Evaluation of Perceptual Accuracy in Simulated Room Impulse Responses

Designing and Implementing a Subjective Testing Methodology for the Perceptual Evaluation of Simulated Room Impulse Responses

Bendik Christensen

Supervisor(s): Jorge Martinez[1], Dimme de Groot[1]

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Bendik Christensen
Final project course: CSE3000 Research Project
Thesis committee: Jorge Martinez, Dimme de Groot, Sole Pera

Abstract—**The accurate simulation of Room Impulse Responses (RIRs) is important in a variety of applications in acoustics such as automatic speech recognition, speech enhancement, and architectural acoustic design. While objective metrics for evaluating RIRs have been researched extensively, the subjective perceptual accuracy of the simulations is largely overlooked. This paper seeks to address this gap, designing a subjective testing methodology for evaluating the perceptual accuracy of simulated RIRs. A framework is proposed that combines the ABX testing methodology with a modified Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) approach, measuring attributes such as clarity, warmth, environment, and reverberation. The study involved 50 participants evaluating audio samples convolved with both real and simulated RIRs. Results seem to indicate that participants could reliably distinguish between real and simulated RIRs, with perceptual differences observed in the "clarity" and "reverberation" attributes. The findings suggest that current simulation methods for RIRs do not fully capture the perceptual aspects of acoustic environments.**

Index Terms—**Room Impulse Response (RIR), ABX testing, MUSHRA, Perceptual Accuracy, Subjective Evaluation**

I. Introduction

Research on how humans perceive sound across various environments has been ongoing for decades [1], and a thorough understanding of this field has a range of applications. Improving the robustness of our Automatic Speech Recognition (ASR) systems [2], implementing Speech Enhancement techniques [3] and integrating an acoustic understanding into the architectural design process [4] are but a few examples of use-cases for this knowledge. Many more can be found within the entertainment industry, specifically for cinema [5], and for music; for both live performances and recordings [6]. One way of modelling these environments is through Room Impulse Responses (RIRs), which is a "transfer function that aims to characterize the acoustic environment of a room" [7, p. 436], given specified source and listener positions. As such, convolving an anechoic signal with the measured RIR for a given room would model the room's effect on the signal [8].

RIRs themselves have been extensively researched, with different methods of construction and objective evaluation of the simulation performance being the main focus [9] [10] [11]. What seems to be missing from the literature however, is an understanding of the subjective component of the simulated RIRs. Both gpuRIR presented in [9] and Fast-RIR in [10] focus on purely on objective metrics such as the performance speedup and the simulation accuracy. Objective methods such as these tend to overlook the individuality of the auditory experience, facilitating a need for subjective testing. Additionally, for many applications, a physically accurate RIR isn't even required. Methods that attempt to achieve an accurate and realistic simulation such as wave-based methods often result in too high complexities, resulting in a trade-off between accuracy and complexity [12]. This leads to further investigation into perceptual accuracy being prompted, as the focus of many applications is to tailor to the auditory experience of humans. As a result, this research paper aims to fill this gap by exploring what constitutes a perceptually accurate Room Impulse Response simulation and how its efficacy can be tested. The main research question is: "How can subjective testing methodologies be designed to evaluate the perceptual accuracy of simulated room impulse responses?" The aim is to have a repeatable and valid experimental procedure that could be employed to ascertain a simulated RIR's intersubjective perceptual accuracy.

To address this, the paper is structured as follows. Firstly, various subjective methodologies currently employed in acoustics are reviewed to identify common practices and essential considerations. Subsequently, existing objective metrics are examined to understand their development and application. Thirdly, certain aspects of sound that a simulated RIR should accurately preserve are detailed. Subsequently, a novel subjective methodology designed to assess the perceptual accuracy of simulated RIRs is proposed, providing a framework for evaluating their effectiveness in reproducing real-world acoustic experiences. Finally, the experimental methodology is employed and the results are analysed and interpreted.

II. Background

This section discusses the background information based on a literature review, decomposing the theoretical foundation into three main components.

1) Which subjective methodologies are currently employed in acoustics.
2) Existing objective/subjective metrics used for the evaluation of audio.
3) Aspects of sound that a simulated RIR should accurately preserve.

A. Current Subjective Methodologies

A book on Perceptual Audio Evaluation [13] was looked at, in combination with a review of studies and standards applied to the field [14] [15] [16], to elicit a testing methodology workflow. This workflow considers the common practices and considerations in perceptual evaluation of audio, building up the methodology step by step. Figure 1 shows the elicited workflow.

A more thorough explanation of each component is provided in the Methods section, coupled with a description on how they were applied in this paper. Additionally, a key insight made in this theoretical component was that including a screening section allows for the possibility of conducting listening tests online, significantly speeding up data collection [17] and improving the overall results [18].

B. Existing Metrics

This subsection is split into two parts: the purely objective metrics that currently exist to evaluate RIRs, and the metrics that incorporate subjectivity in some way, applied to other domains.

1) Objective Metrics for RIRs: Existing objective metrics for RIRs primarily focus on physical accuracy or RIR applicability, without considering human perception. One approach is applying the Mean Squared Error (MSE) between RIRs, presenting the evaluation as a numerical physical difference [19]. Another commonly used approach is evaluating an RIR

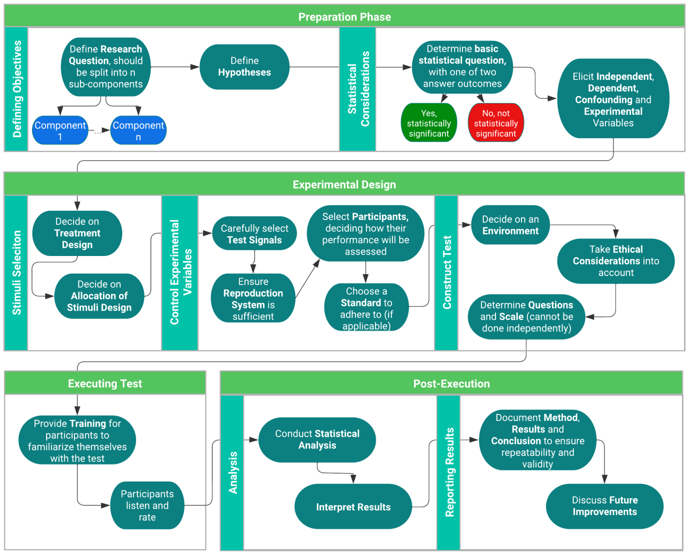

Preparation Phase
Defining Objectives
Define Research Question, should be split into n sub-components
Define Hypotheses
Component 1
...
Component n
Statistical Considerations
Determine basic statistical question, with one of two answer outcomes
Yes, statistically significant
No, not statistically significant
Elicit Independent, Dependent, Confounding and Experimental Variables
Experimental Design
Stimuli Seleciton
Decide on Treatment Design
Decide on Allocation of Stimuli Design
Control Experimental Variables
Carefully select Test Signals
Ensure Reproduction System is sufficient
Select Participants, deciding how their performance will be assessed
Choose a Standard to adhere to (if applicable)
Construct Test
Decide on an Environment
Take Ethical Considerations into account
Determine Questions and Scale (cannot be done independently)
Executing Test
Provide Training for participants to familiarize themselves with the test
Participants listen and rate
Post-Execution
Analysis
Conduct Statistical Analysis
Interpret Results
Reporting Results
Document Method, Results and Conclusion to ensure repeatability and validity
Discuss Future Improvements

Fig. 1: Perceptual Audio Evaluation workflow. Contains a visual representation of the considerations to be made when conducting an experiment on how people perceive various aspects of audio. Section III concretely describes how the paper implements these points.

simulation's effectiveness for an ASR use-case, for example by considering the Word Error Rate (WER) of ASR systems using those simulation methods to account for the environment [20] [21]. Additionally, a comparison of T60 times was employed in the Fast-RIR paper [10], comparing the T60 acquired from the paper's simulation method with the method proposed in [21]. The Reverberation Time (T60) is the "time taken for the sound to decay to 60 dB below its value at cessation" [22, p. 2877].

Finally, it is worth noting that a paper conducting a perceptual evaluation of simulated RIRs was found [23]. Expert listeners as defined by the International Telecommunication Union (ITU) standard ITU-R BS.1543-3 [24] participated in an online listening test, with the results being analyzed with an Analysis of Variance (ANOVA) test. The paper's focus was however on "perceptual thresholds of BRIR parameters", using purely simulated data in an attempt to distinguish simulated RIRs from each other. The perceptual accuracy of the RIRs themselves wasn't properly considered however, since all the RIRs used were simulated. Despite the paper's approach not being fully applicable to this particular use-case, key components such as the staircase and AB testing methods were considered for integration into the final methodology. The AB method consists of giving participants two audios, having them determine whether or not they are the same [25], whereas the staircase method starts from a detectable difference between the audios, reducing it in steps until the participant can no longer perceive it [26].

2) Metrics that take Human Perception into account for other Domains: In other domains such as evaluation of Basic Audio Quality (BAQ) and Overall Listening Experience (OLE) [27], metrics that incorporate human perception are more comprehensive and interpretable. PEAQ [28], an ITU standard [29], models human auditory perception and validates its results against subjective listening tests using the Subjective Difference Grade (SDG) metric. PEMO-Q [30] enhances PEAQ with auditory and cognitive models to produce a Perceptual Similarity Measure (PSM). PEASS [31] (Perceptual Evaluation methods for Audio Source Separation) takes it a step further, using PEMO-Q to determine the subjective significance of specific error estimation components, then refining them with subjective opinion scores collected through listening tests.

Although these metrics exemplify the combination of subjective and objective aspects of evaluation, the complexity of the psychoacoustic models and the metrics' corresponding techniques led to them being deemed excessive to fully research in the given time frame.

C. Attributes for Evaluating Perceptual Accuracy in Sound

In determining the final attributes for evaluating perceptual accuracy in sound, several sources were considered. A variety of works indicated that the main categories of attributes in the perceptive domain that could be considered subjective were timbral qualities and encoding of spatial information in the sound [16] [14] [13]. A journal article by Łetowski [32] formalizes these findings as a partition of sound into 5 components: loudness, pitch, duration, spaciousness and timbre. Pitch and duration can be objectively measured, significantly reducing their relevance for this paper. Loudness, although slightly more subjective [33], has methods such as the ITU-R BS.1770 algorithm [34], a widely used objective metric for loudness measurement [35]. Spaciousness and timbre however, are considered complex n-dimensional concepts, and given their subjective nature, further research into what they consist of is prompted.

In order to determine the timbral attributes that were to be considered in the methodology, the mural found in [36] was looked into. The mural presents a set of concepts relating to sound quality, and was used to elicit attributes such as: clarity, brightness, coloration and richness. Although all of the concepts could be worth looking into and could each provide valuable insights into how sounds differ, due to time and resource constraints only a select few are tested for. Brightness is disregarded as it was found to be too correlated with pitch [37]. Coloration seemed interesting but is slightly too similar to the spatial attributes as it is caused by the interference of the reflected sound with the original [38]. Richness and Texture have the problem of being too complex, with Richness being understood quite differently between different people [39] and texture being too encompassing of a term [40]. The final chosen timbral attributes are Clarity and Warmth. Clarity is the "perceived resolution of the auditory image" [41, p. 41], and relates to how distinct and understandable sounds are in a given acoustic environment [42]. It is highly correlated with intelligibility [43], without being constrained to communication. Warmth is related to pleasantness of the auditory experience, and was presented as one of the main subcomponents of timbre in [44], alongside Roundness, Brightness and Roughness. Since Roughness relates to Clarity and

Brightness to pitch, Roundness and Warmth remain. Finally, Warmth was selected as it is a simpler concept to explain and has a significant overlap with Roundness as well, as they both relate to the pleasantness aspect.

The spatial attributes were simpler to determine, as they were more uniformly defined among sources. A paper on "Spatial Quality Evaluation" [45] elicited twenty different attributes grouped into three categories: dimensional, immersion and miscellaneous spatial attributes. Due to the complexity of considering all the above, the attributes in this paper pertain to the categories themselves rather than some of the twenty specific attributes. Additionally, the ITU-R BS.1116 standard [14] confirms these categories, classifying the main important spatial attributes to be "Localization quality" and "Environment quality" respectively. Including a localization attribute in the experiment was deemed too complex due to the intricate nature of Head-Related Transfer Functions (HRTFs), which describe the acoustic transfer function between a point sound source in an open environment and a specific location within the listener's ear canal [46]. HRTFs contain the acoustic cues that are needed to localize sound sources and are highly individual [47]. Thus, this attribute is marked as potential future work and will be discussed further in the corresponding Future Work section .

One final attribute that is added to the evaluation is Reverberation. Reverberation refers to the lengthening of the sound duration due to the environment [48]. This attribute is interesting as it relates to the the other projects within the project group of this thesis, as they aim to explore different aspects of the T60. Given that the T60 is an "essential factor that reflects how reverberation affects a signal" [49, p. 1013], reverberation is considered in this paper as a means to obtain results that relate to the T60.

Through the decomposition of sound into the 4 subjective components: Clarity, Warmth, Environment and Reverberation, a variety of important subjective aspects of sound are covered. Clarity for the overall auditory resolution, Warmth for the pleasantness of experience, Environment for the immersion and Reverberation for the T60. Although these by no means cover the full range of elicitable attributes, they comprise of the most relevant and interesting ones that were feasible to accomplish.

III. Methods

This section will discuss the methodology used for the experimentation. The aforementioned workflow will be iterated through, explaining the different considerations made for this particular study. To understand the structure of this section better, refer to the Perceptual Audio Evaluation Workflow (Figure 1).

The Perceptual Evaluation Workflow 1 is split into four main phases: the preparation phase, the experimental design phase, the test execution and finally the post-execution phase.

Preparation Phase

In the preparation phase, the objectives are initially defined, starting with the research question. As aforementioned, the research question for the overall paper is "How can subjective testing methodologies be designed to evaluate the perceptual accuracy of simulated room impulse responses?" Having looked at the theoretical sub-questions, the overall research question can be de-composed into sub-components for the experimentation. Firstly, one component is whether or not the participants can perceptually distinguish between the simulated and the real RIRs. If so, the question of what attributes in which they differ is asked. Thus, the two main things the experiment will attempt to answer is: whether or not a simulated RIR is distinguishable from a real one, and on what attributes (elicited in the previous section II-C). The hypotheses of the experiment will thus reflect on these components, with the null hypotheses being that participants cannot reliably distinguish between real and simulated RIRs and that the ratings for the subjective attributes on the simulated ones are not significantly different enough from the real ones.

For the statistical considerations, the basic statistical question applied to this paper is: "Is the observed variability in the subjective impression a result of a perceptual distinguishability between the simulated RIRs and the actual RIRs, or just random fluctuations?" This basic statistical question summarizes the efforts of the experiment; answering it is the main goal. As a result, the Independent Variable is the type of RIR; whether or not it is simulated, and what simulation method is used if it is. The Dependent Variable is the detection accuracy, i.e the participants' ability to detect the real RIR, as well as the ratings given by participants on the different subjective attributes. There are many Confounding Variables, including the listening environment, the audio reproduction system and participants' hearing ability.

Experimental Design

The experimental design section starts with selecting the treatment design, which determines which stimuli will be shown and in what configurations. A properly chosen treatment design should control for the RIRs and the audio samples, since these can impact the results in different ways.

Four real RIRs with varying speaker placements are selected. Two of these are from the MeshRIR dataset [50], with a 0.38 s T60, a fixed room dimension of 7 x 6.4 x 2.7 meters, a fixed source position of [3.8 m, 2.9 m, 1.15 m] and the receiver positions [3.8 m, 2.9 m, 1.15 m] and [2.7 m, 2.7 m, 1.15 m] respectively. The other two RIRs are made in an audio lab at the Delft University of Technology, using time-stretched pulses to measure the impulse response of the room. The recording equipment used is:

- Microphones: AKG C417 PP
- Loudspeakers: Auratone 5C Super Sound Cube
- Audiocard: RME Fireface UFX+
- Microphone Preamp: RME OctaMic II
- Loudspeaker Power Amplifier: Auratone A2-30

The room dimensions for the audio lab are 8.1 m x 6.8 m x 3.07 m, with a fixed [6.01 m, 2.019 m, 1.175 m] source position and a fixed [2.11 m, 3.33 m, 0.99 m] receiver position. Two different configurations of the room are used, one with curtains drawn and one with curtains open, to make one RIR each. The T60s of the RIRs are 0.2 s and 0.42 s respectively. The aforementioned parameters are then used to simulate RIRs with two different methods, Pyroomacoustics [11] and the Mirror Image Source Method [51] as implemented by Emanuël Habets [52]. When simulating the RIRs, the room parameters are kept the same, and the absorption coefficients are selected such that the simulated RIR's T60 matches the real one. This calculation is done with the pyroomacoustics implementation of the Schroeder method [53]. The code for generating these simulations is made available in [54]. Four different anechoic speech excerpts are taken from the Pyramic Dataset [55], with two different male and female voices voicing phonetically rich sounding sentences. Four anechoic classical music segments are used as well, as classical music contains rich auditory information that could aid participants in identifying the real RIR. These classical music excerpts are found here [56]. Note that the excerpts themselves contain individually recorded instruments, so these are combined using "GarageBand for Mac". Finally, a sine sweep generated in python is the final audio sample used. The code and the excerpts may be found in [54].

Since the amount of participants is not enough, the treatment design is fractional and random; each excerpt is convolved with one of the real RIRs at random. The corresponding simulated RIRs are then convolved with the same audio. The sine sweep audio is convolved with two different RIRs, to ensure that there was some variation within that audio sample as well. The randomization combined with the choice of multiple stimuli and multiple RIRs should to some extent control the influence of the confounding variables. The allocation of stimuli design chosen is between-subjects design. This is purely due to the infeasibility of having all participants go through all nine samples. Instead, each participant is shown three, one of the classical music audios, one of the speech samples and the sine sweep. Each of these samples are shown to the participants three times, as they are convolved with three different impulse responses.

The experimental variables that are controlled are the test signals, the reproduction system, the participants and the environment. The controlling of the test signals is mentioned in the previous paragraphs, as a variety of samples were selected. The reproduction system used is the "Sennheiser HD-200 Pro". Participants are recruited through university channels, so neither the expertise level or the diversity is controlled. The mean age of participants is 28.5, the standard deviation is 6.5 with an age range of 19-51, with 35 male participants and 15 female participants. The testing standards used are the ABX testing methodology [57] in the first test, where participants are given a reference audio, as well as two audios A and B and asked to determine which one is an exact match with the reference. The second standard used is the ITU-R BS.1543-3 [24], otherwise known as Multiple Stimuli with Hidden Reference and Anchor (MUSHRA). A modified version of this standard is applied, where the participants rate the audio out of one-hundred on the subjective attributes rather than on basic audio quality, and no anchor is included, due to the complexity of defining one.

The listening test itself is designed by modifying the existing webMUSHRA software framework [58]. The modified code used for the test can be found here [54]. Ethical considerations are elaborated on in the Responsible Research section. In order to keep the test within a reasonable length, the ABX testing and the MUSHRA test were split into separate tests. The test flow of the software for the ABX test is as follows:

- Participants are prompted an explanation page that describes the test.
- Participants are then shown a training page to get them accustomed to the User Interface.
- ABX testing is executed, such that the participant must attempt to distinguish between an anechoic signal convolved with a real RIR and a simulated one for a classical music excerpt, a speech excerpt and the sine sweep excerpt. This will be repeated two times for each excerpt, as each simulation method is tested.
- Participants' age and gender are submitted and the test is completed.

The test flow for the MUSHRA test is very similar, with an additional explanation of the subjective attributes that were being rated being included in the beginning of the test. The orders of the appearances of the stimuli is randomized, to control for learning effects. The tests themselves are accessible by running the software in [54]. A ReadME.md file is provided which explains how to reproduce and use both tests.

A. Test Execution

The test is conducted in a small room (3m x 3m x 2.5m) in person, with the researcher present in case any questions about the software arise. Initially, as mentioned in section II-A, a screening section was considered such that the test could be conducted online, but due to GDPR and server constraints the tests are conducted in person instead. This is however possible in theory and would allow for faster and more efficient recruiting, allowing for more data to be gathered.

B. Post Execution

This component of the workflow will be discussed in detail in the Results and Conclusion sections V VI, consisting of a statistical analysis, interpretation of results as well as a discussion of potential future improvements.

IV. Responsible Research

Since this research involves human participants, additional considerations are needed to ensure that the research conducted is responsible. This section will detail these considerations, describing what steps are taken to mitigate the different kinds of risk.

Firstly, during the test execution itself, a variety of ethical issues were addressed. Initially, the participants were handed "Participant Information" and "Informed Consent Checklist" forms, that served to ensure that participants were informed of the nature of the study, as well as to explicitly obtain their written consent for participation and usage of the elicited data in the study. These forms may be found in [54]. In both forms, it is explicitly stated that the data will be anonymized, ensuring confidentiality for the participant. The "Participant Information" form also specifies that participation is entirely voluntary, and that participants can withdraw from the study at any time. To avoid potential bias in the results, participants were asked if they had any medically diagnosed hearing impairments prior to participating, and were omitted from the study in the event that they did. This was also important as the test favors "typical" hearing abilities over people with hearing impairments, which would make the test unfair for this group of participants had they been able to join. Throughout the test and after test completion, the corresponding researcher was available for questions and explanations on anything unclear. This also opened up for the possibility of debriefing participants after test completion, further describing the purpose and details of the study, including what will be done with the data. The debriefing was non-mandatory however, so mainly related to participants who showed an interest and had further questions out of curiosity.

Secondly, after the test execution, additional concerns such as adhering to the Findable, Accessible, Interoperable and Reusable (FAIR) principles [59] were addressed. In order to do so, significant changes were made to the code used in the study to ensure its readability and understandability. One such example is the addition of "ReadME" files detailing the structure of the published data, with clear comments in the code explaining how future researchers could re-use it to either achieve the same results, or to build on them. To allow for this, a detailed description of the methodology is also presented in the Methods section, promoting the replicability of the attained results. An explanation of bias mitigation can also be found in that section, as bias mitigation is a useful method to increase the validity of the results [60]. The results and the code themselves are findable in [54]. A "Data Management Plan" (DMP) form was also made prior to test execution, describing what would be done with the data. The DMP was reviewed by a Faculty Data Steward on the 28/05/2024 and was adhered to to the best of the corresponding researcher's ability. The DMP can be found in [54] in the corresponding "Docs" folder. The documents in this folder were submitted for approval to the Human Research Ethics Committee of the Delft University of Technology.

Finally, one major consideration, given the amounts of software and data taken from external sources, is licensing. A list of external software/data in this study and their corresponding licensing information is:

- WebMUSHRA software [58]. Modified and used for the data collection, this software has its own license titled "Software License for the webMUSHRA.js Software". In accordance with the license, it is made clear that the software is a third-party modification, with an additional file "modifications.md" being added specifying the modifications made to the code. Additionally, the software has three inherent dependencies, two of them carrying the MIT license and one of them with the Apache license. These are also considered, but as the licenses are quite permissive they didn't contribute to additional constraints for the purposes of this paper.
- MeshRIR Room Impulse Response Dataset [50]. This dataset contains two of the four real room impulse responses used, and goes under the CC BY 4.0 license.
- Pyramic dataset [55]. This dataset was used for the anechoic speech segments listened to in the experiment, and the data is under the CC BY 4.0 license.
- Pyroomacoustics library [11] and Habets' implementation of the Mirror Image Source Method [52], the libraries used to simulate the simulated RIR conditions, both going under the MIT license.
- "Anechoic recording system for symphony orchestra" dataset [56], was used for the four anechoic classical music recordings that were used in the experiment. The audios are saved separately for each instrument, but were combined in a digital audio workstation for the experiment. No license is specified, but it is explained that a citation is sufficient for academic research purposes.

Although this isn't a comprehensive list of all the external dependencies, it provides the ones that could potentially provide licensing restrictions. The license of the data published in this paper will thus have to be of a similar level of restriction to the aforementioned ones, so a CC BY 4.0 is deemed sufficient.

V. Results

This section details the elicited results from the conducted experiment. Firstly, the results of the paired comparison are brought forth. Subsequently, the ratings of the subjective attributes are analysed. Note that this section will not interpret the results, as the interpretation is done in VI.

A. *Paired Comparison (ABX) Results*

27 participants for the paired comparison were shown 6 audio samples each (3 per simulation method), producing 81 samples per simulation method. The overall combined results are given in the following table:

	Pyroomacoustics	**Habets**
Correct	56	61
Incorrect	6	6
Undecided	19	14

TABLE I: Overall results of paired comparison tests. The "Pyroomacoustics" and "Habets" refer to the corresponding simulation methods used to simulate the RIRs that were convolved with the anechoic sound. The rows describe whether the participant was able to correctly distinguish between the simulated RIR and the real one. "Correct" implies that they selected the correct audio as the reference, "Incorrect" means the wrong audio was selected and "Undecided" means that the participant selected the "I don't know" option.

A binomial test [61] is conducted on the overall results for the paired comparison tests to determine the extent to which the results could be attributed to random error. The p-value attained for the pyroomacoustics results for a binomial test, where 56 out of 81 times the correct reference was identified, is $3.76 * 10^{-4}$. Similarly, the binomial test results for the "habets" simulation method gives a p-value of $3 * 10^{-6}$. Additionally, a chi-squared test [62] is conducted between the two simulation method results, giving a chi-squared statistic of 0.971 and a p-value of 0.615.

To provide a more in detail analysis of the RIR differentiation based on the audio samples, an additional figure is provided:

	Classical		**Sine Sweep**		**Speech**	
	Pyroomacoustics	**Habets**	**Pyroomacoustics**	**Habets**	**Pyroomacoustics**	**Habets**
Correct	20	17	12	18	24	26
Incorrect	2	2	4	3	0	1
Undecided	5	8	11	6	3	0

TABLE II: Results of paired comparison tests per audio sample category. The columns "Pyroomacoustics" and "Habets" refer to the corresponding simulation methods used to simulate the RIRs that were convolved with the anechoic "Classical", "Sine Sweep" and "Speech" correspond to the audio categories the used samples belong to. The rows describe whether the participant was able to correctly distinguish between the simulated RIR and the real one. "Correct" implies that they selected the correct audio as the reference, "Incorrect" means the wrong audio was selected and "Undecided" means that the participant selected the "I don't know" option.

A binomial test and chi-squared between simulation methods, conducted per attribute, elicits the following values:

- Classical:
  - Pyroomacoustics binomial p-value: $9.6 * 10^{-3}$
  - Habets binomial p-value: 0.124
  - Chi-squared statistic: 0.934
  - Chi-squared p-value: 0.626
- Sine Sweep:
  - Pyroomacoustics binomial p-value: 0.779
  - Habets binomial p-value: 0.061
  - Chi-squared statistic: 2.813
  - Chi-squared p-value: 0.245
- Speech:
  - Pyroomacoustics binomial p-value: $3 * 10^{-5}$
  - Habets binomial p-value: $2 * 10^{-7}$
  - Chi-squared statistic: 4.08
  - Chi-squared p-value: 0.13

B. Subjective Attribute Results

For the subjective attributes, 23 participants were given 9 audio samples (3 per RIR comparison), leading to 69 samples per RIR comparison. The table of results themselves can be found in [54], but a summary of the data per attribute is seen in the plots in Figure 2.

Upon visual inspection of the plots' interquartile ranges, medians and whiskers, reverberation appears to be the main attribute that differs significantly between the conditions. As such, it's considered first in the further analysis. A "one-factor ANOVA" (Analysis of Variance) [63] is conducted per attribute to determine if post-hoc tests are relevant [64]:

1) Attribute Specific Results: Reverberation: The p-value attained from a single factor ANOVA applied to the combined reverberation data, which can be found in [54], is $1.2 * 10^{-3}$. The critical f-value is 3.04 and the measured f-value is 9.45. Due to the potential implied significance of these results, post-hoc analysis is conducted. A Bonferroni correction is used as it is a simple, conservative adjustment that minimizes the risk of attaining a type I error [65]. Applying pairwise t-tests to the RIR conditions results in the following table:

	Real	Pyroomacoustics	Habets
Real	x	0.789	$1.9 * 10^{-3}$
Pyroomacoustics	0.789	x	$4.5 * 10^{-3}$
Habets	$1.9 * 10^{-3}$	$4.5 * 10^{-3}$	x

TABLE III: Results of pairwise t-tests between RIR conditions for the "Reverberation" attribute. The "Pyroomacoustics" and "Habets" rows and columns refer to the corresponding simulation methods used to simulate the RIRs that were convolved with the anechoic sound, with "Real" representing the real measured RIR.

Applying the Bonferroni correction with 3 pair possibilities leads to a corrected α value of $0.05/3 = 0.017$, where α is the highest p-value that leads to a null hypothesis rejection.

2) Attribute Specific Results: Clarity: The p-value attained from a single factor ANOVA applied to the combined clarity data, which can be found in [54], is 0.028. The critical f-value is 3.04 and the measured f-value is 3.64. Similarly to the reverberation attribute, the Bonferroni Correction post-hoc test with α = 0.017 is conducted to verify the pairs in which the potential statistical difference manifests, leading to the following table:

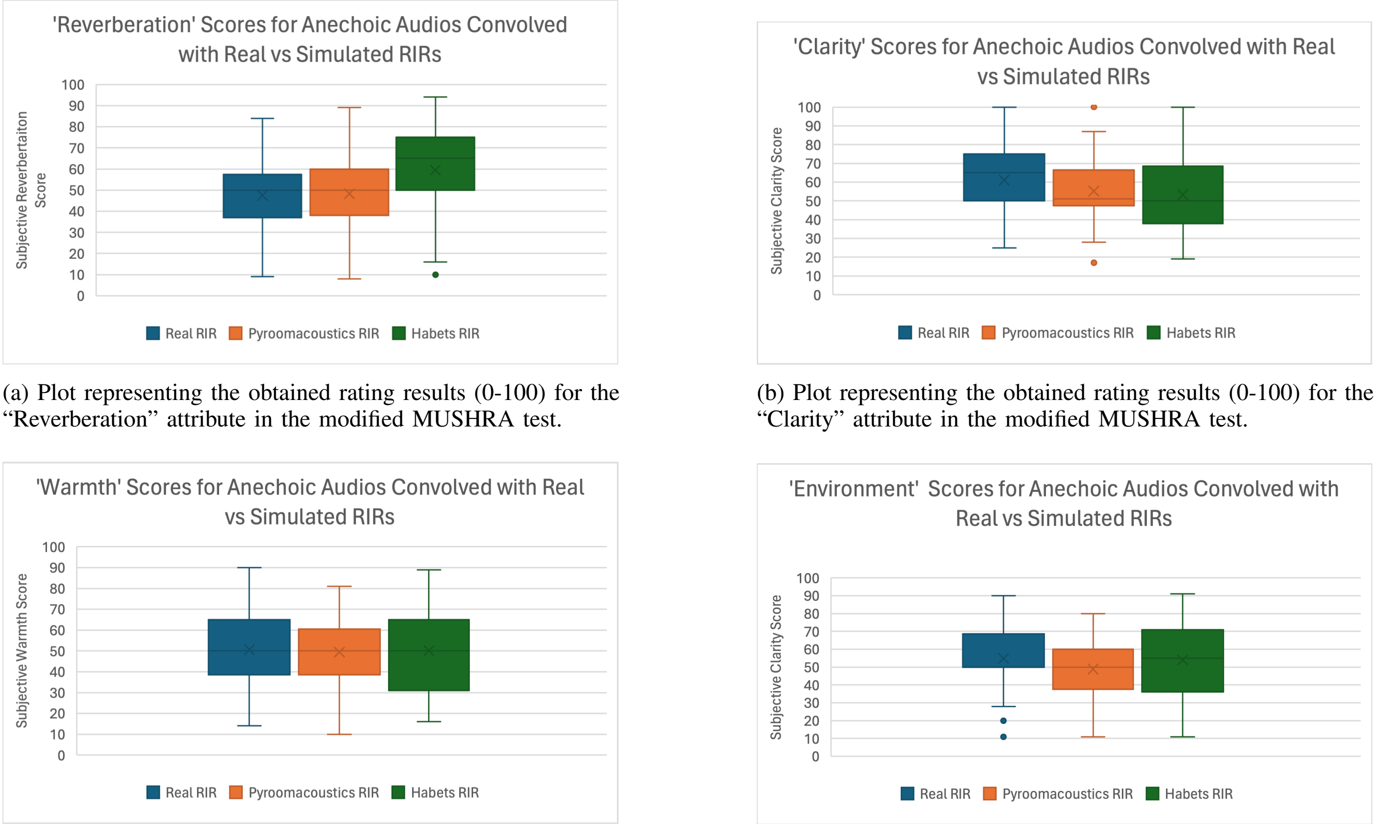
	Real	Pyroomacoustics	Habets
Real	x	0.0441	0.0124
Pyroomacoustics	0.0441	x	0.525
Habets	0.0124	0.525	x

TABLE IV: Results of pairwise t-tests between RIR conditions for the "Clarity" attribute. The "Pyroomacoustics" and "Habets" rows and columns refer to the corresponding simulation methods used to simulate the RIRs that were convolved with the anechoic sound, with "Real" representing the real measured RIR.

Additionally, due to clarity being correlated to intelligibility, as seen in II-C, the speech data could be of specific interest for this attribute. As such, the same post-hoc test is conducted on only the speech data as well, eliciting the following table:

	Real	Pyroomacoustics	Habets
Real	x	0.0347	$2.76 * 10^{-6}$
Pyroomacoustics	0.0347	x	$3.5 * 10^{-3}$
Habets	$2.76 * 10^{-6}$	$3.5 * 10^{-3}$	x

TABLE V: Results of pairwise t-tests between RIR conditions for the "Clarity" attribute considering only the speech data. The "Pyroomacoustics" and "Habets" rows and columns refer to the corresponding simulation methods used to simulate the RIRs that were convolved with the anechoic sound, with "Real" representing the real measured RIR.



(a) Plot representing the obtained rating results (0-100) for the "Reverberation" attribute in the modified MUSHRA test.

(b) Plot representing the obtained rating results (0-100) for the "Clarity" attribute in the modified MUSHRA test.

(c) Plot representing the obtained rating results (0-100) for the "Warmth" attribute in the modified MUSHRA test.

(d) Plot representing the obtained rating results (0-100) for the "Environment" attribute in the modified MUSHRA test.

Fig. 2: 2x2 Grid of "Box and Whisker" plots, representing the obtained rating results (0-100) per attribute in the modified MUSHRA test. The boxes represent the interquartile range, showing the middle 50% of the data, with the line in the middle representing the median. The "whiskers" demonstrate the range of the data excluding outliers, whilst the points that lie outside of this range represent the outliers.

3) Attribute Specific Results: Warmth: For the warmth attribute, the ANOVA test results in a 0.926 p-value and a 0.0772 f-value for a 3.04 critical f-value, implying that no further post-hoc tests should be conducted. However, since warmth is related to pleasantness of auditory experience, as seen in II-C, an additional ANOVA was conducted with only the classical music samples, as this concept is more applicable to music than it is to speech or sine sweeps. Despite this change, the p-value remains above the $\alpha = 0.05$ threshold, with a value of 0.105 and thus prompts no further exploration.

4) Attribute Specific Results: Environment: Similarly to the warmth attribute, the environment attribute seems to indicate a non-significant result with a 0.09 p-value and 2.42 f-value based on a 3.04 critical f-value, when conducting a one-factor ANOVA analysis. However, seeing as immersion as a criteria could also be easier to distinguish for an orchestral classical performance, another ANOVA is conducted purely with those samples. As with the warmth attribute, a 0.066 p-value and 2.82 f-value for a 3.14 critical f-value prompts no further testing.

VI. Conclusions and Future Work

This section interprets the results elicited in V and discusses them in relation to the paper's context. Consequently, limitations and potential future improvements are outlined. Finally, the paper concludes, summarizing the work done.

A. Discussion of Results

Firstly, the paired comparison results are discussed. The low p-values for the overall results, as seen in Table I, indicate that the participants were able to tell the difference between the reference audio (the anechoic sample convolved with the real RIR) and the simulated versions, with a low probability that the elicited results are achieved randomly. One can notice that participants overall were better at determining the simulated version when it was simulated using the "Habets" method, which could suggest that Pyroomacoustics was harder to distinguish in the experiment. However, the chi-squared test values between the simulation method results implies that there is no statistically significant difference between the simulation methods themselves, which could mean that perceptually one isn't favored over another.

Considering the paired comparison results per audio sample category, as seen in table II, potentially interesting new interpretations arise. For the classical samples, it is clear that the p-value is lower for "Pyroomacoustics" than for "Habets", indicating that for the classical samples the participants found it easier to identify the simulated RIR reliably when it was simulated with Pyroomacoustics. This difference between them is however not deemed statistically significant, as the chi-squared statistic is low and the p-value is high. For the sine sweep samples, interestingly participants struggled to differentiate the pyroomacoustics RIR more than the habets one. It is also clear however that this difference isn't statistically significant either, due to the chi-squared values. Finally, the speech samples had the best overall results for identification of simulated RIRs, with the lowest reported p-values by a large margin. This could be because participants are most accustomed to hearing speech in their day-to-day, suggesting that it is easier to hear subtle differences in the audio samples when they're more integrated into real life scenarios.

For the subjective attributes, the main ones that are found to be interesting within the results are "Clarity' and "Reverberation", as seen in Tables III, IV and V. "Warmth" and "Environment" do not seem to yield statistically significant results. This could be due to the complexity and subjectivity of their experience, as both pleasantness of auditory experience and environmental immersion are abstract concepts. Since the participants weren't required to have prior experience with audio, they could have had a difficult time understanding what was meant, as well as relying on fundamentally different preconceived notions of what is "pleasant" or "immersive". This is further enhanced by the imprecise definitions provided in the user interface, using subjective terms such as "full", "rich", "envelopment" and "ambience", highlighting the importance of clear communication and training in studies involving subjective attributes. On the other hand however, one must be wary of doing so as specific definitions could bias participant interpretations, not allowing them the freedom to fully grasp the attribute as is.

The "Reverberation" results, as seen in III, indicate that participants consistently rated the audio convolved with the Habets RIR differently from the real one, which is also visible in 2a. Given that the pairwise p-value of $1.9 * 10^{-3}$ is well below the 0.017 threshold, it's reasonable to conclude that participants rated the reverberation quite differently in the Habets condition from the real one. There doesn't seem to be a significant difference between pyroomacoustics and the real one however, suggesting that Pyroomacoustics could be better at modelling reverberation effects than the Habets method. The difference between the simulation methods themselves isn't too important of a result, but it does serve to back up the claim that they are producing different results.

For the "Clarity" results, only Habets falls under the 0.017 threshold for the elicited overall p-value as seen in Table IV. Participants rated the clarity of the audio convolved with the real RIR as higher than the simulated ones on average, see 2b. In the subsequent table Table V, where only the speech samples are considered, it is seen that the value for pyroomacoustics is reduced slightly, but not enough to fall beneath the threshold, whereas Habets is reduced even more to indicate a strong correlation well below the p-value. This could imply that the Habets method is not sufficiently able to maintain the clarity of the audio, especially when considering speech samples. Although the Pyroomacoustics method doesn't seem to be able to model the Clarity well either based on this low value, the results don't suggest a statistical significance as the p-value for both tables is over 0.017.

To summarize the results, this experiment suggests that both the Habets and the Pyroomacoustics methods of simulating room impulse responses can be perceptually distinguished from a real room impulse response for a space. Additionally, the main attributes that the simulated RIRs seem to fail to preserve are perceived "Reverberation" and "Clarity", due to the difference in subjective ratings between the audio samples. Despite this, it is not possible to definitively conclude that these findings are properly statistically significant. The next subsections will outline why this is the case, as well as provide potential improvements to improve the robustness and validity of the experiment.

B. Limitations of Findings

Although some of the results seen in VI indicate a statistically significant ability for participants to perceive the simulated RIRs differently from the real ones based on certain attributes, certain limitations hinder the paper from reaching a definite conclusion. This subsection will elaborate on these.

One main limitation of this experiment lies in the selection of participants. Since the experiment was conducted in person in a controlled setting, the recruitment was more difficult and thus the sample size isn't sufficient to generalize. Additionally, since the gender distribution is skewed, the data could have a bias towards males. Finally, an additional concern that was not addressed is the level of listener expertise, which wasn't asked for in the experiment. If the sample size was large and diverse enough, then it would be possible to dismiss this as the general level of listening expertise, but this is not possible with this experiment.

Some more limitations lie within the experimental design itself. One major consideration is the way in which the RIRs were simulated, as it's not guaranteed that the simulation methods were used optimally. Although the room parameters remained the same, the T60s were estimated and the individual absorption coefficients of the walls weren't considered. It is thus conceivable that a better use of the simulation methods could lead to better simulation results, improving the perceptual accuracy of the RIRs. Additionally, the headphones used were the same for all participants. All headphones have a unique frequency response, which is a function of the amplitude vs frequency for the output of the system [6]. Since different headphones have different frequency responses, it isn't necessarily guaranteed that the results with these headphones will generalize to other ones. A similar argument can be used for the listening environment, since the environment was mostly controlled, differing listening environments aren't accounted for. This limits the natural validity of the experiment. Finally, although multiple T60s were used, the low range of them from 0.18-0.42 s could also limit the extent to which the T60's effect on the results was controlled.

The final concerns in validity relate to the attributes elicited in the background section II-C. Considering differences in audio only on 4 attributes is a major oversimplification and was only done because it was infeasible and complex to do more. For reference, the sound quality mural in [36] contains eighteen different aspects, and it doesn't even have all the attributes that were elicited. It would thus be more valid to have more attributes and also make them more specific. The complexity of the attributes chosen is also an issue that was highlighted throughout the experiment, as participants struggled to fully understand the descriptions provided.

C. Future Work

As a consequence of the limitations described in VI-B, potential future improvements and additions are elaborated on in this subsection.

The first possibility for additional features for the experiment would be testing for listener expertise. Since this directly impacts the results, it would be helpful to control for this to further analyse the results per listener group. It is possible that an expert listener group would perform better overall, especially for the classical music and sine wave audios, as this target group would be exposed to more of these types of audios. Additionally, including a larger sample size would be a welcome addition, as ascertaining a proper statistical significance would be easier in that case. One method that was mentioned in section III is the possibility of deploying the test online, rendering participant recruitment easier. This idea was successfully implemented with the code as the webMUSHRA software allows for simple deployment with Docker, but was not possible to apply in practice due to server and GDPR constraints imposed by the Delft University of Technology. A larger sample size would also allow for the possibility of a full factorial treatment design, increasing the robustness of the experiment to the variance in audio samples and participants.

Some additional potential future work relates to including more subjective attributes. Since audio can be decomposed into many more subjective attributes than were tested for, it would be interesting to see if there are any other, perhaps more specific, attributes that describe the discrepancies between the RIRs better. An example of such a subjective attribute is "Localization" as mentioned in II-C.

Finally, as mentioned in II-A, the possibility of considering the perceptual components of existing objective measures is an interesting avenue for further research. Since these metrics are extensively researched and incorporate aspects of human perception, a thorough understanding of these could help guide the direction of subjective evaluation of any aspect of audio.

D. Conclusion

This study aimed to investigate subjective methodologies for the evaluation of perceptual accuracy of simulated room impulse responses. To this end, a two-fold subjective methodology was proposed for RIRs simulated with two different methods. Firstly, an ABX test to determine the extent to which participants could differentiate simulated and real room impulse responses was conducted. Subsequently, a modified MUSHRA test was conducted to gain a deeper understanding into what subjective attributes of the audio were perceived to be different. The findings indicate that participants were able to tell the simulated RIRs from the real one, with the main perceptual differences being in the ”reverberation” and ”clarity” attributes. This seems to suggest that current RIR simulation methods cannot accurately maintain the perceived reverberation and clarity in the audio. Despite this, the results must be considered with caution for a variety of reasons, and future work can be undertaken to tackle the limitations with the experimental methodology. In conclusion, this thesis has contributed to a better understanding of how perceptual evaluation of simulated room impulse responses can be conducted, applying a simple experimental design to lay the groundwork for future research. Thus, this work not only advances current methodologies but also paves the way for more refined and comprehensive studies in the perceptual evaluation of audio.

References

[1] W. T. Nelson, R. S. Bolia, M. A. Ericson, and R. L. McKinley, "Spatial audio displays for speech communications: A comparison of free field and virtual acoustic environments," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 43, no. 22. SAGE Publications Sage CA: Los Angeles, CA, 1999, pp. 1202–1205.

[2] C. Weng, D. Yu, S. Watanabe, and B.-H. F. Juang, "Recurrent deep neural networks for robust speech recognition," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 5532–5536.

[3] D. S. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 7, pp. 1492–1501, 2017.

[4] S. Pelzer, L. Aspöck, D. Schröder, and M. Vorländer, "Integrating real-time room acoustics simulation into a cad modeling software to enhance the architectural design process," *Buildings*, vol. 4, no. 2, pp. 113–138, 2014. [Online]. Available: https://www.mdpi.com/2075-5309/4/2/113

[5] M. C. Ward, "The soundscape of the cinema theatre," *Music, Sound, and the Moving Image*, vol. 10, no. 2, pp. 135–165, 2016. [Online]. Available: https://www.liverpooluniversitypress.co.uk/doi/abs/10.3828/msmi.2016.8

[6] F. Toole, *Sound reproduction: The acoustics and psychoacoustics of loudspeakers and rooms*. Routledge, 2017.

[7] W. Yu and W. B. Kleijn, "Room acoustical parameter estimation from room impulse responses using deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 436–447, 2021.

[8] J. Mourjopoulos, "On the variation and invertibility of room impulse response functions," *Journal of Sound and Vibration*, vol. 102, no. 2, pp. 217–228, 1985.

[9] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, “gpurir: A python library for room impulse response simulation with gpu acceleration,” *Multimedia Tools and Applications*, vol. 80, no. 4, pp. 5653–5671, 2021.

[10] A. Ratnarajah, S.-X. Zhang, M. Yu, Z. Tang, D. Manocha, and D. Yu, “Fast-rir: Fast neural diffuse room impulse response generator,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 571–575.

[11] R. Scheibler, E. Bezzam, and I. Dokmanić, “Pyroomacoustics: A python package for audio room simulation and array processing algorithms,” in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 351–355.

[12] J. Martinez, “Low-complexity computer simulation of multichannel room impules responses,” 2013.

[13] S. Bech and N. Zacharov, *Perceptual audio evaluation-Theory, method and application*. John Wiley & Sons, 2007.

[14] International Telecommunication Union, *Methods for the Subjective Assessment of Small Impairments in Audio Systems*, ITU-R Recommendation BS.1116-3, International Telecommunication Union Std., 2015, available: https://www.itu.int/rec/R-REC-BS.1116-3-201502-I/en.

[15] J. Herre and C. Faller, “Spatial sound stability enhancement by advanced user-tracked loudspeaker rendering,” in *Audio Engineering Society Convention 155*. Audio Engineering Society, 2023.

[16] T. Robotham, O. Rummukainen, J. Herre, and E. A. Habets, “Online vs. offline multiple stimulus audio quality evaluation for virtual reality,” in *Audio Engineering Society Convention 145*. Audio Engineering Society, 2018.

[17] M. Cartwright, B. Pardo, G. J. Mysore, and M. Hoffman, “Fast and easy crowdsourced perceptual audio evaluation,” in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 619–623.

[18] S. Moulin, G. Pallone, N. Faure, and S. Bech, “Perceptual evaluation of loudspeaker misplacement compensation in a multichannel setup using mpeg-h 3d audio renderer. application to channel-based, scene-based, and object-based audio materials,” in *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*. Audio Engineering Society, 2019.

[19] D. A. Sanaguano-Moreno, J. F. Lucio-Naranjo, R. A. Tenenbaum, L. Bravo-Moncayo, and G. B. Regattiere-Sampaio, “A deep learning approach for the generation of room impulse responses,” in *2022 Third International Conference on Information Systems and Software Technologies (ICI2ST)*. IEEE, 2022, pp. 64–71.

[20] A. Ratnarajah, Z. Tang, and D. Manocha, “Ts-rir: Translated synthetic room impulse responses for speech augmentation,” in *2021 IEEE automatic speech recognition and understanding workshop (ASRU)*. IEEE, 2021, pp. 259–266.

[21] Z. Tang, L. Chen, B. Wu, D. Yu, and D. Manocha, “Improving reverberant speech training using diffuse acoustic simulation,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6969–6973.

[22] R. Ratnam, D. L. Jones, B. C. Wheeler, W. D. O’Brien Jr, C. R. Lansing, and A. S. Feng, “Blind estimation of reverberation time,” *The Journal of the Acoustical Society of America*, vol. 114, no. 5, pp. 2877–2892, 2003.

[23] H. Mi, G. Kearney, and H. Daffern, “Impact thresholds of parameters of binaural room impulse responses (brirs) on perceptual reverberation,” *Applied Sciences*, vol. 12, no. 6, p. 2823, 2022.

[24] International Telecommunication Union, *Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems*, ITU-R Recommendation BS.1543-3, International Telecommunication Union Std., 2019, available: https://www.itu.int/rec/R-REC-BS.1534.

[25] J. Boley and M. Lester, “Statistical analysis of abx results using signal detection theory,” in *Audio Engineering Society Convention 127*. Audio Engineering Society, 2009.

[26] T. N. Cornsweet, “The staircase-method in psychophysics,” *The American journal of psychology*, vol. 75, no. 3, pp. 485–491, 1962.

[27] M. Schoeffler, A. Silzle, and J. Herre, “Evaluation of spatial/3d audio: Basic audio quality versus quality of experience,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 1, pp. 75–88, 2017.

[28] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, and C. Colomes, “Peaq-the itu standard for objective measurement of perceived audio quality,” *Journal of the Audio Engineering Society*, vol. 48, no. 1/2, pp. 3–29, 2000.

[29] International Telecommunication Union, *Method for Objective Measurements of Perceived Audio Quality*, ITU-R Recommendation BS.1387, International Telecommunication Union Std., 2001, available: https://www.itu.int/rec/R-REC-BS.1387.

[30] R. Huber and B. Kollmeier, “Pemo-q—a new method for objective audio quality assessment using a model of auditory perception,” *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 6, pp. 1902–1911, 2006.

[31] E. Vincent, “Improved perceptual metrics for the evaluation of audio source separation,” in *Latent Variable Analysis and Signal Separation: 10th International Conference, LVA/ICA 2012, Tel Aviv, Israel, March 12-15, 2012. Proceedings 10*. Springer, 2012, pp. 430–437.

[32] T. Łetowski, “Timbre, tone color, and sound quality: concepts and definitions,” *Archives of Acoustics*, vol. 17, no. 1, pp. 17–30, 2014.

[33] S. S. Stevens, “The measurement of loudness,” *The Journal of the Acoustical Society of America*, vol. 27, no. 5, pp. 815–829, 1955.

[34] International Telecommunication Union, *Algorithms to measure audio programme loudness and true-peak audio level*, International Telecommunication Union Std., 2015. [Online]. Available: https://www.itu.int/rec/R-REC-BS.

1770

[35] P. D. Pestana, J. D. Reiss, and A. Barbosa, "Loudness measurement of multitrack audio content using modifications of itu-r bs. 1770," in *Audio Engineering Society Convention 134*. Audio Engineering Society, 2013.

[36] T. Letowski, "Sound quality assessment: Concepts and criteria," in *Audio Engineering Society Convention 87*. Audio Engineering Society, 1989.

[37] E. G. Boring and S. Stevens, "The nature of tonal brightness," *Proceedings of the National Academy of Sciences*, vol. 22, no. 8, pp. 514–521, 1936.

[38] Y. Seki and K. Ito, "Coloration perception depending on sound direction," *IEEE transactions on speech and audio processing*, vol. 11, no. 6, pp. 817–825, 2003.

[39] C. Saitis, C. Fritz, and G. Scavone, "Sounds like melted chocolate: How musicians conceptualize violin sound richness," in *International Symposium on Musical Acoustics*, 2019.

[40] R. L. Klatzky and S. J. Lederman, "Multisensory texture perception," *Multisensory object perception in the primate brain*, pp. 211–230, 2010.

[41] T. Rościszewska, A. Miśkiewicz, T. Rogala, T. Rudzki, and T. Fidecki, "Concert hall sound clarity: A comparison of auditory judgments and objective measures," *Archives of Acoustics*, pp. 41–46, 2012.

[42] D. H. Griesinger, "What is clarity, and how it can be measured?" in *Proceedings of Meetings on Acoustics*, vol. 19, no. 1. AIP Publishing, 2013.

[43] L. S. Eisenberg, D. D. Dirks, S. Takayanagi, and A. S. Martinez, "Subjective judgments of clarity and intelligibility for filtered stimuli with equivalent speech intelligibility index predictions," *Journal of Speech, Language, and Hearing Research*, vol. 41, no. 2, pp. 327–339, 1998.

[44] V. Rosi, "The Metaphors of Sound : from Semantics to Acoustics. A Study of Brightness, Warmth, Roundness, and Roughness," Theses, Sorbonne Université, Jul. 2022. [Online]. Available: https://theses.hal.science/tel-03994903

[45] F. Rumsey, "Spatial quality evaluation for reproduced sound: Terminology, meaning, and a scene-based paradigm," *Journal of the Audio Engineering Society*, vol. 50, no. 9, pp. 651–666, 2002.

[46] H. Møller, "Fundamentals of binaural technology," *Applied acoustics*, vol. 36, no. 3-4, pp. 171–218, 1992.

[47] S. Li and J. Peissig, "Measurement of head-related transfer functions: A review," *Applied Sciences*, vol. 10, no. 14, p. 5014, 2020.

[48] V. Valimaki, J. D. Parker, L. Savioja, J. O. Smith, and J. S. Abel, "Fifty years of artificial reverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1421–1448, 2012.

[49] Y. Li, Y. Liu, and D. S. Williamson, "A composite t60 regression and classification approach for speech dereverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1013–1023, 2023.

[50] S. Koyama, T. Nishida, K. Kimura, T. Abe, N. Ueno, and J. Brunnström, "Meshrir: A dataset of room impulse responses on meshed grid points for evaluating sound field analysis and synthesis methods," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2021, pp. 1–5.

[51] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[52] E. Habets, "Rir-generator," https://github.com/ehabets/RIR-Generator, 2024, accessed: 2024-06-09.

[53] M. R. Schroeder, "New method of measuring reverberation time," *The Journal of the Acoustical Society of America*, vol. 37, no. 6_Supplement, pp. 1187–1188, 1965.

[54] B. Christensen, "Data and code underlying the research project: Evaluation of Perceptual Accuracy in Simulated Room Impulse Responses," 2024. [Online]. Available: https://doi.org/10.4121/9208260b-8625-4917-8ad4-f56190187070

[55] R. Scheibler, J. Azcarreta, R. Beuchat, and C. Ferry, "Pyramic: Full stack open microphone array architecture and dataset," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018, pp. 226–230.

[56] J. Pätynen, V. Pulkki, and T. Lokki, "Anechoic recording system for symphony orchestra," *Acta Acustica united with Acustica*, vol. 94, no. 6, pp. 856–865, 2008.

[57] W. Munson and M. B. Gardner, "Standardizing auditory tests," *The Journal of the Acoustical Society of America*, vol. 22, no. 5_Supplement, pp. 675–675, 1950.

[58] M. Schoeffler, S. Bartoschek, F.-R. Stöter, M. Roess, S. Westphal, B. Edler, and J. Herre, "webmushra—a comprehensive framework for web-based listening tests," *Journal of Open Research Software*, vol. 6, no. 1, p. 8, 2018.

[59] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[60] N. Norris, "Error, bias and validity in qualitative research," *Educational action research*, vol. 5, no. 1, pp. 172–176, 1997.

[61] H. Abdi, "Binomial distribution: Binomial and sign tests," *Encyclopedia of measurement and statistics*, vol. 1, 2007.

[62] R. L. Plackett, "Karl pearson and the chi-squared test," *International statistical review/revue internationale de statistique*, pp. 59–72, 1983.

[63] L. St, S. Wold *et al.*, "Analysis of variance (anova)," *Chemometrics and intelligent laboratory systems*, vol. 6, no. 4, pp. 259–272, 1989.

[64] M. L. McHugh, "Multiple comparison analysis testing in anova," *Biochemia medica*, vol. 21, no. 3, pp. 203–209, 2011.

[65] R. A. Armstrong, "When to use the b onferroni correction," *Ophthalmic and Physiological Optics*, vol. 34, no. 5, pp. 502–508, 2014.