized, Linear, and Mixed Models*, Wiley Series in Probability and Statistics (John Wiley & Sons, Hoboken, NJ, 2011).

[6] J. Bruin, "Introduction to Generalized Linear Mixed Models," UCLA: Statistical Consulting Group, https://stats.oarc.ucla.edu/other/mult-pkg/introduction-to-generalized-linear-mixed-models/ (accessed February 25, 2025).

[7] P. McCullagh, *Generalized Linear Models* , 2nd ed. (Chapman and Hall/CRC, Boca Raton, FL, 1989).

[8] G. N. Wilkinson and C. E. Rogers, "Symbolic Description of Factorial Models for Analysis of Variance," *J. R. Stat. Soc. Ser. C Appl. Stat.*, vol. 22, no. 3, pp. 392–399 (1973 Nov.). https://doi.org/10.2307/2346786.

[9] K. Backhaus, B. Erichson, S. GenslerR. Weiber, and W. Thomas, *Multivariate Analysis: An Application-Oriented Introduction*, 2nd ed. (Springer Gabler, Wiesbaden, Germany, 2023). https://doi.org/10.1007/978-3-658-40411-6.

[10] P. K. Ozili, "The Acceptable R-Square in Empirical Modelling for Social Science Research," *Soc. Res. Methodol. Publ. Res.* (2022 Jun.). http://dx.doi.org/10.2139/ssrn.4128165.

[11] DIN, "Akustik - Messung von Parametern der Raumakustik - Teil 1: Aufführungsräume (ISO 3382-1:2009); German Version EN ISO 3382-1:2009," (2009 Jun.).

[12] DIN, "Akustik - Messung von Parametern der Raumakustik - Teil 2: Nachhallzeit in Gewöhnlichen Räumen (ISO 3382-2:2008); German version EN ISO 3382-2:2008," (2008 Sep.).

[13] E. Larsen, N. Iyer, C. R. Lansing, and A. S. Feng, "On the Minimum Audible Difference in Direct-to-Reverberant Energy Ratioa)," *J. Acoust. Soc. Am.*,

vol. 124, no. 1, pp. 450–461 (2008 Jul.). https://doi.org/10.1121/1.2936368.

[14] F. Martellotta, "The Just Noticeable Difference of Center Time and Clarity Index in Large Reverberant Spaces," *J. Acoust. Soc. Am.*, vol. 128, no. 2, pp. 654–663 (2010 Aug.). https://doi.org/10.1121/1.3455837.

[15] F. Klein, L. Treybig, C. Schneiderwind, S. Werner, and T. Sporer, "Just Noticeable Reverberation Difference at Varying Loudness Levels," presented at the *154th Convention of the Audio Engineering Society* (2023 May), paper 67. http://www.aes.org/e-lib/browse.cfm?elib=22092.

[16] H.-P. Seraphim, "Untersuchungen über die Unterschiedsschwelle Exponentiellen Abklingens von Rauschbandimpulsen," *Acta Acust. united Ac.*, vol. 8, no. 4, pp. 280–284 (1958 Jan.).

[17] M. G. Blevins, A. Buck, Z. E. Peng, and L. Wang, "Quantifying the Just Noticeable Difference of Reverberation Time With Band-Limited Noise Centered Around 1000 Hz Using a Transformed Up-Down Adaptive Method," in *Proceedings of the International Symposium on Room Acoustics* (Toronto, Canada) (2013 Jun.).

[18] I. Frissen, B. F. Katz, and C. Guastavino, "Effect of Sound Source Stimuli on the Perception of Reverberation in Large Volumes," in *Proceedings of the International Symposium on Computer Music Modeling and Retrieval*, pp. 358–376 (Copenhagen, Denmark) (2009 May).

[19] F. Klein, S. V. A. Garí, J. M. Arend, and P. W. Robinson, "Towards Determining Thresholds for Room Divergence: A Pilot Study on Detection Thresholds," in *Proceedings of the Immersive and 3D Audio: From Architecture to Automotive (I3DA)*, pp. 1–7 (Bologna, Italy) (2021 Sep.).

[20] N. Meyer-Kahlen, S. V. A. Garí, I. Ananthabhotla, and P. Calamia, "A Two-Dimensional Threshold Test for Reverberation Time and Direct-to-Reverberant Ratio," in *Proceedings of the Immersive and 3D Audio: From Architecture to Automotive (I3DA)*, pp. 1–8 (Bologna, Italy) (2023 Sep.). https://doi.org/10.1109/I3DA57090.2023.10289365.

[21] S. A. Wirler, N. Meyer-Kahlen, and S. J. Schlecht, "Towards Transfer-Plausibility for Evaluating Mixed Reality Audio in Complex Scenes," in *Proceedings of the AES International Conference on Audio for Virtual and Augmented Reality* (2020 Aug.), paper 3-4. http://www.aes.org/e-lib/browse.cfm?elib=20881.

[22] N. Meyer-Kahlen, S. J. Schlecht, and T. Lokki, "Clearly Audible Room Acoustical Differences May Not Reveal Where You Are in a Room," *J. Acoust. Soc. Am.*, vol. 152, no. 2, pp. 877–887 (2022 Aug.). https://doi.org/10.1121/10.0013364.

[23] N. Meyer-Kahlen, S. Amengual Garí, T. McKenzie, S. J. Schlecht, and T. Lokki, "Transfer-Plausibility of Binaural Rendering With Different Real-World References," in *Proceedings of the 48th Annual German Conference on Acoustics (DAGA)* (Stuttgart, Germany) (2022 Mar.).

[24] N. Meyer-Kahlen, S. J. Schlecht, S. V. A. Garí, and T. Lokki, "Testing Auditory Illusions in Augmented Reality: Plausibility, Transfer-Plausibility, and Authenticity," *J. Audio Eng. Soc.*, vol. 72, no. 11, pp. 797–812 (2024 Nov.).

[25] T. McKenzie, N. Meyer-Kahlen, and S. J. Schlecht, "The Role of Source Signal Similarity in Distinguishing Between Different Positions in a Room," in *Proceedings of the AES International Conference on Spatial and Immersive Audio* (2023 Aug.), paper 25.

[26] H. Helmholz, I. Ananthabhotla, P. T. Calamia, and S. V. Amengual Gari, "Towards the Prediction of Perceived Room Acoustical Similarity," in *Proceedings of the AES International Conference on Audio for Virtual and Augmented Reality* (2022 Aug.), paper 20. http://www.aes.org/e-lib/browse.cfm?elib=21850.

[27] B. Burnett, A. Neidhardt, Z. Cvetković, H. Hacıhabiboğlu, and E. De Sena, "User Expectation of Room Acoustic Parameters in Virtual Reality Environments," in *Proceedings of the Immersive and 3D Audio: From Architecture to Automotive (I3DA)*, pp. 1–10 (Bologna, Italy) (2023 Sep.). https://doi.org/10.1109/I3DA57090.2023.10289314.

[28] J. Steffman and S.-A. Jun, "Perceptual Integration of Pitch and Duration: Prosodic and Psychoacoustic Influences in Speech Perception," *J. Acoust. Soc. of Am.*, vol. 146, no. 3, pp. EL251–EL257 (2019 Sep.).

[29] G. Reardon, G. Zalles, A. Genovese, P. Flanagan, and A. Roginska, "Evaluation of Binaural Renderers: Externalization, Front/Back and Up/Down Confusions," presented at the *144th Convention of the Audio Engineering Society*, (2018 May), paper 9989. https://aes2.org/publications/elibrary-page/?id=19506.

[30] A. Taghipour, S. Athari, A. Gisladottir, T. Sievers, and K. Eggenschwiler, "Room Acoustical Parameters as Predictors of Acoustic Comfort in Outdoor Spaces of Housing Complexes," *Front. Psychol.*, vol. 11 (2020 Mar.). https://doi.org/10.3389/fpsyg.2020.00344.

[31] Y. G. Redman, J. D. Glasner, D. D'Orazio, and P. Bottalico, "Singing in Different Performance Spaces: The Effect of Room Acoustics on Singers' Perception," *J. Acoust. Soc. Am.*, vol. 154, no. 4, pp. 2256–2264 (2023 Oct.). https://doi.org/10.1121/10.0021331.

[32] M. von Berg, P. Schwörer, L. Prinz, and J. Steffens, "Analysis of Physical and Perceptual Properties of Room Impulse Responses: Development of an Online Tool," in *Proceedings of the Forum Acusticum: 10th Convention of European Acoustics Association* (Turin, Italy) (2023 Sep.).

[33] L. Treybig, U. Sloma, S. Werner, J. Höbel-Müller, and A. Nürnberger, "Towards Perceptual Modeling in Room Acoustics - Connecting Acoustic Parameters With Auditory Perception," presented at the *156th Convention of the Audio Engineering Society*, (2024 Jul.), paper 10700. https://aes2.org/publications/elibrary-page/?id=22513.

[34] F. Klein, L. Treybig, S. V. A. Garí, and W. Stephan, "Modeling the Perceptibility of Intra-Room Acoustic Variations," presented at the *156th Convention of the Audio Engineering Society* (2024 Jul.), paper 10710. https://aes2.org/publications/elibrary-page/?id=22523.

[35] L. Treybig, F. Klein, G. Stolz, and S. Werner, "A High Spatial Resolution Dataset of Spatial Room Impulse Responses for Different Acoustic Room Configurations," (2024 Jan.). https://doi.org/10.5281/zenodo.10450779 (accessed February 25, 2025).

[36] L. Treybig, F. Klein, G. Stolz, S. Werner, and S. V. A. Garí, "A High Spatial Resolution Dataset of Spatial Room Impulse Responses for Different Acoustic Room Configurations," in *Proceedings of the 50th German Annual Conference on Acoustics (DAGA)* (Hanover, Germany) (2024 Mar.).

[37] L. Treybig, J. Höbel-Müller, S. Werner, and A. Nürnberger, "Acoustic Inter- and Intra-Room Similarity Based on Room Acoustic Parameters," in *Proceedings of Engineering for a Changing World: 60th Ilmenau Scientific Colloquium*, (Ilmenau, Germany) (2023 Sep.). https://doi.org/10.22032/DBT.58929.

[38] ITU-R, "Methods for the Subjective Assessment of Small Impairments in Audio Systems," Recommendation ITU-R BS.1116-3 (2015 Feb.).

[39] S. Tervo, A. Pätynen, J. and Kuusinen, and T. Lokki, "Spatial Dcomposition Method for Room Impulse Responses," *J. Audio Eng. Soc.*, vol. 61, no. 1/2, pp. 17–27 (2013 Jan.). http://www.aes.org/e-lib/browse.cfm?elib=16664.

[40] L. Treybig, S. Werner, F. Klein, and S. V. A. Garí, "Robust Reverberation Time Estimation for Audio Augmented Reality Applications," presented at the *154th Convention of the Audio Engineering Society* (2023 May), paper 10640. http://www.aes.org/e-lib/browse.cfm?elib=22047.

[41] H. Kuttruff, *Room Acoustics*, 5th ed. (CRC Press, Boca Raton, FL, 2009).

[42] S. Amengual Garí, W. O. Brimijoin, H. G. Hassager, and P. Robinson, "Flexible Binaural Resynthesis of Room Impulse Responses for Augmented Reality Research," in *Proceedings of the EAA Spatial Audio Signal Processing Symposium* (Paris, France) (2019 Sep.). https://doi.org/10.25836/sasp.2019.31.

[43] H. Braren and J. Fels, "A High-Resolution Head-Related Transfer Function Data Set and 3D-Scan of KEMAR," RWTH Aachen University (2020 Nov.). https://doi.org/10.18154/RWTH-2020-11307 (accessed April 12, 2023).

[44] D. L. Clark, "Ten Years of A/B/X Testing," presented at the *91st Convention of the Audio Engineering Society* (1991 Oct.), paper 3167. http://www.aes.org/e-lib/browse.cfm?elib=5549.

[45] S. Göring, R. R. R. Rao, S. Fremerey, and A. Raake, "AVRate Voyager: An Open Source Online Testing Platform," in *Proceedings of the IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–6 (Tampere, Finland) (2021 Oct.).

[46] C. Schneiderwind and A. Neidhardt, "Perceptual Differences of Position Dependent Room Acoustics in a Small Conference Room," in *Proceedings of the International Symposium on Room Acoustics (ISRA)* (Amsterdam, the Netherlands) (2019 Sep.). https://doi.org/10.18154/RWTH-CONV-240138.

[47] C. J. V. Rijsbergen, *Information Retrieval*, 2nd ed. (Butterworth-Heinemann, Oxford, UK, 1979).

[48] G. W. Brier, "Verification of Forecasts Expressed in Terms of Probability," *Mon. Wea. Rev.* vol. 78, no. 1, pp. 1–3 (1950 Jan.). https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

[49] A. P. Bradley, "The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms," *Pattern Recognit.*, vol. 30, no. 7, pp. 1145–1159 (1997 Jul.). https://doi.org/10.1016/S0031-3203(96)00142-2.

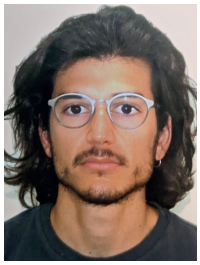
[50] F. Klein, T. Surdu, L. Treybig, and S. Werner, "The Ability to Memorize Acoustic Features in a Discrimination Task," *J. Audio Eng. Soc.*, vol. 71, no. 5, pp. 254–266 (2023 May).

THE AUTHORS


Lukas Treybig


Florian Klein


Stephan Werner


Sebastià V. Amengual Garí


Steve Göring

Lukas Treybig was born in 1988 in Zwickau, Germany. In 2014, he received an M.Sc. degree in media technology from Technische Universität Ilmenau (TU Ilmenau), Germany. From 2015 to 2020, he worked as a research assistant in the Audiovisual Technology Group at TU Ilmenau. Since 2021, he has been a research assistant in the Electronic Media Technology Group at the same university and is currently pursuing his doctorate. His research interests include the parametric and perceptual classification of room acoustics.

•

Florian Klein was born in 1986 in Bad Salzungen, Germany. He received his Dr.-Ing. in media technology from Technische Universität Ilmenau (TU Ilmenau). He is currently a postdoctoral researcher, and his main research area includes auditory adaptation processes and perception of room acoustics in the scope of spatial hearing. He is also working on solutions for six-degrees-of-freedom spatial sound rendering for augmented reality/virtual reality applications.

•

Stephan Werner was born in 1981 in Merseburg, Germany. Since 2020 he is the acting head of the Electronic Media Technology Group and senior researcher at Technische Universität Ilmenau (TU Ilmenau). He received his Dr.-Ing. (equivalent to a Ph.D.) from TU Ilmenau in the field of spatial audio and context-dependent quality parameters on the perception of externality and auditory event location. His focus of research is in the fields of spatial audio for augmented/mixed reality, psychoacoustics, quality evaluation, acoustic measurements, and analysis.

•

Sebastià V. Amengual Garí is currently a research scientist at Meta Reality Labs Research, working on room acoustics, spatial audio, and auditory perception. He received his Diploma in Telecommunications with a major in sound and image from the Polytechnic University of Catalonia (UPC), Barcelona, Spain, in 2014. He completed his Master's thesis at the Norwegian University of Science and Technology (NTNU), Trondheim, Norway. He received his doctoral degree (Dr.-Ing.) from the University of Music, Detmold, Germany, in 2017, focusing on the interaction of room acoustics and live music performance using virtual acoustic environments. His research interests lie at the intersection of audio, perception, and music.

•

Steve Göring is a computer scientist working in the Audiovisual Technology Group at Technische Universität Ilmenau (TU Ilmenau). Currently, his focus is on data analysis problems for video quality models and video streams. He finished his Ph.D. in 2022, with the focus of visual quality prediction using machine learning. Before he started in 2016 at the Audiovisual Technology Group, he was working at Bauhaus University Weimar in the Big Data Analytics Group. From 2008 to 2013, he studied at TU Ilmenau and graduated with a B.Sc. and M.Sc. in computer science. His specializations are data analytics/machine learning, video quality, and distributed communication/information systems.