Adjusting the delay and gain of late reverberation in the context of real-time auralization

March 3 - August 11, 2025
Internship report
August 19, 2025

Alice Pain

Master ATIAM
Sorbonne Université, Télécom Paris, IRCAM

Supervision

Brian F.G. Katz, *brian.katz@sorbonne-universite.fr*
David Poirier-Quinot, *david.poirier-quinot@sorbonne-universite.fr*
Institut Jean Le Rond ∂'Alembert, Sorbonne Université, CNRS
4 place Jussieu, Paris

Abstract

This work introduces and evaluates an algorithm for dynamically adjusting the delay and gain of artificial late reverberation in a real-time auralization system combining the image source method with a Feedback Delay Networks. We propose a mixing time estimator that operates solely on the early reflections of a Room Impulse Response (RIR), enabling its use in real-time dynamic auralization where full RIRs are unavailable. The estimator is based on both signal-based and model-based predictors and exploits spatial information from Ambisonic-encoded early reflections. We demonstrate that our method yields mixing time estimates that align with those obtained from complete RIRs, while significantly reducing the required computational load. Building on this estimate, we propose a gain-matching strategy that ensures smooth continuity of the RIR energy envelope, avoiding perceptual discontinuities and masking effects. The algorithm was implemented in a real-time-compatible auralization framework and evaluated on a dataset of measured and simulated room responses. Objective metrics and perceptual tests confirm that the proposed method preserves spatial impression, source localization, and immersion. This makes it suitable for applications in virtual and augmented reality, interactive audio systems, and real-time acoustic simulation. Its performance on a broader range of room geometries and acoustic scenes remains to be fully evaluated.

***Keywords*— Room Acoustics, Spatial Audio, Auralization, Virtual Reality, Reverberation, Perception.**

Résumé

Ce travail présente et évalue un algorithme d'ajustement dynamique du délai et du gain de la réverbération artificielle dans un système d'auralisation combinant la méthode des sources images avec un Feedback Delay Network (FDN). Nous proposons un estimateur du temps de mélange opérant uniquement sur les premières réflexions d'une réponse impulsionnelle de salle (RIR), ce qui permet son utilisation dans un contexte d'auralisation dynamique en temps réel où les RIR complètes ne sont pas disponibles. L'estimateur repose sur des prédicteurs à la fois basés sur le signal et sur le modèle géométrique, et exploite l'information spatiale contenue dans les premières réflexions encodées dans le format ambisonique. Nous montrons que la méthode fournit des estimations du temps de mélange cohérentes avec celles issues des RIR complètes, tout en réduisant considérablement le temps de calcul. À partir de cette estimation, nous proposons une stratégie d'ajustement du gain permettant d'assurer la continuité de l'enveloppe énergétique de la RIR, évitant ainsi les discontinuités perceptibles et les effets de masquage. L'algorithme a été implémenté dans un cadre d'auralisation en temps réel et évalué sur un ensemble de réponses impulsionnelles mesurées et simulées. Les évaluations objectives et perceptives montrent que la méthode préserve l'impression spatiale, la localisation des sources et l'immersion. Elle se prête ainsi à des applications en réalité virtuelle et augmentée, dans les systèmes audio interactifs et la simulation acoustique en temps réel. La généralisation de cette méthode à une grande diversité de géométries et de scènes acoustiques demande à être validée sur une base de données plus complète.

***Mots clés*— Acoustique des salles, Audio spatialisé, Auralisation, Réalité virtuelle, Réverberation, Perception.**

Contents

Chapter 1

Introduction

This report presents the work that I conducted during my internship at the Institut Jean le Rond ∂'alembert (Sorbonne Université) as part of the ATIAM Master 2 (IRCAM, Sorbonne Université, Télécom Paris).

1.1 Context of the internship

My internship took place in the "Sound and Space" group of the Lutheries Acoustique Musique (LAM) team, located on the Jussieu campus, from March 3 to August 11, 2025, under the supervision of Brian F.G. Katz and David Poirier-Quinot. The Sound and Space group conducts research on room acoustics and spatial sound, with a focus on real-time dynamic rendering of acoustic scenes. The ATIAM Master's degree is part of my research curriculum at École Normale Supérieure de Paris. This internship served as an introduction to the research field I will specialize in. I will be starting a PhD position in October 2025 at IRCAM under the supervision of Benoît Alary, Olivier Warusfel and Carlos Agon on the topic of room acoustic simulations and spatial audio technologies.

1.2 Scientific context and motivation

The auralization of acoustics aims at reproducing the perceptual experience of sound by simulating its propagation from a source to a receiver in a room. With the rise of immersive and binaural sound technologies, there is growing demand for accurate virtual acoustics methods for applications such as Augmented Reality and Virtual Reality (AR/VR) [43], archeoacoustics [42], video games, and virtual telepresence [5, 24]. In these scenarios, virtual sound sources must seamlessly integrate with real or virtual environments, allowing the listener to freely explore the scene while the simulation is dynamically updated.

The real-time dynamic auralization of acoustic spaces poses computational challenges. In fact, precise numerical simulations of the wave equation are generally incompatible with real time requirements and room acoustic modelers rely on geometric simplifications of the wave phenomena, treating the sound waves as sound rays [49]. Real time auralization engines also rely on the simplified Room Impulse Response (RIR) model depicted in Fig. 1.1. A Room Impulse Response between a source and a receiver is the time-domain pressure signal obtained at the receiver after a Dirac excitation at the source. It can be used to reproduce the reverberation effect by means of a convolution with an anechoic signal [23]. In this model, the RIR consists in three parts:

- the direct sound, arriving at time t_0 (with respect to the sound source emission at time $t = 0$);
- the early reflections, which consist in sparse echoes of the direct sound arriving typically 20 to 200 milliseconds after it;
- the late reverberation, starting at time t_1, when too many reflections start arriving at once at the receiver so that they cannot be perceptually separated. After t_1, the reverberation is completely characterized by its time-frequency energy envelope, following an exponential decay.

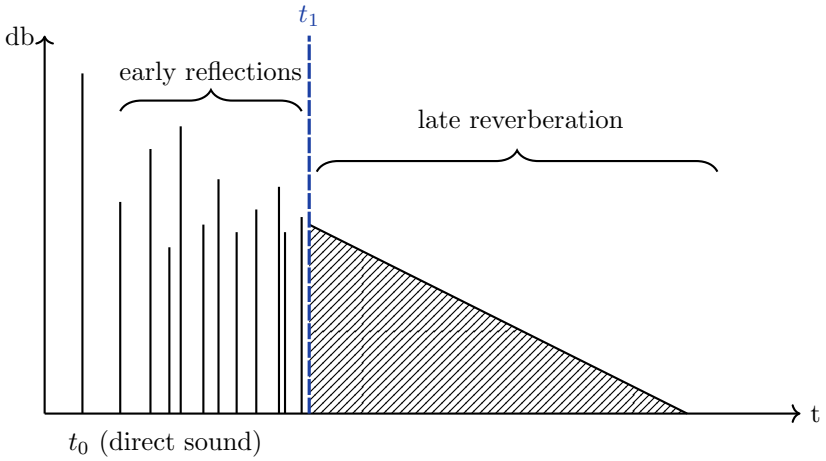


Figure 1.1: Room impulse response model [23].

This model is a simplification that is not valid for all room geometries or source and receiver positions. It is known that coupled volumes do not display an exponential decay, but rather a double-slope decay [30]. The closeness of the receiver to a room boundary might also cause artifacts such as comb filter effects that hinders a smooth exponential energy decay [29].

The RIR model separating the early reflections from the late reverberation is the basis of many existing auralization systems, comprising two independent modules, one for the early reflections and one for the late reverberation. In particular, a widely-used combination uses the Image Source method (ISM) to compute the early reflections and a Feedback Delay Network (FDN) to render the late reverberation [39, 43]. While the ISM is computationally challenging when computing high orders of reflections, one may compute only the first few orders and have an artificial reverberator produce the reverberation tail. FDNs [22] are one of the most popular artificial reverberators, giving a precise control over the frequency-band decay times of the reverberation. Because the late reverberation is completely determined by its exponentially-decaying energy envelope, according to the stochastic model, and must not be updated dynamically with source and receiver movements, artificial reverberators are computationally cheap. This duality leads to efficient systems compatible with real-time purposes.

However, the question of the transition time between the Image Sources and the FDN remains mostly unaddressed in the literature. Following the stochastic model, the transition time should be set to be the mixing time. This makes sense perceptively if we assume that the mixing time is equal to the perceptual mixing time. The perceptual mixing time t_{pm} is defined as the smallest time τ at which the impulse response is perceptually indistiguishable from the same impulse response but with the reverberation tail after τ replaced by a generic reverberation tail with the same gain and slope [29]. Generally, no spatially precise modeling is required after the mixing time to preserve plausibility and source localization. The RIR manipulation process is illustrated in Fig. 1.2. Here, the reverberation tail is replaced by Gaussian white noise modulated by an amplitude envelope with the same decay slope as that of the RIR, as it has been shown that this is perceptually similar to a room's late reverberation [36, 23].

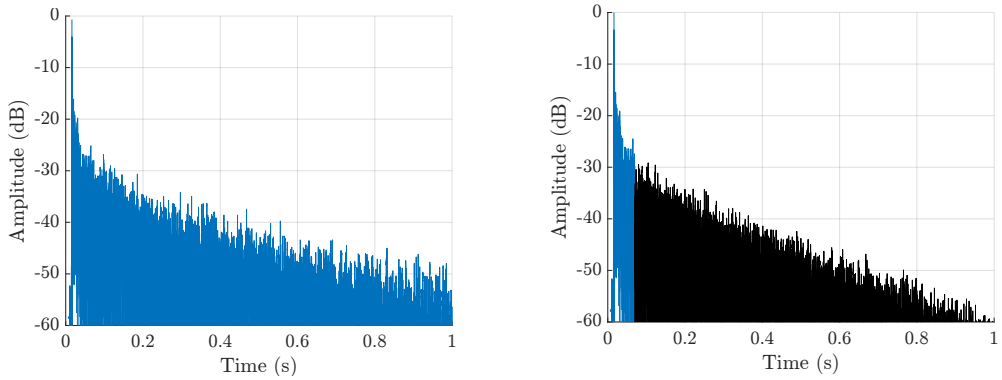


Figure 1.2: Measured RIR (left) and the same RIR with tail replaced by an exponentially decaying noise (right).

However, the mixing time is not the transition time used in most implementations. This is because no accurate method exists to estimate the mixing time based solely on model parameters on of the first early reflections. Fur-

thermore, computing early reflections up to the mixing time might have a high computational cost, for instance in very slowly mixing rooms. For instance, a classical value for the mixing time in a room with a volume of around 2000 m^3 and a total surface area of around 1000 m^2 is 200 milliseconds; that might require computing up to 10 orders of reflections. Instead, for implementation eases and to reduce the number of image sources that must be computed, the FDN delay is usually set to an arbitrary number around 40-80 milliseconds [36], or simply to the initialization time of the FDN [29]. This implementation, lacking theoretical justification, may degrade auralization quality, as the FDN might mask the image sources if introduced too early, and thus prevent source localization [18].

The question of the FDN gain is also not widely addressed in the literature. Reverberation gain critically influences spatial impression, such that a higher gain can create the illusion of a larger space, even if the decay time remains unchanged [27]. It may also bias distance perception by altering the early-to-late energy ratio, which is correlated to perceived distance [35]. In most systems, there is a user control over the gain [5], but no algorithm that tries to optimize the gain for maximal plausibility and immersivity. Like the reverberation delay, a reverberation gain set too high might create a masking effect masking the precise spatial information contained in the image sources. Conversely, a reverberation gain set too low will create a perceptually implausible effect of absence of reverberation.

Adjusting the delay and gain of the artificial late reverberation is critical for creating efficient and immersive auralization systems. The solution proposed must be based on theoretical considerations. More precisely, to match the stochastic model and its perceptual consequences, the FDN delay should be set to the mixing time, while making sure that the FDN initialization time is compensated for. The question of estimating the mixing time from model parameters and/or early reflections is thus critical and constitutes the gist of this work. Following the stochastic model, the gain value may be deduced based on the gain of the early reflections around the mixing time, with the idea of guaranteeing the smoothness of the exponential decay curve at all frequencies.

1.3 Aim of the internship

The aim of this work is to propose an algorithm for adjusting the delay and the gain of an artificial reverberator so that it smoothly integrates to an image source simulation. This algorithm should have theoretical foundations based on the physical properties of reverberant sound fields. It should be compatible with a real-time implementation for integration into a dynamic auralization framework. The goal is to enhance the auralization framework by improving localization, plausibility and immersivity. Its performance will be evaluated both with objective metrics and with perceptual tests. The work intersects acoustics (sound propagation modeling), computer science (real-time algorithm optimization), signal processing (implementation in the Max softmare for real-time auralization), and auditory perception (subjective evaluation of the auralization).

The structure of this report mirrors the successive steps of the research. First, a test dataset was gathered, comprising both measured RIRs and room models for which RIRs were simulated using different simulation methods and parameters (Chapter 2). Then, the main work of this internship was focused on the problem of mixing time estimation. Chapter 3 describes methods for the estimation of the mixing time both on complete RIRs and on early reflections only, based on theoretical properties. Mixing time estimation is the first step of the algorithm developed for the delay and gain adjustment of the artificial reverberation, which is described in Chapter 4. The following two chapters are dedicated to the evaluation of the algorithm, based first on objective metrics (Chapter 5) and on a perceptual evaluation (Chapter 6). Finally, remaining fields of inquiry and open questions are discussed in Chapter 7.

Chapter 2

Dataset and simulation methods

The test dataset is used through the study to test and quantify the behavior of the diffuseness metrics. It consists in two types of data:

- Room Impulse Responses (RIRs) measured in various existing rooms
- Geometrical acoustic models that can serve as input of room acoustic modelers to simulate RIRs

Ideally, the room models would correspond to the rooms the measurements were made in, so that measured and simulated RIRs could be compared for given source and receiver configurations. However, such datasets are very hard to acquire in practice. A room model corresponding to a real room must be calibrated to fit the reality, as for instance the absorption coefficients of the materials are not precisely known [6]. This is a time-consuming process that was judged incompatible with the time constraints of the internship. Moreover, real rooms tend to have complex geometries that lead to computationally expensive simulations that are difficult to work with. Systematic variation of room properties in order to cover a wide range of room types and ensure a thorough evaluation of the method is difficult for real rooms and enabled by designing room models that do not correspond to real rooms. Thus, to produce simulations, room models were used that either did not correspond to existing spaces, or if they did, were not calibrated. Independently, measured RIRs were taken from open-source datasets in order to evaluate the behavior of the proposed metrics on real-world data. We will first describe the measured RIR dataset and then turn to the room models.

2.1 Measured Room Impulse Responses

The measured RIRs gathered for the study were Spatial RIRs (SRIRs). SRIRs are RIRs measured using Spherical Microphone Arrays (SMAs) in order to capture the spatial information of the sound fields. In Chapter 3, we will define metrics that take as input SRIRs and make use of the spatial information contained in them in order to estimate the mixing time. The SRIRs were encoded into higher-order Ambisonic (HOA). The HOA formalism relies on the spherical harmonic decomposition of signals, where spherical sound fields are expressed in a basis of spherical functions. A B-format signal comprises $(N+1)^2$ channels which correspond to the coefficients of the SH expansions, where N is the Ambisonic order, with the convention that the first $(M+1)^2$ channels always correspond to the Mth-order encoding of the signal (the first 4 channels correspond to the 1st order encoding, the first 9 to the 2nd order encoding, etc.).

SRIRs were gathered from three academic datasets:

- McKenzie et al. [34] released a dataset of SRIRs measured with an Eigenmike and encoded into 4th-order Ambisonic. These were measured in a variable acoustics room with five different levels of absorption. In the following, those are treated as five different rooms referred to as Variable 0%, Variable 25%, Variable 50%, Variable 75% and Variable 100% and 21 RIRs per level of absorption were analyzed.
- The Motus dataset Götz et al. [19] is a series of measurements made in the same room with varying furniture and source and receiver configurations. They were recorded with an Eigenmike and encoded to 4th-order Ambisonic. Five furniture configurations were selected from this dataset with four different source-to-receiver configurations per room configuration.
- Finally, an internal dataset of 13 SRIRs measured with an OctoMic in the Saint-Elizabeth church and encoded to 2nd-order Ambisonic was used.

There were in total 123 measured RIRs with various acoustic parameters whose variability is presented in Fig. 2.1. The standard acoustic parameters that were estimated are [21]:

- Early Decay Time (EDT): the reverberation time extrapolated from a linear fit between the arrival of the direct

sound and the -10 dB point, i.e. the time when the energy has decayed by 10 dB compared to the initial energy;
- 30 dB reverberation time (rt_{30}): the -60 dB reverberation time extrapolated from a linear fit between the -5 dB point and the -35 dB point.
- clarity (c_{80}): the ratio of the total energy in the first 80 milliseconds of the RIR to the total energy after 80 milliseconds:

$$c_{80} = 10 \log_{10} \left[\frac{\int_{0ms}^{80ms} h^2(\tau) d\tau}{\int_{80ms}^{+\infty} h^2(\tau) d\tau} \right] \tag{2.1}$$

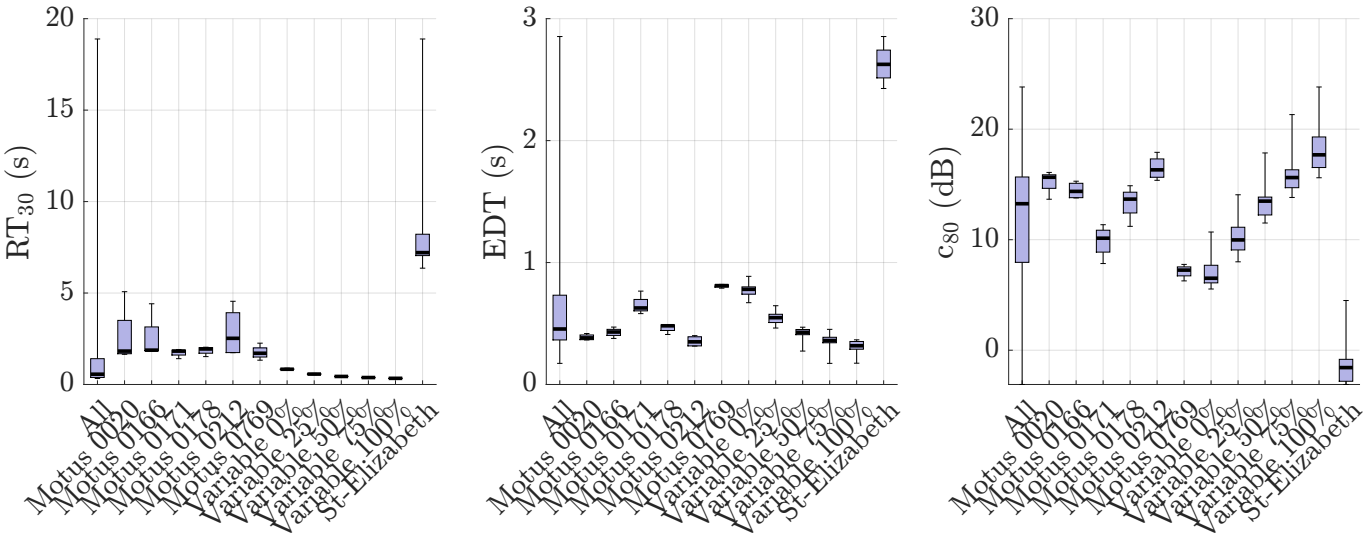


Figure 2.1: Distribution of RT_{30}, EDT and c_{80} values in total and per room.

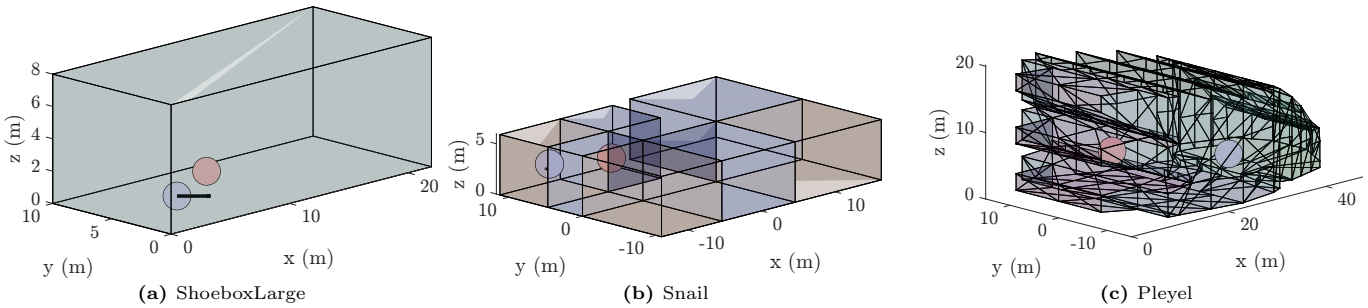
2.2 Room models and room acoustic scenes

The dataset of geometric models consisted in 16 room models with various geometries, volumes, surfaces, and wall absorption. For some rooms, several source-receiver configurations were modeled, for a total of 21 distinct room acoustic scenes. The complete configurations are listed in Table 2.1 along with the mean values.

For each configuration, the complete room acoustic scene consists in:

- Room geometry: walls and furniture positions and dimensions
- Room absorption: material for each surface with corresponding absorption coefficient per frequency band
- Source and receiver position and orientation

In terms of geometry, the room complexity shows great variability, ranging from the simplest shoebox rooms to large concerts halls and complex coupled volumes showing nonexponential decay, such as the Snail. Examples of room models are depicted on Fig. 2.2.



(a) ShoeboxLarge **(b)** Snail **(c)** Pleyel

Figure 2.2: Room models examples, with source (red) and receiver (blue) positions. Surface colors indicate various wall materials.

Thus, while relatively small, the dataset provides enough variability in the room acoustic parameters to evaluate the performance of the presented methods on various acoustic spaces with various complexities. This will enable us to estimate the level of generality of the methods in terms of geometric complexity.

Room	Rec.	Src.	Volume (m^3)	Surface (m^2)	Abs. (%)	RT_{60} (s)	S-R dist. (m)
Amst	1	1	19578	5507	17.7	2.49	21.0
Coupled	1	1	12000	3600	32.1	1.40	15.5
Cube	1	1	1000	600	50.0	0.39	5.0
Fogg	1	1	2613	1472	17.1	1.37	10.9
HalteresFurnished	1	1	2114	1635	43.2	0.49	19.2
HalteresFurnished	1	2	2114	1635	43.2	0.49	7.1
Halteres	1	1	2227	1427	43.0	0.60	19.2
Halteres	1	2	2227	1427	43.0	0.60	7.1
Morgan	1	1	2089	2679	9.1	1.12	6.3
Orsay	1	1	2122	1357	15.7	1.27	12.0
Pleyel	1	1	16342	5384	22.4	1.69	24.5
ShoeboxIsoRefl	1	1	220	238	14.2	0.97	4.9
ShoeboxIso	1	1	220	238	10.0	1.41	4.9
ShoeboxLargeIso	1	1	1760	952	10.0	2.82	4.9
ShoeboxLarge	1	1	1760	952	10.0	2.82	4.9
Shoebox	1	1	220	238	10.0	1.41	4.9
Snail	1	1	3573	2038	21.4	1.17	8.5
Snail	1	2	3573	2038	21.4	1.17	13.1
Snail	1	3	3573	2038	21.4	1.17	24.3
Snail	1	4	3573	2038	21.4	1.17	20.1
Vienne	1	1	17095	5641	15.1	2.62	32.8
Mean			**4762**	**2054**	**23.4**	**1.36**	**12.9**

Table 2.1: Room model parameters for each acoustic scene of the dataset. The 60 dB reverberation time RT_{60} is estimated with Eyring's formula.

2.3 Simulation methods: geometrical acoustics

The geometrical acoustic models, more precisely the room geometry, absorption material and source and receiver location and orientation, serve as input to room acoustic modeling algorithms aimed at simulating sound propagation. One of the focuses of this work is the comparison between different simulation methods. The various room acoustic modeling methods are generally divided into two main families [49]. The first family consists in numerically solving the wave equation using appropriate discretization schemes [7]. These methods are computationally expensive and unadapted to real-time purposes. A second family of methods, called geometrical acoustics, enables to considerably speed up computations, at the cost of physical exactitude. Sound waves are approximated as rays traveling in straight lines. Whereas in optics this assumption is mostly valid because the wavelengths are much smaller than the physical objects they reflect on, acoustical waves may be of the same order of magnitude as these objects, particularly at low frequencies [45], and in that context wave phenomena such as diffraction may not be neglected. Moreover, the phase components of the waves are neglected, so that rays do not superpose coherently and interferences are not simulated.

In the following work, we will compare two geometrical simulations methods, namely Ray Tracing and the Image Source method (ISM). Both methods were run on all the room acoustic models presented above.

2.3.1 CATT-Acoustic simulations: Ray Tracing

In the ray tracing method, inherited from geometrical optics, a certain number of rays are cast from the sound source and their individuals reflection paths are computed until they reach the receiver or their sound energy becomes too low. The rays are reflected from the walls either specularly or according to a given random law (such as Lambert's law) [55, 49]. The ray tracing method allows extensions that aim to model wave phenomena such as scattering (by enabling a division of a sound ray in many rays when it reflects on a surface) and diffraction on edges [49].

The ray tracing method is considered highly accurate for simulating room acoustics. Schroeder [55] shows the shortcomings of the traditional reverberation formulas (Sabine, Eyring and Millington) on the basis of reverberation times estimated from RIRs simulated with ray tracing. Ray tracing simulations are often considered as groundtruth for estimating room acoustic parameters and are used to study the variability of those parameters e.g. across source or receiver positions [8], since it is very easy to run many simulations while modifying only one parameter, whereas perfect control over the parameters is impossible in measurement setups. However, the accuracy of ray tracing depends

on the number of rays cast from the sound, and the computational cost grows with the number of rays cast. As a result, accurate simulations require an important amount of computations, which makes this method incompatible with real-time requirements.

The ray tracing method was not re-implemented during this internship. Instead, we relied on a closed-source commercial ray tracing software called CATT-Acoustic [11]. The CATT-Acoustic software is an accurate ray-tracer widely used in academia and industry.

2.3.2 Evert simulations: the Image Source Method

The ISM is a widely used method in the context of real-time room acoustic modeling and auralization because of its algorithmic simplicity. This method was originally introduced by Allen and Berkley [4] in the case of a rectangular room, and then extended to arbitrary room geometries [9]. It consists in recursively creating virtual mirror sources of the original sound source, positioning them symmetrically with respect to the surfaces of the room, as shown on Fig. 2.3. The image sources are virtually located outside of the room, with the property that their distance to the receiver is equal to the length of the reflection path of the sound ray from the real source to the receiver. For rigid wall boundaries in a shoebox room, the image source solution converges to the wave equation solution when the number of image sources increases [49]. The ISM assumes specular reflections: the angle of incidence equals the angle of reflection. No scattering, diffuse reflection, or surface roughness is considered, although recent implementations try to integrate these wave phenomena [57].

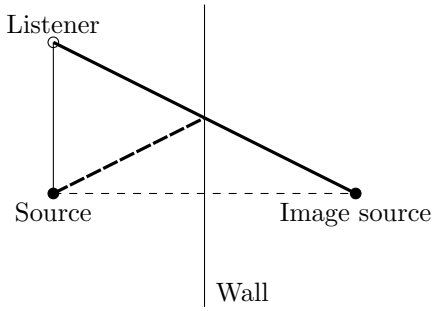


Figure 2.3: First order image source solution. The thick line from the image source to the listener represents the same distance as the path from the real source to the listener reflecting on the wall.

The image source algorithm was not re-implemented in the course of the internship. The implementation included in the IRCAM SPAT [10] software, called Evert, was used. It outputs a list of all image sources up to a certain reflection order. For each image source, its time of arrival, location, and per-frequency band sound attenuation due to distance and absorbing material are given.

The auralization of the image sources was re-implemented. The auralization consists in outputing a RIR based on the list of image sources. The monophonic (single-channel) implementation is straight-forward. For each frequency band, the image sources are added together at their respective times of arrival with the adequate attenuation. The frequency band auralizations are summed together to produce a broadband auralization. However, this neglects the receiver orientation and prevents the localization of the sound source. With the development of the Higher Order Ambisonic (HOA) formalism and of binaural technologies, spatially accurate methods have been developed for the auralization of image sources. The method implemented in SPAT and re-implemented in the course of this work consists in projecting the input signal corresponding to one image source onto the basis of spherical harmonic functions, an orthonormal basis of functions defined over the sphere. The ambisonic-encoded signal vector $\mathbf{a}(t) \in \mathbb{R}^{(N+1)^2}$ for a source signal $s(t)$ coming from direction (θ, ϕ) is given by:

$$\mathbf{a}(t) = s(t) \cdot \mathbf{y}(\theta, \phi), \tag{2.2}$$

where $\mathbf{y}(\theta, \phi)$ is the vector of $(N+1)^2$ real spherical harmonics up to order N, evaluated at azimuth θ and elevation ϕ. Each image source is thus encoded based on its direction of arrival, then all image sources are summed together to produce the final, spatialized auralization. A higher Ambisonic order leads to a higher spatial resolution. Here we chose $N = 3$, which results in 16-channel signals. In order to listen to the Ambisonic signals and perceive the desired spatial impression, one must decode them into a given loudspeaker setup or to the binaural (2-channel) format for headphone rendering [38]. Binaural decoding will be more thoroughly discussed in Chapter 5. This was implemented in SPAT and re-implemented in the course of the internship with good agreement.

Table 2.2 shows the maximum orders on which the simulations were run and the corresponding computation times and RIR durations. For practical reasons, the maximum order was chosen based on the running time in order not to exceed one hour of computation. Predicting the running time based on the room model parameters is a non-trivial problem. In the general case, the number of image sources grows exponentially with the reflection order: the number of image sources of order up to K is $\sum_{k=1}^{K} N(N-1)^{k-1}$ for N surfaces [49]. Borish [9] shows that the most costly computations come from the visibility check: given an image source, one must check whether its sound actually reaches the receiver. In a shoebox room, no visibility check is necessary, which makes the computations very efficient for those rooms, as shown by Table 2.2. In other room geometries, this visibility check influences the computation time in unpredictable ways.

Room	**Rec**	**Src**	**Max Order**	**Running time**
Amst	1	1	12	32'50
Coupled	1	1	20	5'39
Cube	1	1	20	0'26
Fogg	1	1	4	26'56
HalteresFurnished	1	1	11	22'42
HalteresFurnished	1	2	10	11'20
Halteres	1	1	14	28'40
Halteres	1	2	14	30'17
Morgan	1	1	6	1h15'41
Orsay	1	1	12	50'06
Pleyel	1	1	10	57'05
ShoeboxIsoRefl	1	1	18	15'11
ShoeboxIso	1	1	18	15'14
ShoeboxLargeIso	1	1	12	13'56
ShoeboxLarge	1	1	20	0'29
Shoebox	1	1	20	0'28
Snail	1	1	13	18'28
Snail	1	2	14	13'45
Snail	1	3	14	12'42
Snail	1	4	13	14'15
Vienne	1	1	12	35'46

Table 2.2: Simulation parameters and computation times for each room acoustic scene.

Thus, ray tracing and image source simulations were run on the same room models, yielding 3rd-order Ambisonic RIRs. Due to the exponential increase in the number of image sources with reflection order, the ISM quickly becomes computationally infeasible for high orders. Consequently, it is typically limited to modeling only early reflections. In contrast, ray tracing approximates the entire RIR by simulating a large number of rays that propagate and reflect within the environment. The accuracy and computational cost of ray tracing depend heavily on the number of rays emitted. Using too few rays risks missing some sound paths, which can lead to an underrepresentation of energy and reflection density in the simulated RIR. However, the ISM deterministically computes all specular reflection paths up to the chosen order, ensuring completeness within that limit. Moreover, the ISM models only specular reflections, whereas ray tracing naturally accommodates more complex phenomena such as scattering by allowing reflections with randomized incidence directions, thus better capturing diffuse reflection effects. The differences between the two simulation methods are summarized in Table 2.3.

These datasets will serve as a basis to evalute the metrics and algorithms that will be defined throughout this work, with an emphasis on the contrasted behaviors between measured and simulated RIRs and between ISM and ray tracing simulated RIRs.

	Image Source Method	**Ray Tracing**
Computational Complexity	Exponentially grows with reflection order; becomes intractable at high orders.	Depends on number of rays; scales linearly with rays cast.
Reflection Order	Computes all specular reflections up to a specified order exactly.	Approximates reflections probabilistically; can model higher-order reflections with enough rays.
Completeness	Guaranteed to find all reflection paths up to the chosen order.	May miss some paths if too few rays are used, leading to reduced energy/density.
Reflection Type	Models only specular (mirror-like) reflections.	Can model specular and diffuse/scattered reflections via randomized incidence.
Physical Phenomena Modeled	Limited to specular reflections and direct paths.	Includes scattering, diffraction (to some extent), and diffuse reflections.
Typical Use Case	Accurate early reflection modeling, useful for detailed analysis of initial reflections.	Simulates complete RIR including late reverberation and diffuse field effects.

Table 2.3: Comparison between Image Source Method and Ray Tracing.

Chapter 3

Mixing time estimation

In the context of a real-time auralization system combining the ISM for the early reflections and a FDN for the late reverberation, a natural choice for the FDN delay is the mixing time. In fact, the mixing time is defined as the time from which the RIR contains no more spatial information and may be fully modeled by its exponentially-decaying time-frequency envelope. The ISM produces a spatially precise auralization, where each image source is auralized based on its virtual position relative to the listener, while the FDN produces a generic auralization which takes as input only the room reverberation time and does not depend on source or receiver position or orientation. Therefore, the mixing time seems to be the perfect tradeoff between computational efficiency and perceptual authenticity. Many mixing time estimations methods have been widely addressed in the literature [2, 58, 20, 12, 29, 31], but no consensus has been reached. Moreover, mixing time estimators often take as input full RIRs and do not operate online, making them incompatible with real-time implementations. Thus, this chapter will address the question of mixing time estimation, focusing in particular on the challenge of estimating the mixing time in a real-time context, i.e. when only model parameters and the early reflections are available inputs.

3.1 Theoretical background: mixing rooms, ideally diffuse sound fields and the mixing time

In order to propose estimates for the mixing time, one must define from a theoretical point of view the stochastic model and the properties associated with it, namely the properties of a diffuse sound field.

Schroeder [56] in the frequency domain and Polack [44] in the time domain proposed a stochastic model for reverberation, which serves as the theoretical basis for artificial reverberators. The late reverberation may be modeled as the realization of a Gaussian random process:

$$h(t) = b(t)e^{-\delta t}, \tag{3.1}$$

where $b(t)$ is a centered stationary Gaussian white noise and δ the damping factor. A dependence to frequency may be added by modeling the impulse response's time-frequency energy envelope:

$$W_h(t, f) = P(f)e^{-\delta(f)t}. \tag{3.2}$$

There are two frequency-dependent parameters, the noise initial power spectrum $P(f)$ and the damping factor $\delta(f)$. This stochastic model is based on two densities: modal density and echo density [53, 23].

$$D_m(f) \approx 4\pi V \frac{f^2}{c^3}, \tag{3.3}$$

$$D_e(t) \approx 4\pi c^3 \frac{t^2}{V}, \tag{3.4}$$

where c is the speed of sound, V the room's volume and t is the time since the sound source emission.

The stochastic model is valid for high modal and echo densities. In the frequency domain, the stastical model relies on the assumption of high modal overlap, which is not verified at low frequencies [23]. The cutoff frequency, based on the criterion that the average spacing between normal frequencies must be less than one third of the bandwidth of a mode, is called the Schroeder frequency [53]:

$$f_{sch} \approx 2000\sqrt{\frac{RT}{V}} \text{ Hz}, \tag{3.5}$$

where RT is the -60 dB reverberation time in seconds and V the room volume in m^3. The Schroeder frequency marks the transition from individual, well-separated resonances to many overlapping normal modes.

In the time domain, the stochastic model assumes a high density of arriving reflections at any point. The transition time between early individual reflections and a diffuse sound field is called the physical mixing time t_m and is a property of the room. The threshold for sufficient echo density D_e to reach the physical mixing time remains to be determined.

Eq. 3.4 shows that the density of echos increases with time after a Dirac excitation. After a sufficient time, if the right conditions are met, there are many reflections arriving from every direction at any point of the enclosure. This property is called diffuseness, or mixing. The physical mixing time is the time required for a sound field to become diffuse after a Dirac excitation; it is the duration of the diffusion process [29], during which the sound energy continuously spreads over the whole volume.

An ideally diffuse sound field satisfies the following properties [26, 29]:

- isotropy: a uniform angular distribution of sound energy flux (rate of energy flow per unit area) at any point
- homogeneity: a constant acoustical energy density (energy per unit volume) over the whole space

Mixing is linked to the stochastic model because during mixing, the sound field becomes increasingly stochastic over time. A condition for a room to be mixing, i.e. to reach an ideally diffuse sound field after a sufficient amount of time, is its ergodicity [45]. Ergodicity ensures that long-term time-averaged measurements (e.g., decay rates, energy distribution) are representative of the entire room. The statistical behavior at all points in the space equals that at one point over time. After a sufficiently long time, the system's state becomes statistically independent of its starting configuration.

The theoretical conditions for a room's ergodicity are not fully understood. Some room geometries are known to be non-mixing: for example, perfectly rectangular rooms, since particles traveling in them take only eight possible directions and thus might not explore the whole space.

In practice, an ideally diffuse sound field is never reached. In a reverberant chamber, the diffuse-field conditions are not perfectly fulfilled due to the statistical superposition of wall reflections (time domain) or of room modes (frequency domain) [37]. Polack [45] showed that absorbing rooms can never be perfectly isotropic because there always remains a net energy flow towards the absorbing walls, as is illustrated by Fig. 3.1. Diffuse-field conditions are not usually met everywhere in the enclosure, as there are for instance comb filters caused by the nearness of room boundaries.

Therefore, although the diffuse-field properties (i.e. isotropy and homogeneity) can serve as a basis for measuring the diffusion process, they are never completely met. Thus, we must rely on perceptual thresholds to estimate a mixing time based on these properties.

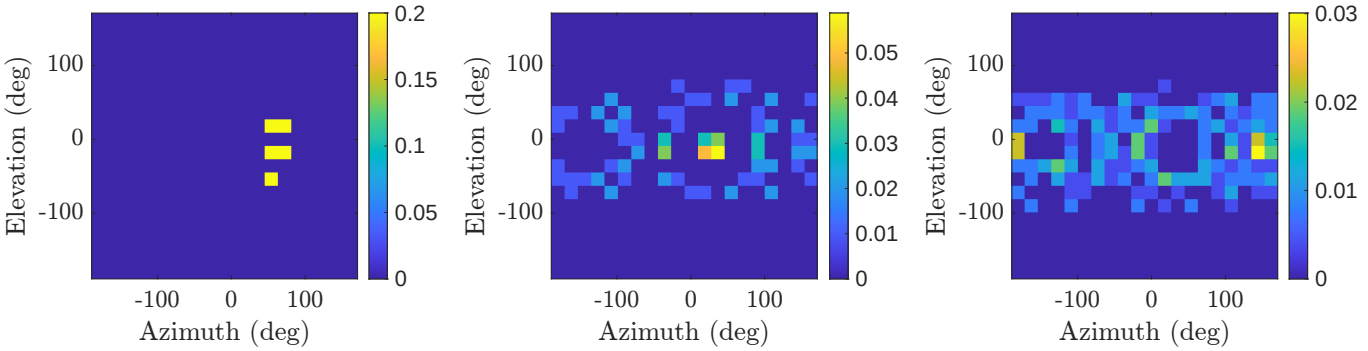


Figure 3.1: Angular distribution of energy at the receiver during an image source simulation, at time windows 43-64 ms (left), 234-256 ms (center), and 405-426 ms (right).

3.2 Model-based mixing time estimators

Model-based estimators of the mixing time take as input one or several room model parameters such as volume, surface, or absorption coefficients. Lindau et al. [29] provided a comprehensive review of existing model-based estimators. A simple criterion for estimating the perceptual mixing time is a sufficient echo density. The echo density is given by Eq. 3.4. The echo density threshold is determined based on psychoacoustic knowledge. Based on the criterion that at least 10 reflections overlap within the auditory system's time resolution (24 ms), the mixing time may be estimated from equation 3.4 as [23]:

$$t_{mp} \approx \sqrt{V} \text{ in milliseconds,} \tag{3.6}$$

where V is the room volume in m^3. Other model-based estimators of the perceptual mixing time have been proposed, for instance:

$$t_{mp} \approx 47 \times \frac{V}{S}, \tag{3.7}$$

where S is the room surface area in m^2 [48].

Model-based estimators provide computationally cheap estimates of the mixing time based on room properties. After determining perceptual mixing times on a dataset of measured RIRs via a perceptual test, Lindau et al. [29] showed that $\sqrt{V}$ and V correlated linearly with perceptual mixing times with statistical significance on their dataset. However, their dataset was limited to classical shoebox rooms with uniform distribution of absorption, which are also the types of rooms from which the model-based formulas have been derived [23]. It is expected that the model-based estimators will perform poorly on non-Sabinian rooms [46, 60], i.e. rooms that do not obey the assumptions of Sabine's formula (uniform distribution of absorption across surfaces, homogeneous and isotropic sound field, and exponential energy decay). More refined estimation methods, based on RIRs, have been proposed.

3.3 Signal-based mixing time estimators: echo density

Several signal-based methods have been proposed to estimate the mixing time from RIRs. A comprehensive review is provided by Lindau et al. [29], comparing the perceptual relevance of various estimators. Among those reviewed, the method of Abel and Huang [2] shows the strongest correlation with perceptual mixing time and is thus the primary focus of our discussion. Other approaches such as those by Defrance et al. [12], Stewart and Sandler [58], Hidaka et al. [20] perform less well in this regard.

The echo density profile $\eta(t)$ proposed by Abel and Huang [2] quantifies the time-varying diffuseness of an RIR. Their approach is based on the assumption that, in a fully mixed (diffuse) field, the pressure amplitude values in a short window follow a Gaussian distribution. The profile measures the fraction of samples in a window that exceed one standard deviation, normalized by the expected value under the Gaussian assumption.

The echo density at time t is given by:

$$\eta(t) = \frac{1}{\operatorname{erfc}(1/\sqrt{2})} \sum_{\tau=t-\delta}^{t+\delta} w(\tau)\mathbf{1}\{|h(\tau)| > \sigma(t)\}, \tag{3.8}$$

where $h(t)$ is the room impulse response, $w(\tau)$ is a positive windowing function (typically rectangular or Hanning), $2\delta + 1$ is the window length in samples, $\sigma(t)$ is the windowed standard deviation of h defined as:

$$\sigma(t) = \left[\sum_{\tau=t-\delta}^{t+\delta} w(\tau)h^2(\tau) \right]^{1/2}, \tag{3.9}$$

and $\operatorname{erfc}(1/\sqrt{2}) \approx 0.3173$ is the expected proportion of samples outside one standard deviation in a standard Gaussian distribution. The window duration is typically around 20 ms, long enough to contain at least a few reflections within each frame and short enough to be psychoacoustically relevant.

This profile starts near zero in the early part of the RIR (where few strong reflections dominate, so the standard deviation is high) and increases as overlapping reflections cause the pressure distribution to approximate Gaussian noise, as shown by Fig. 3.2. In fully diffuse conditions, $\eta(t) \approx 1$. Although Abel and Huang [2] do not explicitly mention the mixing time, they interpret the time at which a value of one is first attained as the start of the late field. Following this remark, Lindau et al. [29] propose using the echo density profile to estimate the mixing time, simply by setting the estimated mixing time as the first time at which the profile reaches unity. An example of a measured RIR, its corresponding echo density profile and resulting estimated t_m is given on Fig. 3.3, along with the echo density profile of measurement noise.

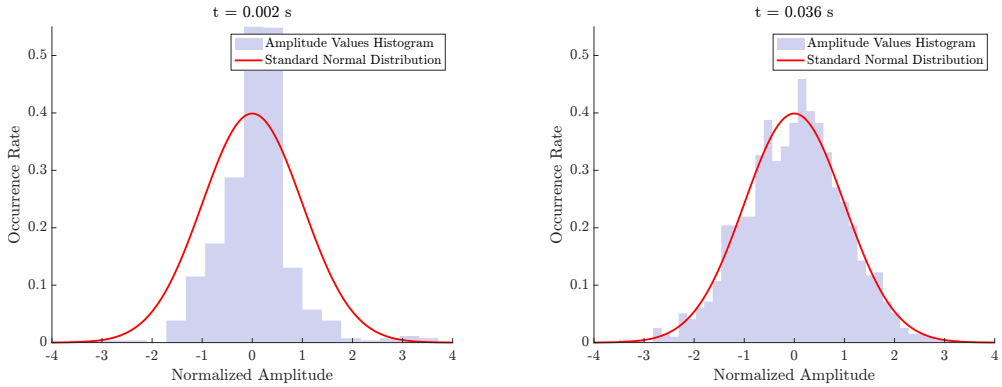


Figure 3.2: Amplitude histograms before and at the estimated mixing time

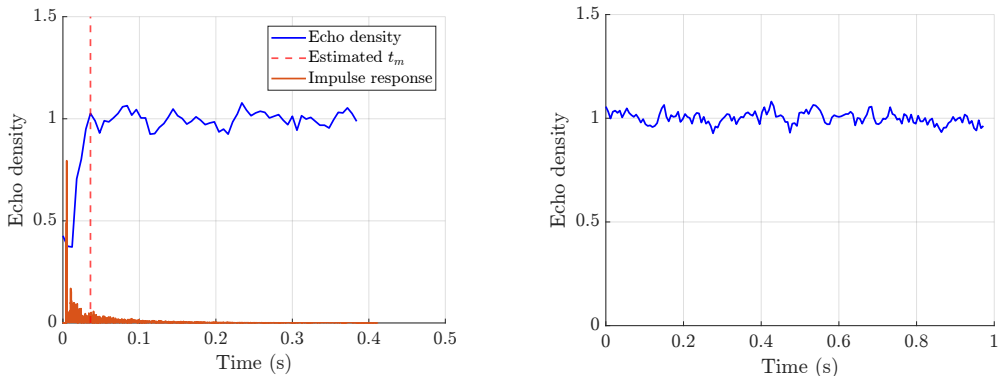


Figure 3.3: Echo density on a measured RIR (left) and on measurement noise (right).

The echo density metric shows good performance on measured RIRs but does not give satisfactory results on early RIRs simulated with the ISM. For early RIRs, the echo density metric never reaches unity on image source auralizations, as they are too sparse to take on a Gaussian distribution. In the classical implementation, the ISM does not implement scattering. Scattering is a phenomenon caused by the reflection of sound on rough surfaces, and causes the reflections to spread in space and time [57]. In contrast, the ray tracing simulations used in this study implemented scattering. Thus the echo density metric applied on ISM and ray tracing simulations for the same acoustic scene produce very different results, as can be seen on Fig. 3.4.

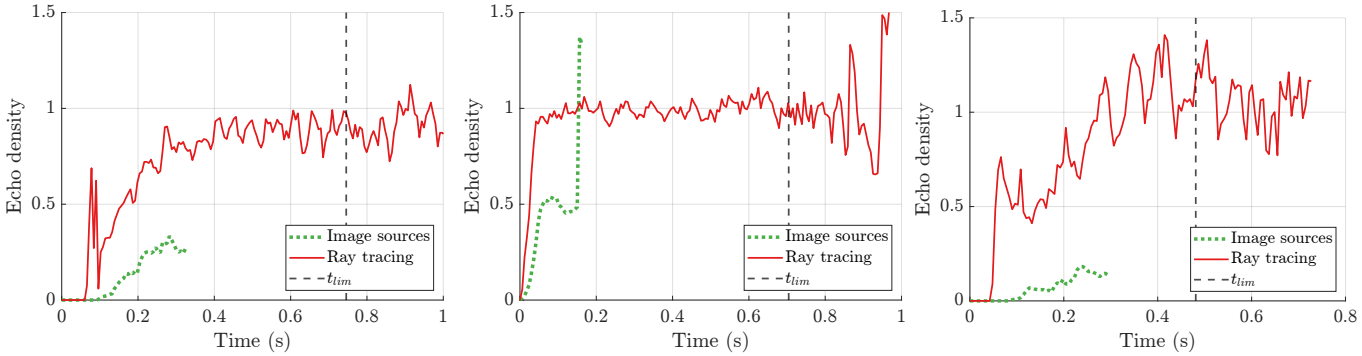


Figure 3.4: Comparisons of echo density profiles between ISM and ray tracing simulations. From left to right: Snail 1 3, ShoeboxIsoRefl 1 1, and Halteres 1 1. t_{lim} is the noise level.

As such, the echo density metric is not usable to estimate mixing time from an image source auralization, since in general the echo density metric does not reach unity on early RIRs.

3.4 Signal-based mixing time estimators: spatial incoherence

The echo density metric seems unadapted to RIRs simulated with the ISM. It relies only on the arrival times and amplitudes of the image sources and does not take into account their direction of arrival relative to the listener. Other

metrics must be proposed that fully exploit the spatial information contained in the ISM simulations.

In fact, both the ISM and ray tracing may simulate Spatial RIRs (SRIRs). SRIRs are encoded using the Spherical Harmonic formalism. The acoustic pressure measured at the spherical coordinates $\mathbf{r} = (r, \theta, \phi)$ may be expressed in the basis of the SH functions:

$$p(\mathbf{r}, k) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} i^l j_l(kr) Y_l^m(\theta, \phi) b_{l,m}(k), \tag{3.10}$$

where $k = 2\pi f/c$ is the wavenumber, j_l the order-l spherical Bessel function, and Y_l^m the order-l, degree-m, real-valued SH function. The order-L SH expansion consists in keeping the coefficients up to order L. The higher the order, the larger the area of accurate description of the sound field around the origin. The time-domain SH expansion is the inverse Fourier transform of the frequency-domain coefficients $b_{l,m}(k)$. The coefficients are classically ordered by order and then by degree, and the total number of coefficients is $(L+1)^2$. Each coefficient is a time series $b_{l,m}(t)$. This format is called Ambisonics. Typical values for L are 1 and 3, and the SH signals are thus encoded using respectively 4 and 16 channels. For $L > 1$, the format is referred to as Higher Order Ambisonics (HOA).

Epain and Jin [13] proposed to use the spatial information contained in SRIRs encoded in the SH domain in order to estimate the spatial incoherence of the sound field. In the case of an ideally diffuse sound field, which may be modeled by Gaussian noise or by an infinite number of uncorrelated plane waves, the covariance matrix of the SH signal channels is close to $\rho\mathbf{I}_{(L+1)^2}$, where $\mathbf{I}$ is the identity matrix, which means that the SH signals are mutually uncorrelated. This is due to the fact that the sound field, being perfectly diffuse, is spatially incoherent, with plane waves arriving from every direction. On the contrary, in the case of a maximally non-diffuse sound field, consisting for instance of only one plane wave or of several correlated plane waves, the SH signals with have non-zero covariance outside of the diagonal. The homogeneity of the eigenvalues of the covariance matrix is a good indicator for the degree of spatial incoherence caputed by the SH signals. This is illustrated by Fig. 3.5.

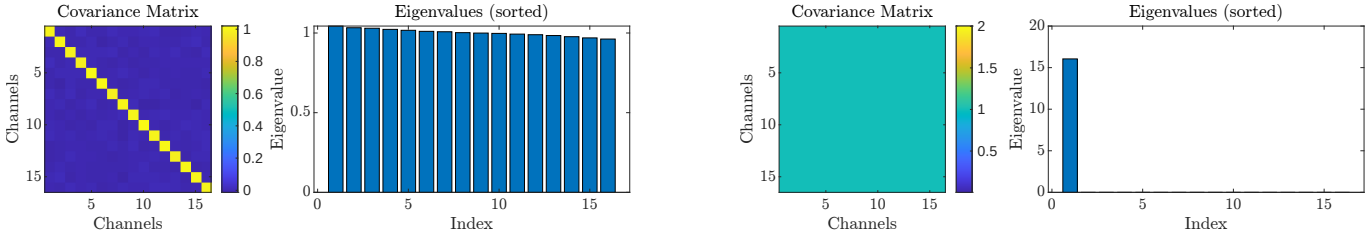


Figure 3.5: SH signals covariance matrix and eigenvalues for uncorrelated Gaussian white noise (left) and perfectly correlated Gaussian white noise (right).

As a result, the spatial incoherence metric coined CoMEDie is defined as [13]:

$$d = 1 - \frac{\gamma}{\gamma_0}, \tag{3.11}$$

where γ is the deviation of the eigenvalues of the covariance matrix from their mean:

$$\gamma = \frac{1}{\langle v \rangle} \sum_{i=1}^{(L+1)^2} |v_i - \langle v \rangle|, \quad \text{with } \langle v \rangle = \frac{1}{(L+1)^2} \sum_{i=1}^{(L+1)^2} v_i \tag{3.12}$$

and $\gamma_0 = 2\left[(L+1)^2 - 1\right]$ the value in the most non-diffuse case.

If we simulate an ideally diffuse sound field by mutually uncorrelated Gaussian white noise across the SH channels, we obtain a diffuseness value close to 1, since the covariance matrix approximates $\rho\mathbf{I}_{(L+1)^2}$ and its eigenvalues are all close to ρ. The maximum value for this metric is thus 1 ($\gamma = 0$).

Massé [31] extended the CoMEDiE measure to a spatial incoherence profile by proposing to apply it on a short sliding window of the RIR. A typical value for a window length is 24 milliseconds. This results in a spatial incoherence profile $\Gamma(t)$. Examples of the spatial incoherence profile of two measured RIRs and of measurement noise are given on Fig. 3.6.

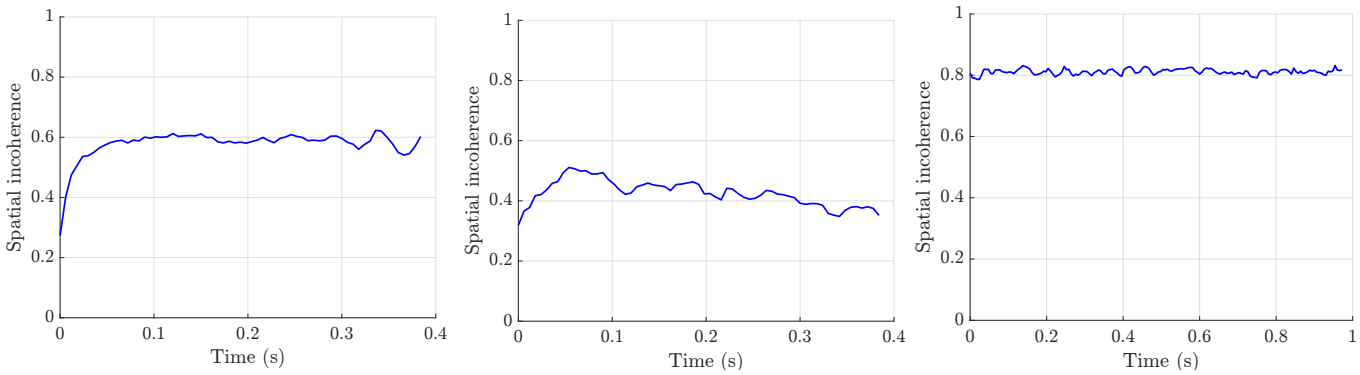


Figure 3.6: Spatial incoherence on two measured RIRs (left, center) and on measurement noise (right).

As such, the spatial incoherence metric is more adapted to ISM simulations than the echo density metric. In fact, the spatial incoherence profile of the ISM simulations closely matches that of the ray tracing simulations for the same room acoustic scene, as can be seen on Fig. 3.7. The RMS error between the ray tracing and the ISM simulations averages only 0.09 for the spatial incoherence metric whereas it is 0.4 for the echo density metric. Fig. 3.9 shows closely matching spatial incoherence values. This can be explained by the fact that the spatial incoherence metric makes full use of the spatial information contained in image source simulations in order to capture the phenomenom of increasing diffuseness. Although the two simulations methods (ray tracing and image source) have critical differences, such as the fact that ray tracing takes into account scattering while image source assumes purely specular reflections, the spatial incoherence metric, by using the spatial information contained in the image sources, manages to overcome the limitations of image source and capture the increasing diffuseness of the sound field even after a few reflections.

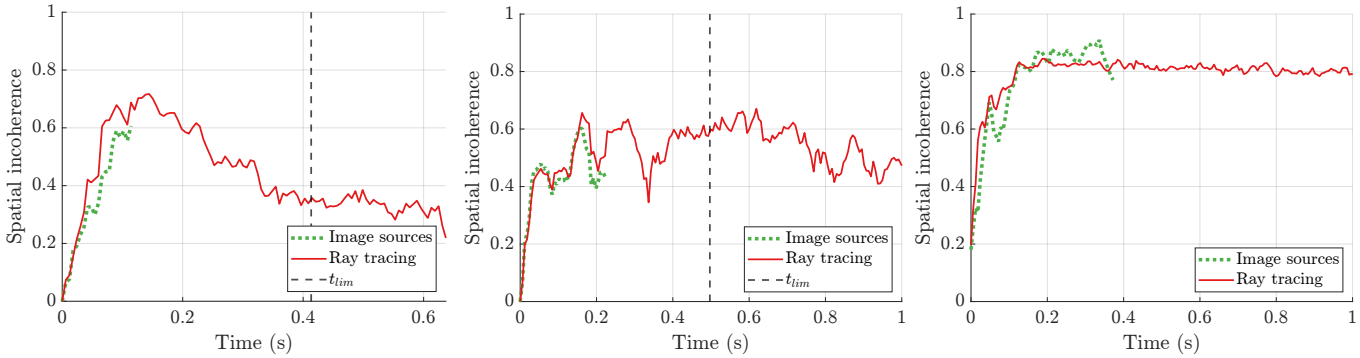


Figure 3.7: Spatial incoherence profiles computed on ISM and ray tracing simulations. From left to right: HalteresFurnished 1 1, Snail 1 1, and ShoeboxLarge 1 1.

Unlike the echo density measure, the spatial incoherence profile cannot be straightforwardly used to estimate the mixing time. In fact, as we can see on Figs. 3.6 and 3.7, a clear plateau is not always reached, and when it is, there is no predictable value for the spatial incoherence plateau. The spatial incoherence value is constrained not to exceed unity, but since the sound field is never perfectly diffuse, there is no guarantee that it will reach a certain value in all cases. The spatial incoherence metric was applied on the dataset of measured RIRs and simulations presented in Chapter 2 with great variability. Table 3.1 shows the global spatial incoherence values (i.e. the spatial incoherence metric applied on the full RIR with no windowing) for all room acoustic scenes and the mean spatial incoherence values per measured room. We observe great variability in the case of the simulated RIRs, with values ranging from 0.2 to 0.84. For the measured RIRs, we observe slightly less variability, with values ranging from 0.31 to 0.58.

As higher orders induce higher spatial resolution, the metric is sensitive to Ambisonic order. This was already observed by Epain and Jin [13] who noted that the sound field generated by a few uncorrelated sources evenly distributed in space might be interpreted as diffuse when looking only at the order-1 Ambisonic signals, while the spatial incoherence profile of the order-3 Ambisonic signals will show the sound field coherence. Typically, when several uncorrelated sources are located in directions opposite to each other, higher Ambisonic orders will lead to smaller spatial incoherence values: the sound field looks diffuse when looking only at the lower order Ambisonic signals.

Figs. 3.8 and 3.9 show the influence of Ambisonic order on the spatial incoherence value for both measured and simulated RIRs. In all cases, the spatial incoherence value tends to decrease as the Ambisonic order increases. Examples of spatial incoherence profiles confirm this observation (Figs. 3.11, 3.10). The figures show that the higher

order signals have a smoother profiles and tend to decrease variability within a room. Therefore, the 3rd-order signals were chosen for mixing time estimation. A more thorough analysis of the influence of Ambisonic order and correlation with perception of diffuseness is left for future work.

Room	Rec.	Src.	S.I. (rt)	S.I. (IS)
Amst	1	1		0.58
Coupled	1	1		0.42
Cube	1	1		0.20
Fogg	1	1		0.34
HalteresFurnished	1	1	0.42	0.38
HalteresFurnished	1	2	0.34	0.34
Halteres	1	1	0.45	0.41
Halteres	1	2	0.34	0.30
Morgan	1	1	0.41	0.32
Orsay	1	1	0.56	0.49
Pleyel	1	1	0.64	0.54
ShoeboxIsoRefl	1	1	0.69	0.73
ShoeboxIso	1	1	0.76	0.84
ShoeboxLargeIso	1	1	0.66	0.78
ShoeboxLarge	1	1	0.72	0.80
Shoebox	1	1	0.78	0.84
Snail	1	1	0.44	0.49
Snail	1	2	0.48	0.53
Snail	1	3	0.56	0.57
Snail	1	4	0.48	0.47
Vienne	1	1	0.59	0.56

Room (# RIRs)	Mean S.I.	Min	Max
Motus 0020 (4)	0.46	0.41	0.55
Motus 0166 (4)	0.44	0.37	0.54
Motus 0171 (4)	0.46	0.39	0.55
Motus 0178 (4)	0.44	0.37	0.54
Motus 0212 (4)	0.47	0.39	0.55
Motus 0769 (4)	0.50	0.47	0.57
Variable 0% (21)	0.50	0.37	0.54
Variable 25% (21)	0.51	0.31	0.58
Variable 50% (21)	0.51	0.35	0.55
Variable 75% (21)	0.50	0.33	0.57
Variable 100% (21)	0.49	0.31	0.58

Table 3.1: Spatial incoherence of the complete RIRs simulated by ray tracing (rt) and image sources (IS) for the different room models (left) and mean spatial incoherence per room of the measured RIRs truncated at order 3 (right). For rooms Amst, Coupled, and Cube, 3-rd order ray tracing simulations were not available.

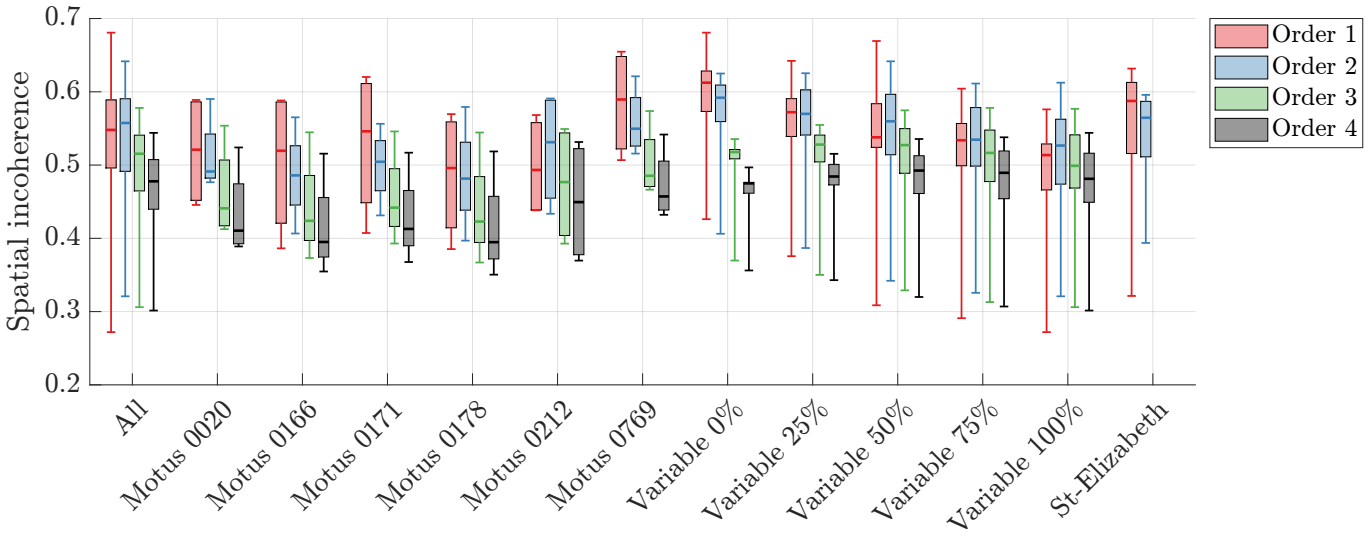


Figure 3.8: Distribution of spatial incoherence values for all measured RIRs and per room, for different Ambisonic orders.

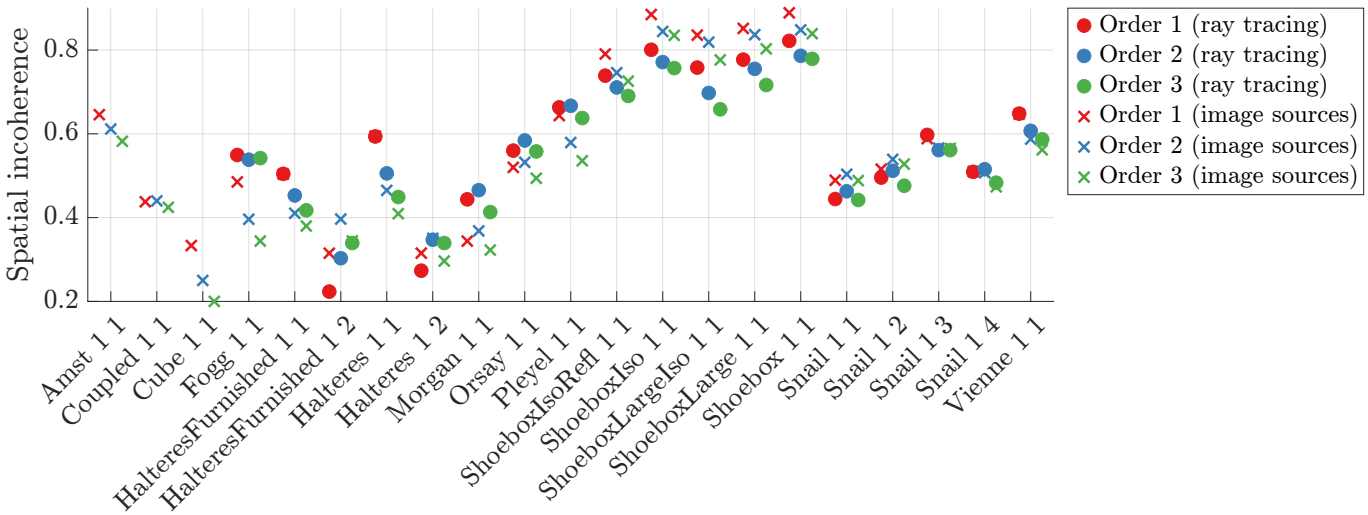


Figure 3.9: Spatial incoherence values for each configuration for ray tracing and image sources, for different Ambisonic orders.

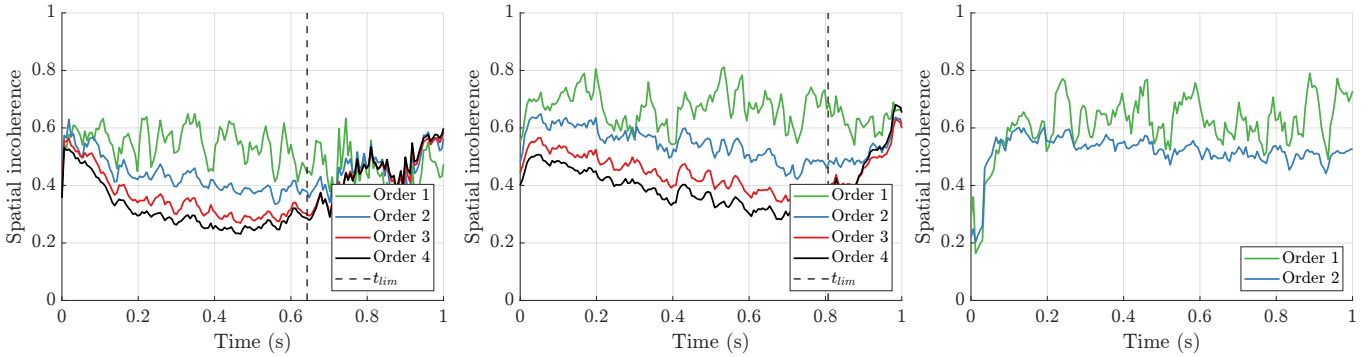


Figure 3.10: Spatial incherence metric for different Ambisonic orders on measured RIRs. From left to right: Motus 0171, Variable 0%, and Saint-Elizabeth.

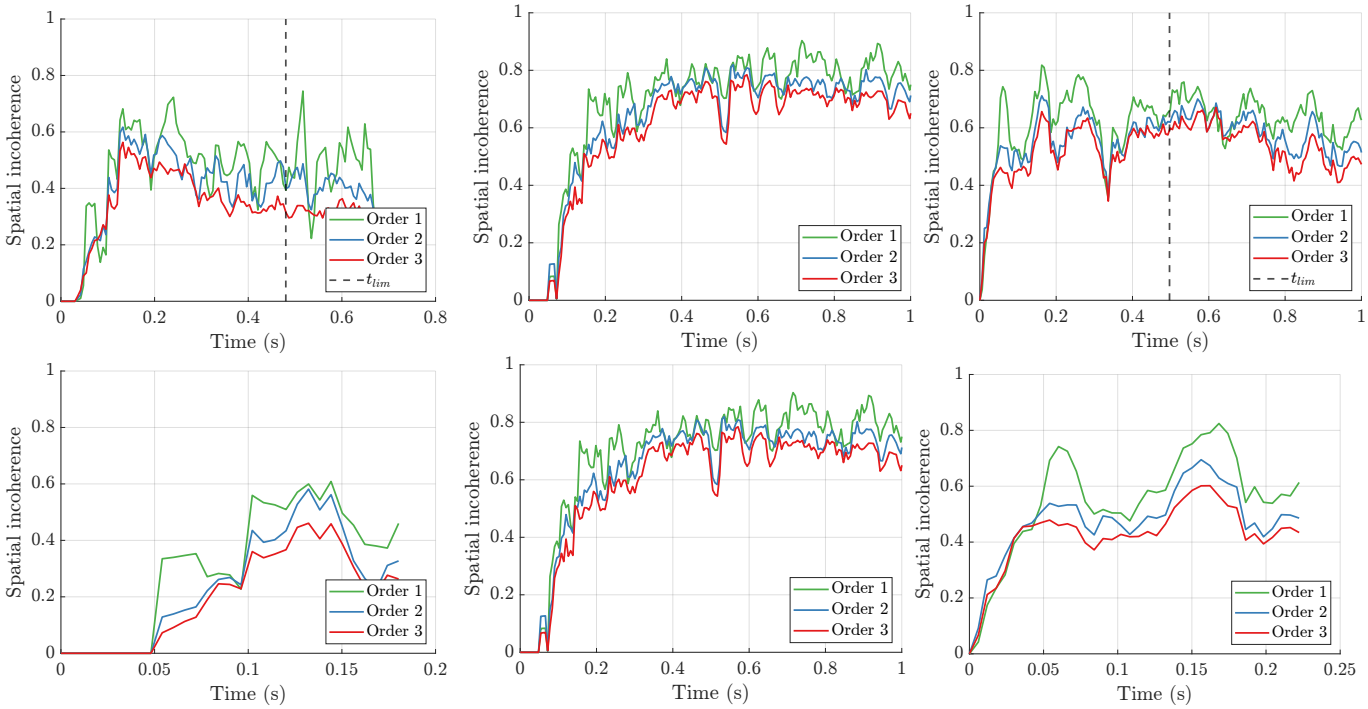


Figure 3.11: Spatial incherence metric for different Ambisonic orders on ray tracing (top) and image source (bottom) simulations. From left to right: HalteresFurnished 1 1, Pleyel 1 1, and Snail 1 1.

Due to the variability in spatial incoherence measure, the general non-smoothness of the profiles and the sensitivity to Ambisonic order, there exists no straightforward algorithm to deduce the mixing time from the spatial incoherence profile. Massé [31] proposed an algorithm to estimate the moment when $\Gamma(t)$ levels out at a near-maximal value after the initial instability due to coherent early reflections. The algorithm consists in segmenting the spatial incoherence profile using an adaptive Ramer-Douglas-Peucker (RDP) algorithm, assigning a score to each segment based on its length, slope and average incoherence, and estimating t_m at the time of the segment with the highest score. The goal is to find the moment the diffuseness profile reaches a maximum value that is maintained throughout the late reverberation tail. This algorithm shows mixed results and is incompatible with a real-time implementation. Other techniques for detecting the start of a spatial incoherence plateau include detecting when the derivative of the profile is close to zero, but this is not robust to small fluctuations in the profile.

Fig. 3.8 hints at the fact that the spatial incoherence value might be linked to properties of the room. In fact, in the case of the variable-acoustics room [34], the spatial incoherence tends to decrease with increased absorption. This correlation is particularly visible for lower Ambisonic orders. The overall level of absorption might not so much have an impact on the spatial incoherence as the uniformity of the distribution of absorbing surfaces. Polack [45] showed that absorbing rooms can never be perfectly diffuse because there is always an energy flow in the direction of the absorbing walls [29]. As such, a non-uniform repartition of absorption is expected to reduce spatial incoherence as rays might travel in coherent directions [60]. Regular room shapes and specular reflections also tend to favor coherent ray trajectories [45]. This can explain the very low spatial incoherence value of the cubic room (Table 3.1).

For these reasons, some work has been devoted to try to predict the value of the spatial incoherence plateau for a given room based on model parameters such as volume, surface, or the repartition of absorption. Preliminary results on the dataset of room acoustic scenes are a quadratic correlation between the room's total surface area and the spatial incoherence value, with a R^2 value of 69%. The associated quadratic formula is:

$$d(S) = 4.7e^{-8}\ S^2 - 3e^{-4}\ S + 0.82, \tag{3.13}$$

where S is the surface and $d(S)$ the estimated global spatial incoherence. The quadratic curve against the data points is shown on Fig. 3.12. Other parameters that were tested are the source-receiver distance, V, $\sqrt{V}$, $\frac{V}{S}$, and the Sabine and Eyring reverberation times; no satisfactory linear or quadratic correlation was found for these parameters.

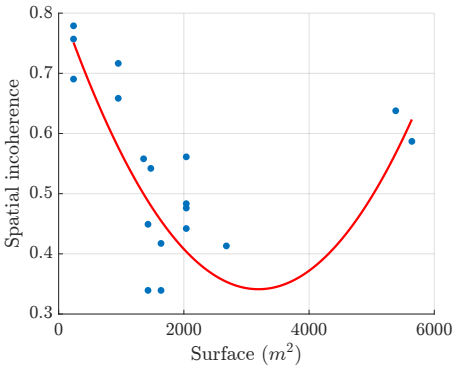


Figure 3.12: Spatial incoherence of the ray tracing simulations against room surface area and best quadratic fit on the dataset of room acoustic scenes.

This is a preliminary result that is not expected to generalize well to any room geometry but that will serve as a basis to propose a mixed mixing time estimation method based on both model and signal parameters. In future work, a more thorough study on a larger dataset of measured RIRs and room models will be conducted with a specific focus on understanding the relationship between the spatial incoherence value and the repartition of absorption and the shape of the enclosure.

This formula gives a prediction of the spatial incoherence value of a room. This enables to derive a real-time compatible mixing time estimation algorithm from the spatial incoherence profile. One can estimate the mixing time as the first time the spatial incoherence value exceeds $d(S) + \epsilon$, where ϵ is a very small adjustable value. We thus obtain a new complete mixing time estimator that operates on early RIRs for rooms whose surface area is known.

3.5 Results

The evaluation of mixing time estimation results is not straightforward because there is no available groundtruth for the mixing time. Lindau et al. [29] reported great variability of the estimated mixing times across the estimation

methods.

Fig. 3.13 compare the results of mixing time estimation on simulated RIRs between the estimator based on echo density and the estimator based on spatial incoherence. The spatial incoherence mixing time is systematically lower than the echo density one. The estimated mixing times vary by factors up to 5. The mean square error between both metrics for ray tracing simulations is of 0.05 s. There is a statistically significative linear correlation between the two estimation methods ($p = 0.002$). The mean squared error between mixing times estimated from the ray tracing RIRs and from the image source RIRs whenever the mixing time estimation succeeds is 0.0004 s. Cases of failure of the algorithm will be discussed later.

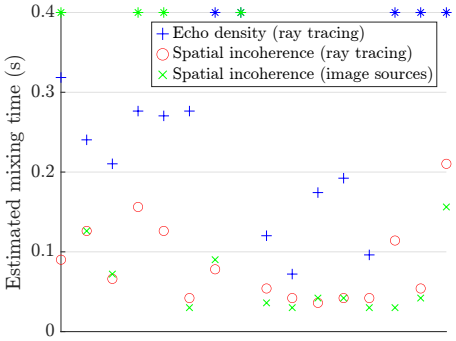


Figure 3.13: Results of mixing time estimation by echo density and by spatial incoherence on the ray tracing and image source simulations.

The parametric threshold, computed using the room surface area and formula 3.13, successfully detects the starting time of the spatial incoherence plateau for almost all acoustic scenes, as illustrated by Fig. 3.14.

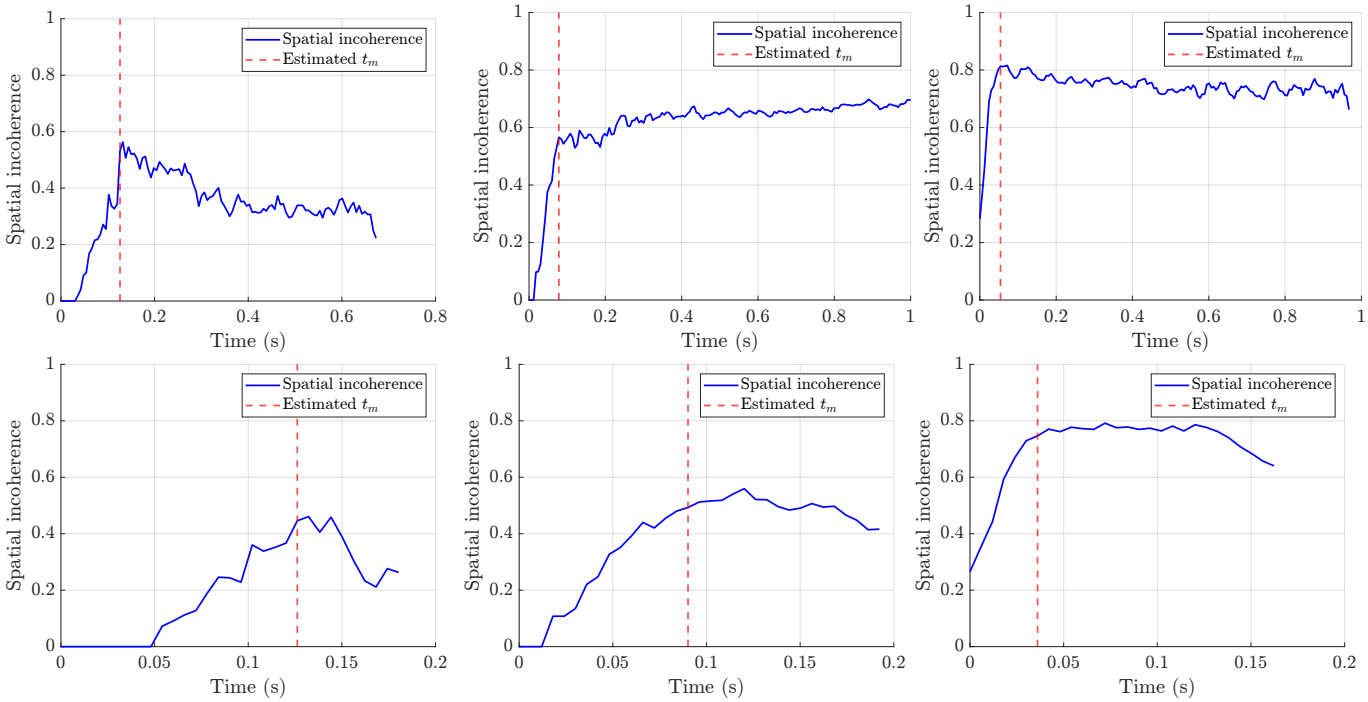


Figure 3.14: Examples of spatial incoherence profiles and estimated mixing times on ray tracing (top) and image source (bottom) simulations. From left to right: HalteresFurnished 1 1, Orsay 1 1, ShoeboxIsoRefl 1 1.

3.6 Limits and discussion

There remain differences in the spatial incoherence between ray tracing and ISM simulations. The ISM spatial incoherence is globally lower than the ray tracing spatial incoherence. The parametric threshold is not reached for some ISM RIRs (whereas it is always reached for the ray tracing RIRs). This is the case for configurations 6, 7, 9 and 19. For those configurations, the spatial incoherence profile of the image source simulation fails to converge to the ray tracing one, as illustrated by Fig. 3.15. The computation of higher orders of reflections might permit to exceed the spatial incoherence threshold.

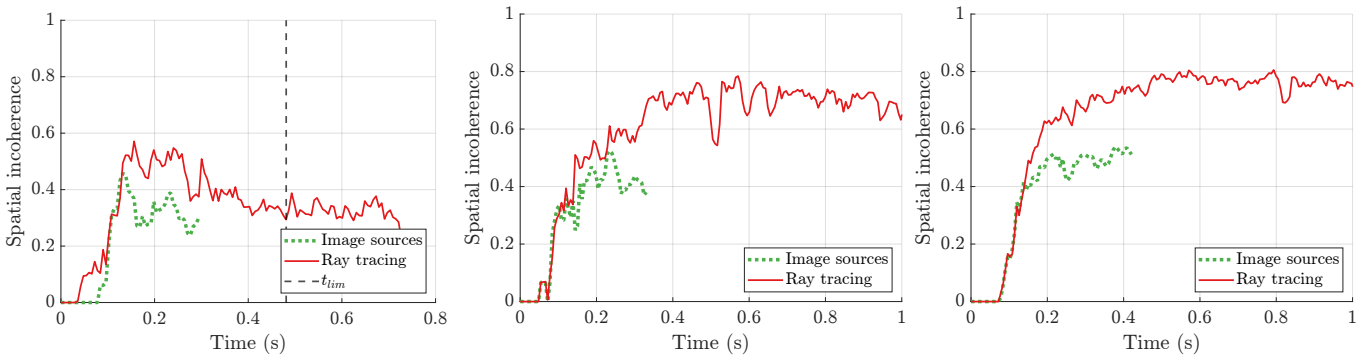


Figure 3.15: Examples of non-matching spatial incoherence profiles. From left to right: Halteres 1 1, Pleyel 1 1, and Vienne 1 1.

This might be explained by the absence of scattering in the ISM in its classical implementation. By considering only specular reflections, we obtain sparse simulations. Further work will include considering extensions of the ISM implementing surface scattering [57] and evaluating the effect of scattering on the diffuseness profiles of the simulated RIRs. The constraints of specular reflections and a non-uniform distribution of absorption may entail a lack of ergodicity, which might favor coherent trajectories for sound rays [55, 45]. For example, in our dataset, the cubic room simulations fails to reach a sufficient echo density and spatial incoherence threshold, as seen on Fig. 3.16. The rooms that show the greatest discrepancy in the spatial incoherence profile between image source and ray tracing simulations are models of concert halls: Halteres, Pleyel and Vienne, that might deviate from a uniform distribution of absorption.

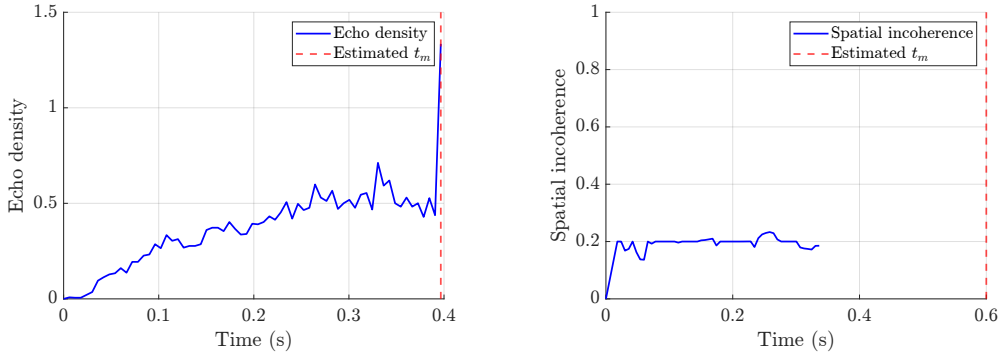


Figure 3.16: Echo density metric on the ray tracing simulation (left) and spatial incoherence metric on the ISM simulation (right) for the cubic room.

In this chapter, we addressed the problem of estimating the mixing exclusively on the early part of the RIR, more specifically the early reflections simulated with the ISM. We have put forward a method for estimating the mixing time based solely on the early reflections and on room parameters, making use of the spatial information contained in the Ambisonic encoding of the simulated RIRs. The results on the dataset of room models shows that the spatial information contained in the image source simulations, when analyzed with an adequate metric, enables to quantify the diffuseness growth even from the first few orders of reflections. We have shown that the spatial incoherence metric may be used to improve mixing time estimation on image source simulations, when room parameters enable to predict the spatial incoherence value. Future work will need to further explore methods for a robust prediction of the spatial incoherence value.

Chapter 4

Late reverberation synthesis and integration

In this chapter, we will build on the mixing time estimation method developed in Chapter 3 in order to propose a full algorithm to adjust the transition time between the early reflections and the late reverberation and their relative gains.

Following the duality between spatialized early reflections and late stochastic reverberation, the mixing time seems to be the ideal time at which to transition from the image source auralization to the artificial reverberator. In principle, after the mixing time, the reverberation tail may be fully described by its time-frequency envelope, which is independent of source and receiver positions. This time-frequency envelope is further constrained to an exponential decay, and thus the late reverberation is described by a reverberation gain and slope for each frequency band. The slope is generally expressed as a -60dB reverberation time, related to the exponential damping coefficient δ by the formula [23]:

$$RT(f) = \frac{3 \ln 10}{\delta(f)}. \tag{4.1}$$

Thus, implementations of artificial reverberators focus on simulating a stochastic process whose energy decay time may be precisely controlled for each frequency band. We will first describe two different implementations of artificial reverberation, before discussing the late reverberation delay and gain adjustment based on the mixing time estimation process described in Chapter 3.

4.1 Static late reverberation synthesis: Gaussian white noise with exponentially decaying envelope

Moorer [36] first noted the perceptual ressemblance between a RIR and Gaussian white noise modulated by an exponentially decaying envelope. Convolving an anechoic signal with such a simulated RIR leads to a natural sounding reverberation effect. In a room, frequencies decay at different rates, with higher frequencies usually having a faster decay rate. He thus proposes a simple method for synthesizing a reverberation tail with frequence-dependent reverberation time:

- first, a sequence of Gaussian white noise is genereated;
- then, it goes through a filter bank. The filter bank consists in 3rd order Butterworth filters in octave bands at frequency centers 125, 250, 500, 1000, 2000, 4000, and 8000 Hz [1];
- each bandpassed signal is modulated by an exponential curve with the adequate damping factor computed according to the 60dB reverberation time (Eq. 4.1);
- the bands are summed together to produce the synthetic reverberation tail.

This produces a monophonic reverberation signal. An Ambisonic-encoded late reverberation tail may be synthesized by repeating this process for each channel, generating a new Gaussian signal each time. Massé et al. [33] showed that synthesizing a Gaussian white noise per spherical harmonic component preserves the properties of the late diffuse sound field, i.e. incoherence and isotropy.

This method was implemented and used to synthesize the late reverberation in 3rd-order Ambisonic encoding. The auralization of anechoic signals then reduces to a convolution with the synthesized RIR. However, in real-time

contexts, convolutions might be too expensive, and artificial reverberators are used that enable cheaper auralization, in particular the Feedback Delay Network (FDN).

4.2 Dynamic late reverberation synthesis: Feedback Delay Networks

Originally proposed by Jot and Chaigne [22], FDNs consist of multiple interconnected delay lines, each possibly followed by attenuation filters, with their outputs mixed and fed back into the system through a feedback matrix. FDNs can achieve reverberation synthesis with far lower computational cost than direct convolution with room impulse responses, especially for long reverberation times and multi-channel scenarios [50]. While convolution (even with fast partitioned methods) requires $O(L \log L)$ operations per block for a RIR length L, an N-delay-line FDN can operate in $O(N)$ or $O(N \log N)$ per sample, independently of the reverberation time. For example, a dense orthogonal feedback matrix (e.g., random orthogonal) requires $O(N^2)$ multiplications per sample, while a Hadamard matrix can be implemented in $O(N \log N)$ using the Fast Hadamard Transform, greatly reducing cost for large N. These computational gains make FDNs scalable for Ambisonic-encoded or multi-source reverberation, where convolution would scale multiplicatively with the number of sources and outputs.

The FDN topology allows for precise control over the decay time in separate frequency bands. Delays and gains may be modulated in real-time for time-varying reverberation without recomputing the entire impulse response.

In the context of this preliminary study on late reverberation gain and delay adjustment methods, exponentially decaying Gaussian white noise was chosen rather than a FDN for the late reverberation synthesis. In fact, precise control over the FDN delay is difficult in practice, as FDNs have an inherent initialization delay during which the decay is not exponential. However, the developed algorithm will in time be integrated into a real-time auralization framework using a FDN, and further work will study on the specific implementation challenges this raises.

4.3 Reverberation time estimation

The per-frequency-band reverberations times given as input to the artificial reverberator were computed using the Eyring formula [15]:

$$T_{60}(f) = \frac{0.164\ V}{-S \times \ln(1 - \alpha(f))}, \tag{4.2}$$

with $\alpha(f)$ the frequency-dependent mean absorption coefficient (the mean absorption of the different materials pondered by the surface they occupy):

$$\alpha(f) = \frac{1}{\sum_i S_i} \sum_i \alpha_i(f) S_i, \tag{4.3}$$

where S_i are the different room surfaces and $\alpha_i(f)$ the corresponding frequency-dependent absorption coefficients.

This formula assumes ideally diffuse conditions as described in Chapter 3, i.e. isotropy and homogeneity of the sound field. We have seen that ideal diffusion is never met in practice. It is a known fact that this formula fails to predict accurately the reverberation time in many cases [55, 8]. However, developing a more accurate method for reverberation time estimation is left to future work and will be discussed briefly in Chapter 7.

4.4 Late reverberation delay adjustment

As hypothesized throughout this work, the mixing time seems to be the optimal time at which to set the late reverberation delay, based on theoretical and perceptual considerations. In fact, the mixing time is the time limit of the validity of the stochastic model, which reduces the late reverberation tail to its exponentially decaying time-frequency envelope. The mixing time is estimated based on the spatial incoherence profile, meaning that it is the time from which enough spatial incoherence is reached to forget the spatial information contained in the early reflections. For all these reasons, the mixing time estimated on the image sources as described in Chapter 3 was chosen as the starting time of the late reverberation tail.

To handle cases where the spatial incoherence does not reach the required threshold within a reasonable time, a maximum mixing time of 0.2 seconds was set. This value was chosen because it is longer than any estimated mixing time observed in the dataset. If the threshold has not been reached by 0.2 seconds, the system automatically starts late reverberation. This ensures a consistent transition when the mixing time estimation method described in Chapter 3 fails.

4.5 Late reverberation gain adjustment

The main theoretical concern when computing the late reverberation gain is to maintain the smoothness of the exponential decay at each frequency band. Solutions have been proposed in the literature to satisfy that constraint. The solution proposed by Gardner [17] is to adjust the gain of the artificial reverberator so that the linear decay slope retrospectively reaches the level of the direct sound. This solution would be valid if the early decay time was always equal to the late reverberation time. In practice, the early reflections tend to follow a steeper decay slope than the late reverberation, even in shoebox rooms. Thus, this method tends to overestimate the late reverberation gain and hinders the continuity of the decay. Another solution proposed by Li and Feng [28] is to set the gain of the artificial reverberator as the mean gain of the auralized image sources in a given window around the mixing time. Since even at the mixing time, the image source auralization remains sparse, this usually results in a very low late reverberation gain because of the many zero values in the image source auralization. Both methods are illustrated on Fig. 4.1.

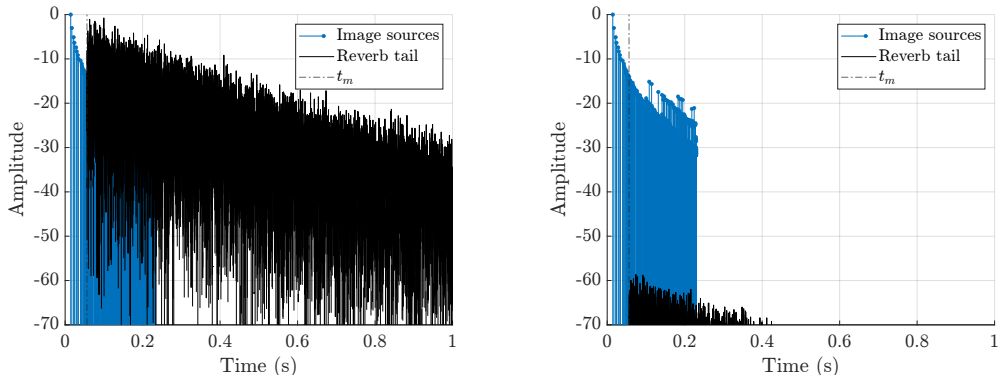


Figure 4.1: Methods for late reverberation gain adjustment proposed by Gardner [17] (left) and Li and Feng [28] (right) applied on ShoeboxLargeIso.

Building on the limits of these methods, the method implemented consisted in setting the gain as the mean amplitude of the non-zero values of the image source auralization in a 5 ms window around the mixing time. The zero values were excluded in order to maintain the smoothness of the decay. This was done for each frequency band to obtain a frequency-dependent gain factor. To account for cases when the mixing time estimation failed and the late reverberation started after the end of the image source auralization, the gain was set to half of the direct sound amplitude in those cases.

This gave visually satisfactory results, where the straight decay line of the dB-scale amplitude was preserved, as can by seen on Fig. 4.2. This gain adjustment method remains to be thoroughly evaluated against other methods by means of a perceptual test focusing on the perception of reverberation loudness.

To conclude, the late reverberation delay and gain adjustment algorithm was directly deduced from the mixing time estimation algorithm developed in Section 3. The delay was set as the estimated mixing time and the gain was set to guarantee the continuity of the decay envelope around the mixing time. We thus obtain a complete delay and gain adjustment algorithm that is compatible with a real-time implementation. We will now turn to the evaluation of the results, first through objective metrics and then through a perceptual test.

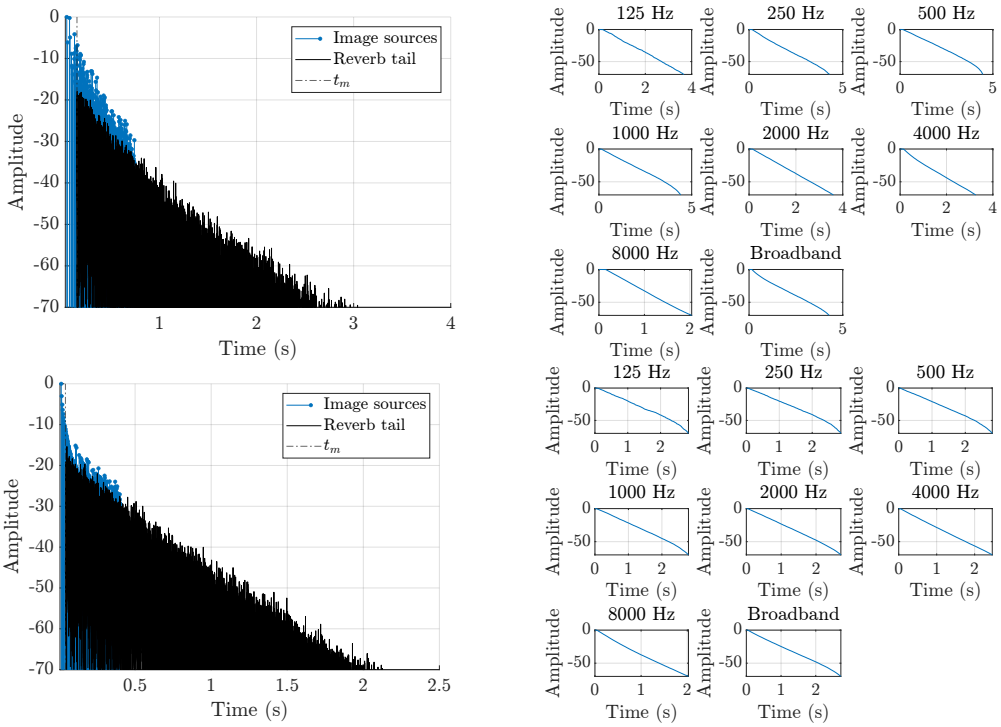


Figure 4.2: Examples of late reverberation delay and gain adjustment on Coupled 1 1 (top) and ShoeboxLarge 1 1 (bottom), and the resulting Reverse Integration Curves per frequency band.

Chapter 5

Objective evaluation

In this chapter, we propose and apply metrics for the objective evaluation of the late reverberation adjustment algorithm. In the first part, we will check that the energy decay and main acoustic parameters are close to that of the ray tracing reference. In the second part, we will evaluate the computational performance of the algorithm and in particular study the question of the number of image source reflection orders that must be computed for the algorithm to succeed. Last, we will turn to an exploratory study of the diffuseness properties of the resulting RIRs, by applying metrics developed in Chapter 3.

5.1 Energy decay properties and acoustic parameters

In this section, we will present some of the main acoustic parameters and estimate them on the CATT-Acoustic ray tracing RIRs as well as on the RIRs obtained by adjusting the late reverberation to the image sources through the algorithm described in Chapter 4.

One of the main concerns in the late reverberation delay and gain adjustment is to preserve the continuity of the energy decay, which can be verified by looking at the Energy Decay Curve (EDC), or Reverse Integration Curve (RIC). This curve, first introduced by Schroeder [54], is given by the formula:

$$EDC_h(t) = \int_t^{+\infty} h^2(\tau) d\tau, \tag{5.1}$$

where $h(t)$ is the RIR. In other terms, $EDC_h(t)$ is the remaining energy in the RIR after time t. The backward integration serves as a smoothing technique on the envelope of the response [23]. In theory, the room response should follow an exponential decay and thus the EDC should follow a straigth line in dB scale, at least after the mixing time. Figure 5.1 gives examples of configurations where the EDC of the IS+FDN RIR is very smooth and closely matches that of the ray tracing RIR. These examples illustrate the fact that the EDCs closely match whenever the ray tracing EDC is nearly a straight line in dB scale.

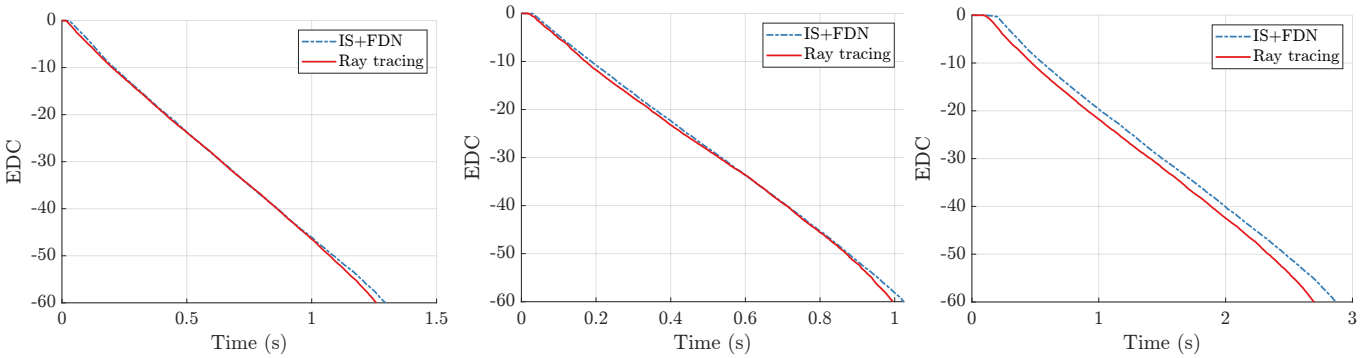


Figure 5.1: Comparisons of Energy Decay Curves of the ray tracing RIR and of the IS+FDN RIR on configurations with exponential decays (from left to right: ShoeboxIso 1 1, Morgan 1 1, and Vienne 1 1).

In cases when the decay is not perfectly exponential, the IS+FDN RIR EDC does not exactly match the ray tracing EDC. This is due to the artificial reverberation synthesis method, which constrains the late reverberation to have an

exponential envelope. The IS+FDN RIR EDC tends to be smoother than the ray tracing EDC. Examples are visible on Fig. 5.2. Possible methods for enabling double decays or other nonexponential late energy decays will be discussed in Chapter 7.

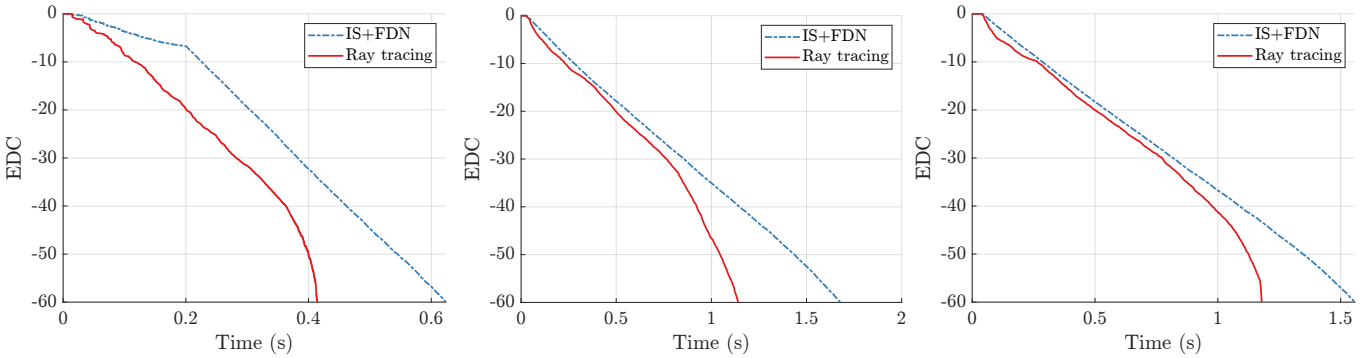


Figure 5.2: Comparisons of Early Decay Curves of the ray tracing RIR and of the IS+FDN RIR on configurations with nonexponential decays (Cube 1 1, Snail 1 1, and Snail 1 2).

There are different acoustic parameters describing the reverberation time. They are all computed from a linear fit on the dB-scale EDC and extrapolated to a -60 dB reverberation time, i.e. the time it would take for the energy level to decay by 60 dB after a Dirac excitation. However, they differ in the part of the EDC on which the linear fit is performed. EDT is the reverberation time extrapolated from a linear fit between the arrival of the direct sound and the -10 dB point, i.e. the time when the energy has decayed by 10 dB compared to the initial energy, while RT_{30} is the reverberation time extrapolated from a linear fit between the -5 dB point and the -35 dB point [21].

Table 5.1 shows the errors of RT_{30}, EDT and c_{80} between ray tracing RIR and IS+FDN RIR for each configuration, compared to the JND. RT_{30} is almost always smaller than one JND, which is not surprising since the ray tracing reverberation times are given as input to the artificial reverberator. EDT error can go up to 12 JNDs and c_{80} to 15 JNDs.

Room	Rec.	Src.	RT_{30}	EDT	c_{80}
JND			5%	5%	1
Amst	1	1	3.1	17.3	14.8
Coupled	1	1	10.6	31.8	5.9
Cube	1	1	20.2	24.1	NaN
Fogg	1	1	8.1	48.5	7.6
HalteresFurnished	1	1	7.4	35.4	5.0
HalteresFurnished	1	2	14.3	57.6	9.3
Halteres	1	1	33.0	47.1	5.9
Halteres	1	2	19.7	44.1	7.2
Morgan	1	1	3.3	17.2	3.1
Orsay	1	1	2.9	7.1	3.4
Pleyel	1	1	2.3	14.5	6.6
ShoeboxIsoRefl	1	1	1.2	7.0	1.9
ShoeboxIso	1	1	1.6	3.1	0.8
ShoeboxLargeIso	1	1	3.3	7.7	3.3
ShoeboxLarge	1	1	4.2	13.7	3.3
Shoebox	1	1	2.3	5.7	0.8
Snail	1	1	3.3	41.9	8.2
Snail	1	2	2.6	44.7	NaN
Snail	1	3	4.2	25.0	NaN
Snail	1	4	3.1	45.0	4.8
Vienne	1	1	5.0	19.5	5.6

Table 5.1: Acoustic parameter errors between the ray tracing simulation and the IS+FDN simulation.

On simple rooms such as shoebox rooms, acoustic parameters are usually quite well preserved in each frequency band, as shown on Fig. 5.3. This is not the case (except for RT_{30}) on more complex rooms, as illustrated by Fig. 5.4. However, we can see that the general relative trend across frequency bands is preserved.

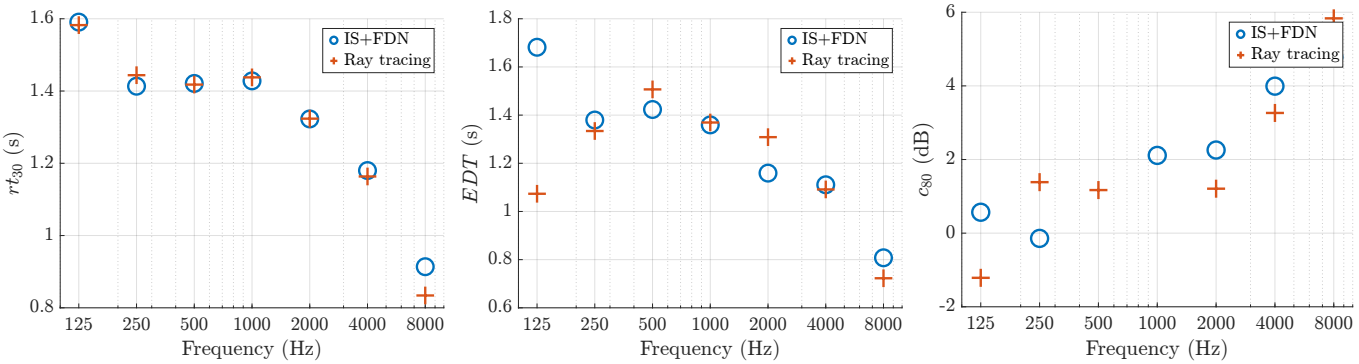


Figure 5.3: rt_{30}, EDT and c_{80} per frequency band on the ShoeboxIso 1 1 configuration.

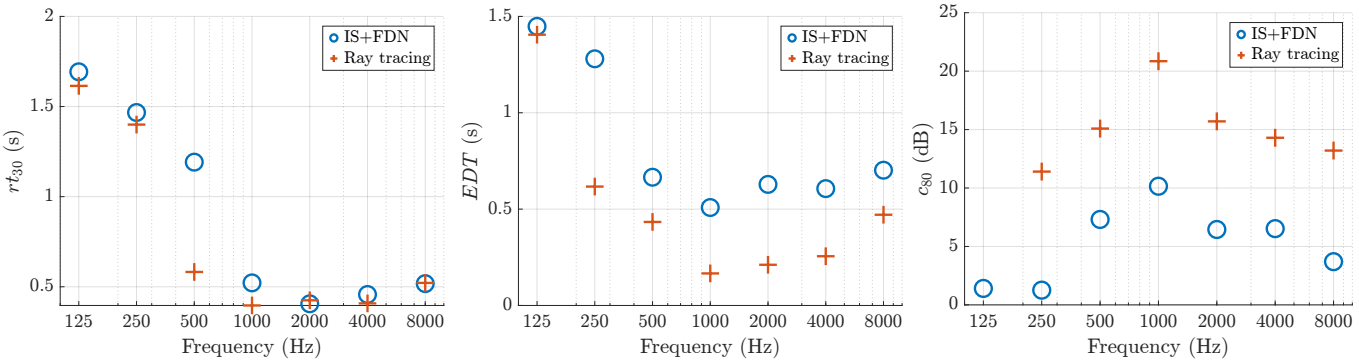


Figure 5.4: rt_{30}, EDT and c_{80} per frequency band on the HalteresFurnished 1 2 configuration.

5.2 Performance evaluation

In this section, we turn to the evaluation of the computational complexity of the algorithm proposed in Chapter 4, in order to test its compatibility with a real-time implementation. In particular, we would like to know how many orders of image source reflections must be computed before reaching the mixing time. Table 5.2 shows the estimated mixing times of each configuration and the corresponding number of reflections that must be computed so that the image source auralization lasts until at least the mixing time. This minimum order is equal to the maximum order of reflection encountered in the set of image sources that arrive before the mixing time. We also give the running time to compute reflections up to that order on a personal computer with 16 GB of RAM and an Apple M1 chip. Some straightforward optimizations could be implemented to reduce the running time. Specifically, the image source implementation used computes all image sources up to the given order; it could be modified to compute all the image sources up to the mixing time, and this would lead to lower computation times. The times indicated also include read/write operations that are unnecessary in a real-time integration.

The expected running time is less than one second for more than half of the configuration and less than ten seconds for three-quarters of the configurations. For these configurations, adequate optimizations are expected to enable real-time compatibility. However, there remain some configurations where the expected duration makes this method out-of-reach for real-time purposes. These are mostly large rooms with complex geometries. Methods must be developed in those cases to enable real-time auralization. For instance, higher image source reflections orders may be computed progressively whenever there is no change in room geometry or source location, as implemented by Poirier-Quinot et al. [43]. Ad-hoc solutions to mask the gap between uncalculated image sources and the FDN include adding partially spatialized clusters [10].

Room	Rec.	Src.	t_m	Orders required	Expected duration
Amst	1	1	0.20	7	**0'44**
Coupled	1	1	0.15	4	649 ms
Cube	1	1	0.20	12	**0'05**
Fogg	1	1	0.20	Inf	**Inf**
HalteresFurnished	1	1	0.13	10	0'01
HalteresFurnished	1	2	0.07	5	0'01
Halteres	1	1	0.20	14	**28'40**
Halteres	1	2	0.20	13	**11'46**
Morgan	1	1	0.03	1	735 ms
Orsay	1	1	0.09	7	**0'36**
Pleyel	1	1	0.20	10	**57'05**
ShoeboxIsoRefl	1	1	0.04	4	991 ms
ShoeboxIso	1	1	0.03	3	420 ms
ShoeboxLargeIso	1	1	0.04	3	906 ms
ShoeboxLarge	1	1	0.04	3	339 ms
Shoebox	1	1	0.03	3	350 ms
Snail	1	1	0.03	1	53 ms
Snail	1	2	0.04	1	54 ms
Snail	1	3	0.16	8	**0'05**
Snail	1	4	0.18	8	**0'07**
Vienne	1	1	0.20	10	**0'07**

Table 5.2: Estimated mixing time, corresponding minimum number of orders of reflections to compute and estimated computation time on a standard computer for that reflection order. Times strictly over one second are highlighted in bold.

5.3 Diffuseness properties of the simulated RIRs

We now apply the echo density and spatial incoherence metrics defined in Chapter 3 to the IS+FDN RIRs and compare the resulting diffuseness profiles with the profiles of the ray tracing RIRs and to profiles of measured RIRs. In particular, we would like to know if the diffuseness profiles ressemble those of measured RIRs. That would tell us something about the realism of the simulated RIRs.

In the case of the echo density measure, it is no surprise that this jumps directly to 1 at the mixing time. In fact, the echo density measure is normalized such that Gaussian white noise has an echo density of exactly 1. This ressembles the late reverberation field of measured RIRs and of ray tracing RIRs. The echo density profile of the IS+FDN RIR is closer to the ray tracing one than that of the image source RIR, thus the late reverberation corrects the lack of echo density inherent to the image sources. However, echo density tends to increase too fast in the IS+FDN RIRs compared to the ray tracing RIRs.

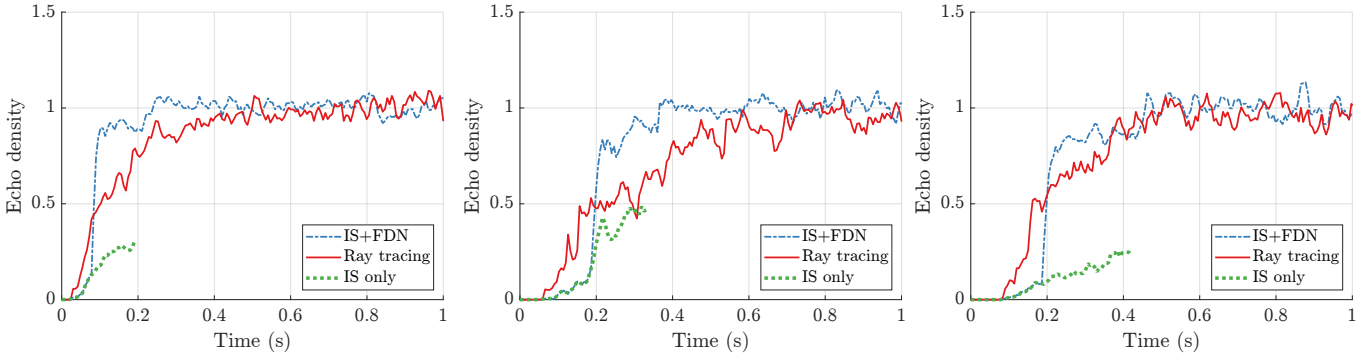


Figure 5.5: Comparisons of echo density profiles of the IS+FDN RIR, the ray tracing RIR and the image source RIR for the same acoustic scenes (Orsay 1 1, Pleyel 1 1, Vienne 1 1).

In the case of the spatial incoherence profile, which shows very different profiles across different measured RIRs and ray tracing simulated RIRs, we then again obtain a predictible diffuseness plateau for the late reverberation. The spatial incoherence value of the synthetic Gaussian reverberation is not of 1 like the echo density value, but around 0.9. In fact, by definition, the spatial incoherence value is constained to not exceed 1. In theory, Gaussian white noise should have a diagonal covariance profile and a uniform distribution of eigenvalues. Pure uncorrelation is not

obtained in practice and the spatial incoherence value of Gaussian white noise goes up to 0.95, with higher values for smaller HOA orders, as can be seen on Fig. 5.6 (left). The spatial incoherence value is further lowered by the exponential envelope which might cause some covariance in the different spherical harmonic components. This effect is more pronounced for lower HOA orders, as illustrated on Fig. 5.6 (center). A non-uniform exponential decay over frequency bands leads to a drastic decrease of spatial incoherence over time, as can be seen on Fig. 5.6 (right). A non-uniform exponential decay causes some covariance in the SH signals; this phenomenon should be further studied from a theoretical point of view and experimentally.

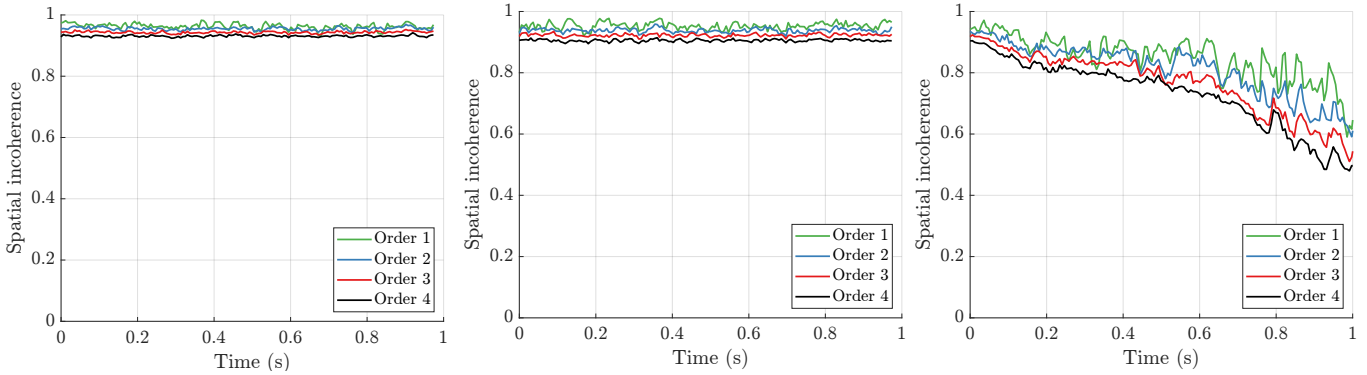


Figure 5.6: Comparisons of spatial incoherence profiles of Gaussian white noise with no exponential decay (left), Gaussian white noise with uniform exponential decay at all frequencies (center), and Gaussian white noise with faster decay at high frequencies (right), for different HOA orders.

In some cases, the spatial incoherence value jumps directly from a low value to 0.9 at the mixing time. In other cases, it has a gradual increase that more closely matches the gradual increase which may be observed in measured RIRs or ray tracing RIRs. However, the value of the spatial incoherence plateau is systematically higher for the IS+FDN RIR than for the ray tracing RIR, which never goes as high as 0.9. One possible explanation for this is that the late reverberation of the ray tracing RIRs deviates from the assumptions of an ideally diffuse sound field, i.e. homogeneity and isotropy. In particular, in some rooms, the late reverberation field might never reach isotropy. Late reverberation synthesized with Gaussian white noise is by nature isotropic and homogeneous [13, 32]. This is not necessarily the case for other artificial reverberators, in particular for FDNs. The possible anisotropy of the late reverberation field and its perceptual implications will be further discussed in Chapter 7.

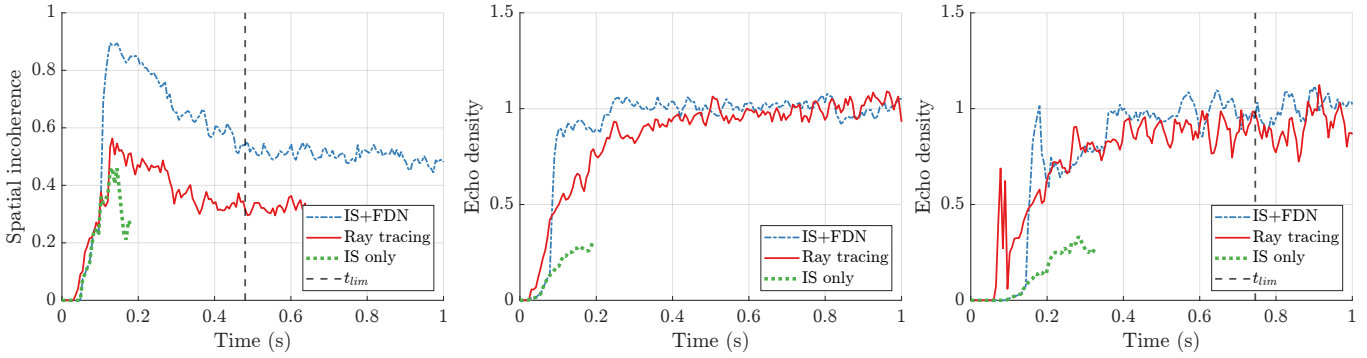


Figure 5.7: Comparisons of spatial incoherence profiles of the IS+FDN RIR, the ray tracing RIR and the image source RIR for the same acoustic scenes (HalteresFurnished 1 1, Orsay 1 1, Snail 1 3).

The objective evaluation of the algorithm shows mixed results, where the acoustic parameters are preserved for only part of the acoustic scenes and the real-time compatibility is not guaranteed for all rooms. In the context of AR/VR applications, however, a perceptual evaluation might be more relevant in order to assess the possible limitations of the algorithm in terms of immersion.

Chapter 6

Perceptual evaluation

This chapter presents the design and results of a perceptual test aimed at evaluating the proposed reverberation gain and delay adjustment algorithm. The algorithm and resulting auralizations will be evaluted with a focus on localization, authenticity, and immersion.

6.1 Test objective and expected results

The goal is to provide a perceptual evaluation of the new late reverberation delay and gain adjustment scheme. The perceptual evaluation is particularly relevant because the goal of the framework is to produce perceptually plausible auralization that enables immersion into Virtual Acoustic Environments. Whereas objective measures might detect differences between distinct auralization schemes, these might be undistinguishable from a perceptual point of view. The design of a formal perceptual test with adequate statistical analysis aims to assess and quantify the possibility of perceptual discrimination between different systems under test and an ideal reference.

This first perceptual evaluation will be limited to a static context, i.e. there will be no dynamic update of the auralization based on listener or source movement and no visual feedback.

As discussed in Chapter 2, in the absence of calibrated geometric models for which measured RIRs could be compared to simulated RIRs, the reference that was chosen for comparison with the different auralization schemes is the CATT-Acoustic ray tracing simulation.

The different auralization conditions that are put to test are:

- CATT: the reference auralization, i.e. the RIRs simulated using the CATT-Acoustic ray tracing software;
- Images: image source auralization only, no late reverberation;
- Gaussian: late artificial reverberation only, i.e. Gaussian white noise modulated by an exponential envelope;
- Basic: image sources and late reverberation with a basic transition scheme: the transition time is chosen to be $\sqrt{V}$ ms and the FDN gain is half of the direct sound gain, as presented in Chapter 4;
- Diffuseness: image sources and late reverberation with the new transition scheme: the transition time is computed based on the spatial incoherence profile as described in Chapter 3, and the gain is computed based on the image source gain around that transition time, as explained in Chapter 4;
- CattGaussian: the CATT RIR with the late reverberation tail replaced by exponentially decaying white noise starting at the mixing time. Evaluating the proximity of these RIRs to the original RIRs is a way of evaluating the validity of the mixing time estimated with the method discussed in the previous chapter;
- Anchor: a lowpass-filtered version of the reference signal.

For all of these schemes, the per-frequency band reverberation times of the late reverberation are aligned to those of the CATT-Acoustic reference estimated by exponential decay curve analysis. Participants were asked to focus on authenticy, plausibility, and localization.

6.2 Test protocol

6.2.1 MUSHRA test and web implementation

The test was designed according to the Recommendation ITU-R BS.1534-3 Method for the subjective assessment of intermediate quality level of audio systems [59], also known as "MUlti Stimulus test with Hidden Reference and Anchor" (MUSHRA), designed to evaluate medium and large impairments of audio systems. The MUSHRA test

presents, for each trial, a high quality reference signal, that must be compared to multiple conditions:

- a certain number of systems under test, expected to introduce impairments to the reference signal;
- a hidden reference signal;
- one or two anchor signals, which are lowpass-filtered versions of the reference signal and should be judged as most different from the reference signal. In our test, we generated a single anchor with a lowpass filter cutoff of 3.5 kHz.

The assessors are asked to grade to ressemblance of each signal to the high quality reference signal. In our case, the reference signal is the CATT condition, i.e. the RIRs generated using the CATT-Acoustic ray tracing software.

The test was implemented using webMUSHRA, an open-source web implementation based on the WebAudio API [52]. The test interface may be visited here[1] and a screenshot is visible on Fig. 6.1. The audio signals under trial were binaurally rendered to be played over headphones, such that anyone could participate remotely with their personal headset. While experimental conditions can not be precisely controlled, it has been shown that this does not significantly impact the results [51]. Due to the flexibility of a web support, the recruitment of participants is facilitated and no experimental setup is required. In addition, binaural rendering over headphones is the most frequent rendering setup for our applications, and thus it is logical that this should be tested first. In future works, evaluation should be performed in a loudspeaker array setup in order to assess the influence of higher spatial resolution.

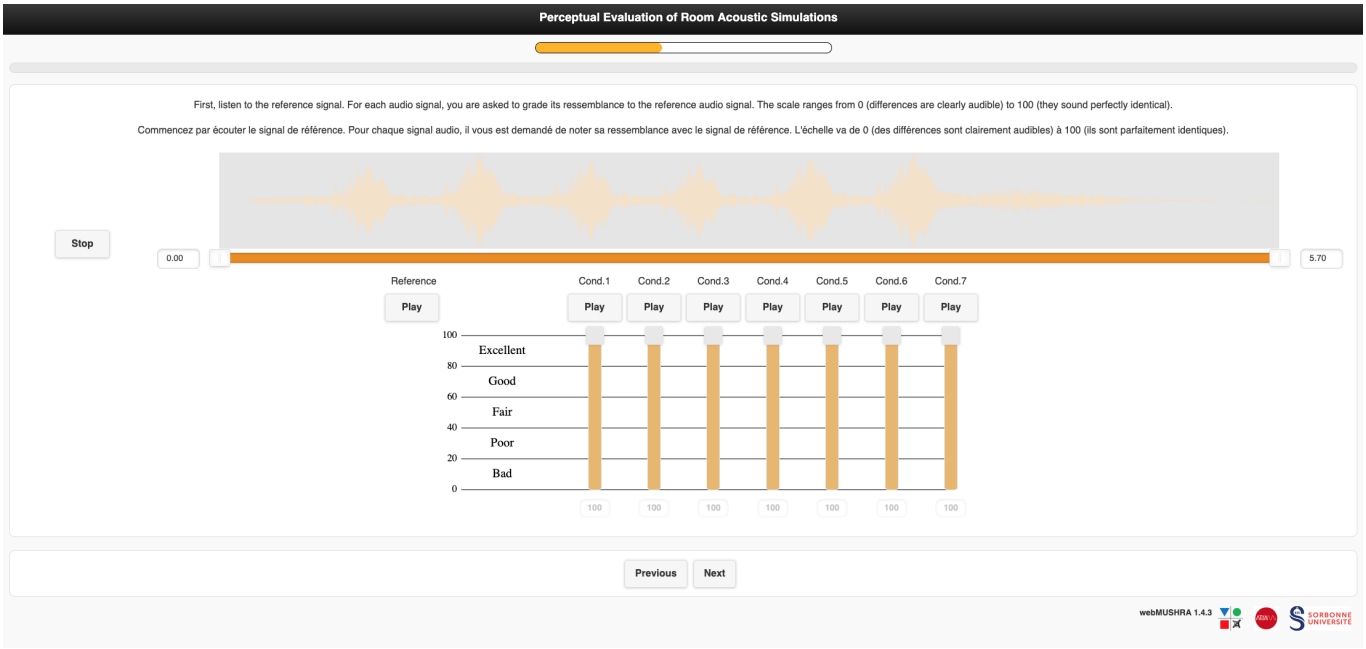


Figure 6.1: Screenshot of a trial page of the online test, using the webMUSHRA framework.

6.2.2 Subjects, number of trials, duration

The maximum test duration was set to 30 minutes in order to avoid tiredness of the participants. All audio stimuli had a duration of 6 seconds. Four different room acoustic conditions were selected, and two different source stimuli, for a total of 8 trials. As such the test was expected to last approximately 24 minutes (3 minutes per trial) excluding the training phase. Each participant was asked to rate all 7 conditions in all 8 trials, resulting in 56 measures per participant in a fully-repeated measures design. Power analysis for a repeated-measures ANOVA with 7 conditions across 8 trials (moderate effect size $f = 0.25$, type I error rate $\alpha = 0.05$, correlation between measures $\rho \approx 0.5$) indicated that approximately 18 participants provide 80% power to detect statistically significant differences between conditions while accounting for inter-listener variability. Therefore, a minimum of 18 participants with normal hearing and experience in listening tests and/or spatial audio was sought for statistical significance. Subjects were recruited within academic research groups specialized in room acoustics, spatial audio, digital signal processing, and musical acoustics.

The training phase was designed according to the recommendation [59]. It was aimed at making the participants familiar with the variety of audio stimuli, the web interface and the grading scale. In the first training phase, the participants could freely listen to some of the audio excerpts that they would be asked to assess. In the second training

[1] `https://experiment.dalembert.upmc.fr/webmushra/`

phase, they were asked to perform a MUSHRA trial whose results were not taken into account in the analysis. The participants were not informed of the presence of the hidden reference signal or of the anchor.

6.2.3 Anechoic source stimuli

Two anechoic excerpts were chosen as source stimuli, each 4.5 seconds long. The first one is a percussive sequence. The second one is a female speech sequence. Both are critical signals for the assessment of reverberation. Percussive signals have sharp transients and clear onsets, which make them highly sensitive to early reflections and changes in temporal smearing, allowing listeners to perceive differences in reverberation duration and clarity. Speech signals are critical because they contain sustained harmonic content, formant structures, and intelligibility cues, which are affected by reverberation in a way that is perceptually relevant for everyday listening [16].

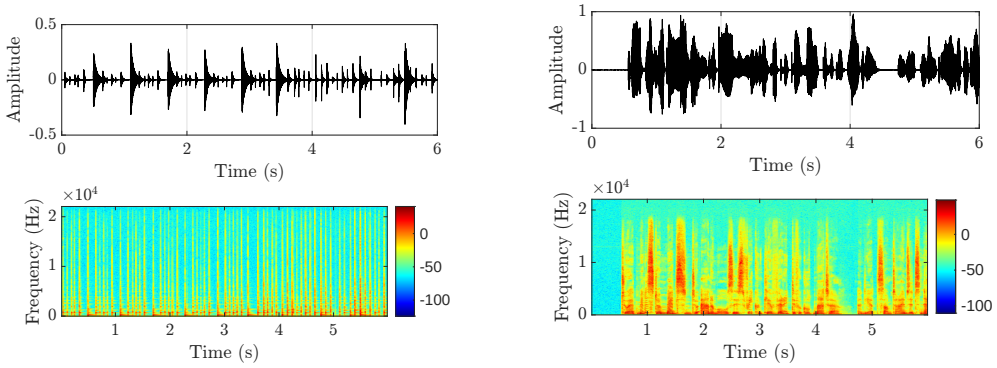


Figure 6.2: Anechoic source stimuli and their STFTs: percussive sounds (left) and female voice (right).

6.2.4 Room acoustic conditions

Four room acoustic conditions were chosen with the aim of systematic variation of some parameters: room geometry, room volume, room surface, source-to-receiver distance, mean absorption, reverberation time, global spatial incoherence, estimated mixing time. The room geometries include coupled volumes with furniture (HalteresFurnished), a simple shoebox room (ShoeboxIsoRefl), a complex succession of rooms (Snail) and a concert hall (Vienne), each visible on Figure 6.3. The parameters of the configurations chosen are presented on Table 6.1.

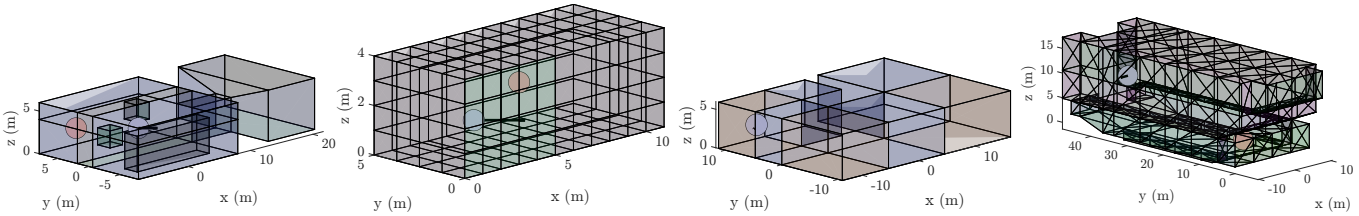


Figure 6.3: The four room models selected for the test.

Room	Rec.	Src.	Volume (m^3)	Surface (m^2)	S-R dist. (m)	Abs. (%)	RT (s)	S.I.	t_m
HalteresFurnished	1	2	2114	1635	7.1	43.2	0.76	0.34	0.07
ShoeboxIsoRefl	1	1	220	238	4.9	14.2	0.94	0.69	0.05
Snail	1	4	3573	2038	20.1	21.4	1.78	0.48	0.20
Vienne	1	1	17095	5641	32.8	15.1	2.57	0.59	0.28

Table 6.1: Parameters for the chosen acoustic scenes. The reverberation time, spatial incoherence and t_m are estimated on the CATT reference signal.

6.2.5 Binaural rendering

The Ambisonic RIRs were decoded to binaural RIRs (BRIRs) using the virtual loudspeaker approach and a set of publicly available HRTFs [25]. The virtual loudspeaker approach decodes a Higher-Order Ambisonics (HOA) signal to

a set of 20 virtual loudspeakers positioned evenly on a sphere using the golden-angle method. Each virtual loudspeaker signal is convolved with the HRTF corresponding to its location to create binaural signals, which are then summed to produce the final stereo output. Convolution of the left BRIR with the source signal was performed to obtain the left-ear channel of the binaural auralization, and likewise for the right BRIR.

6.3 Data analysis

6.3.1 Subject post-screening

At the end of the test, participants were asked their gender, if they self-reported a hearing loss, and if they had expertise in listening tests and in spatial audio. One participant reported a hearing loss and was excluded from the results analysis.

The hidden reference was exploited to further exclude some participants. As stated in the recommendation [59], an assessor should be excluded from the results if he or she rates the hidden reference condition lower than a score of 90 for more than 15% of the trials, i.e. for at least two trials in our case. This rule led to the exclusion of two participants. This resulted in a total of 23 participants, with a repartition reported on Table 6.4.

	Total	23
Gender	female	8
	male	15
Hearing Loss	yes	0
	no	23
Listening tests	3+times	10
	1-2 times	7
	never	6
Expert spatial audio	yes	8
	no	15

Figure 6.4: Summary of participants after post-screening.

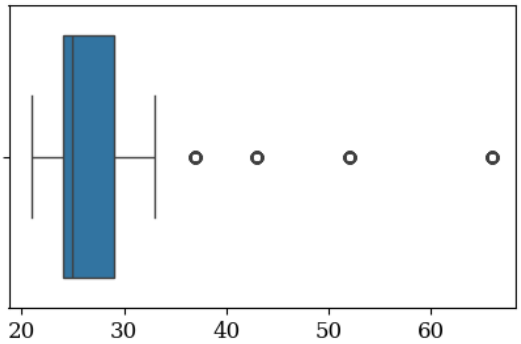


Figure 6.5: Age distribution after post-screening.

6.3.2 Statistical data analysis

Following the recommendations [59], a Linear Mixed Effects Model was applied on the data, treating the participant as a random effect. Each participant gave multiple ratings (one per trial and per condition) and we are not interested in individual differences in rating systems, which are thus treated as random effects. The response was the rating score and the terms were the different factors (age, gender, room, source stimulus, condition) and their interactions.

The non-significant interactions ($p > 0.05$) were discarded and the significant effects were further analyzed with post-hoc pairwise Tukey tests.

6.4 Results

First, we provide some visualization of the results, which will give us intuition on the performance of the different conditions in the different trials, before turning to statistical analysis with the aim of showing significant effects. Fig. 6.6 shows the distribution of ratings given to each condition, without separating by acoustic scene or source stimulus. The boxplots show the median and quartiles 1 and 3. The whiskers show the value range of points that are not outliers, and outliers are displayed separately. Outliers are data points lying below $Q_1 - 1.5 \times IQR$ or above $Q_3 + 1.5 \times IQR$, where $IQR = Q_3 - Q_1$ and Q_1 and Q_3 are the first and third quartiles.

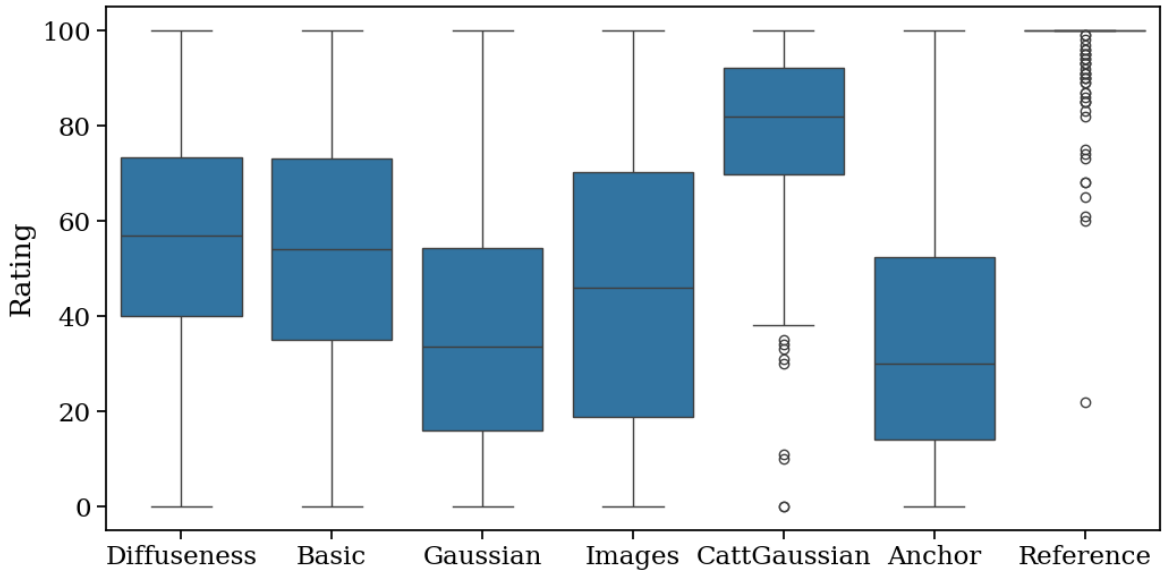


Figure 6.6: Distribution of the rating scores for each condition, over all acoustic scenes and source stimuli.

As we can see, the anchor does not serve its role as an anchor since its median rating is 30/100. The anchor was a lowpass-filtered version of the reference signal, and although it was perceptually quite far from the reference signal, participants were asked to focus on authenticy, plausibility, and localization, and were judging the perception of reverberation across the different conditions. The plot illustrates the absence of anchor, as no condition was uniformly poorly rated across all trials. In terms of reverberation, the anchor was often closer than other conditions. The question of finding a good anchor for that specific test will be further discussed in the results analysis.

Apart from the reference, the condition that stands out is the CattGaussian condition, with a median score of 82/100. This tends to validate that the estimated mixing time is generally not smaller than the real perceptual mixing time, since the reference and the CattGaussian condition seem perceptually close.

Fig. 6.7 further shows the distribution of grades for each condition across acoustic scenes and for both source stimuli. The plot displays the mean value and standard deviation of the ratings. We can notice several trends:

- The hidden reference has a high mean rating, with some fluctuations across the trials, which can be explained by confusion with the CattGaussian condition: whenever the CattGaussian condition is rated quite high (ShoeboxIsoRefl and Snail), the hidden reference is rated a bit less than 100.
- The average ratings for one condition and one acoustic scene are usually similar between both source stimuli, except for condition Gaussian where the mean rating of the voice is systematically higher than the mean rating of the bongos.
- Except for the hidden reference, all conditions perform very differently in the different acoustic scenes. In particular, the Images condition is very well rated in HalteresFurnished and poorly rated in the other rooms. The CattGaussian condition is on average well rated in ShoeboxIsoRefl, Snail, and Vienne and relatively poorly rated in HalteresFurnished.
- The CattGaussian is systematically on average the best rated after the hidden reference, except in HalteresFurnished.

To test for significant effects, the data was further analyzed via a Linear Mixed Effects Model. A type III ANOVA test was performed using Satterthwaite's method, treating the participant as a random effect. The results are summarized in Table 6.2. The factors tested for influence on the rating are:

- `expert_audio`: their experience in listening tests (3 groups)
- `expert_spatial`: whether they worked in spatial audio (2 groups)
- `room`: the room acoustic scene (4 groups)
- `stimulus`: the source stimulus (2 groups: voice or bongos)
- `condition`: the condition under test (7 groups)

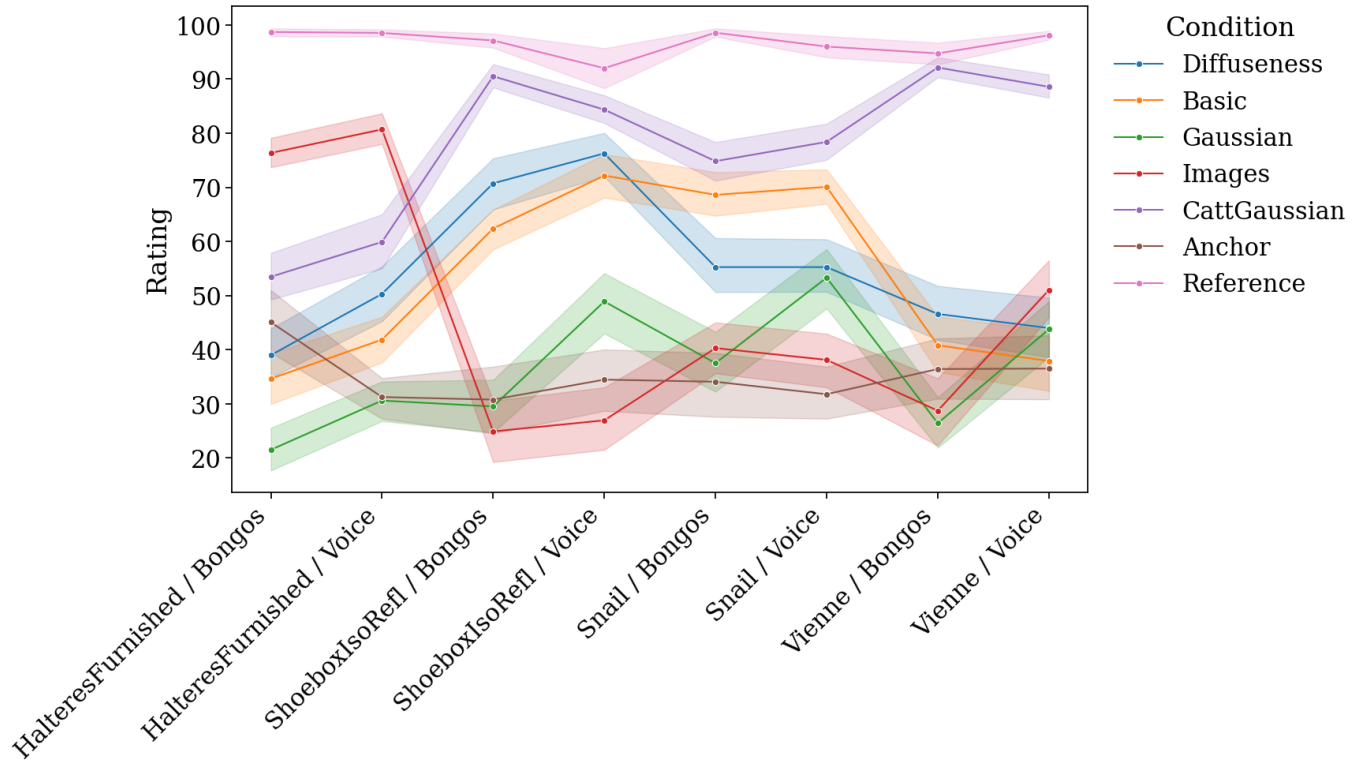


Figure 6.7: Distribution of the rating scores for each condition, for each acoustic scene and source stimulus. The highlighted zones indicate 68% confidence intervals.

Factor	F value	p-value
expert_audio	0.25	0.79
expert_spatial	0.99	0.35
room	9.7	$< \epsilon$
source_stimulus	14	$< \epsilon$
condition	3.1e2	$< \epsilon$
room:source_stimulus	0.41	0.75
room:condition	31	$< \epsilon$
source_stimulus:condition	5.8	$< \epsilon$
room:source_stimulus:condition	1.9	0.01

Table 6.2: Type III Analysis of Variance Table applied to the different factors to test their influence on the rating score. $\epsilon = 10^{-3}$. The semi-colons indicate an interaction effect.

Experience in listening tests and/or spatial audio had no significant influence on the rating scores. The interaction between the room and the source stimulus (voice or bongos) was also analyzed as non-significant.

The source stimulus (voice or bongos) had a significant influence on the rating score. The voice source stimulus was significantly better rated than the bongos ($F = 12.1, p = 0.0001$). The significant difference in rating scores between the voice and bongos stimuli can be attributed to the distinct perceptual characteristics of these sound types in reverberant environments. Speech signals allow for better adaptation and intelligibility in reverberant conditions. Conversely, percussive sounds, such as those produced by bongos, lack the sustained tonal components of speech, making them more susceptible to degradation in reverberant spaces [41].

The room also had a significant impact on the rating ($F = 9.7, p = 2.4e^{-6}$), as well as the condition under test ($F = 311.6, p < 2.2e^{-16}$). However, those correlations may not be analyzed separately because of the interaction effects. The main statistically significant interaction was the `room:condition` interaction ($p < 2.2e^{-16}$), so we further analyzed this interaction with post-hoc pairwise comparisons. In particular, we computed the estimated marginal means for each `room:condition` pair. The adjusted marginal means for each acoustic scene and each condition are depicted on Fig. 6.8. The whiskers represent the standard error around the marginal mean. The model being based on a balanced design (there is an equal number of observations per `room:condition` pair, since all participants evaluate all conditions in all trials), the standard errors are the same (equal to 3.61). Intuitively, if whiskers of two lines do not

overlap, then the difference between their respective ratings is statistically significative: one condition is significantly better than the one.

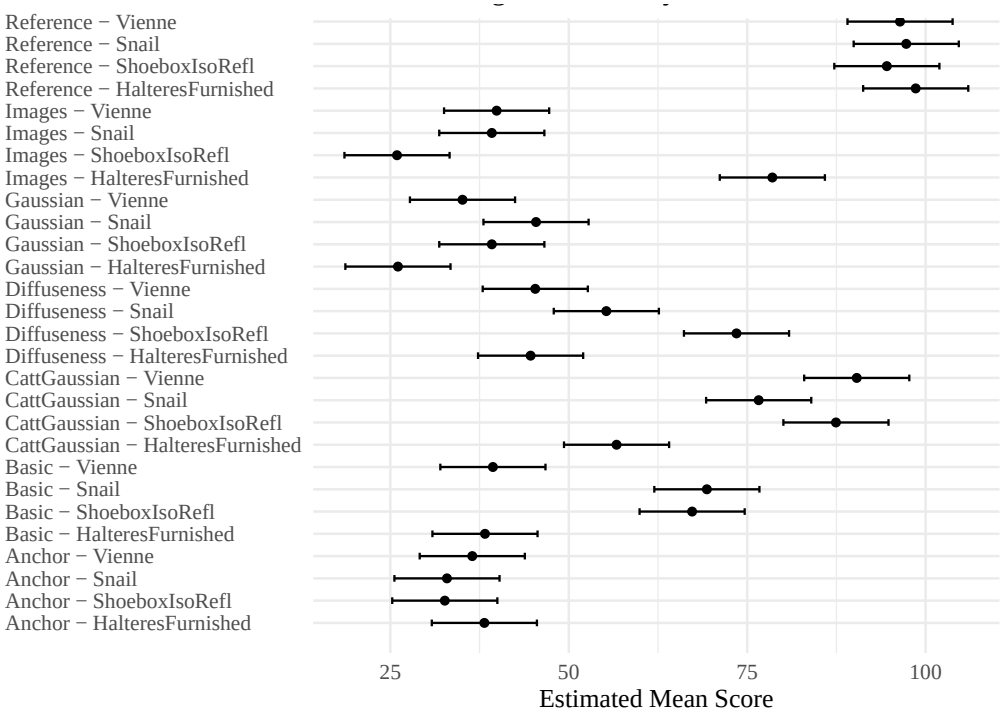


Figure 6.8: Estimated marginal means for each acoustic scene and each condition.

A thorough analysis of the significance of the difference between the mean scores for pairs of conditions in the same room acoustic scene enabled to draw some conclusions on the test results. For sanity check, it was verified that:

- in each acoustic scene, it was never the case that the anchor was rated significantly better than another condition;
- in each acoustic scene, it was never the case that the reference was rated significantly worse than another condition.

Now, we will first turn to the analysis of the significative pairs in the perspective of validating the mixing time estimation algorithm (condition CattGaussian), and then with the aim of validating the reverberation delay and gain adjustment algorithm (condition Diffuseness).

6.4.1 Partial validation of the mixing time estimation algorithm

The CattGaussian condition was used in order to partially validate the mixing time estimation algorithm presented in Chapter 3. In fact, if it is judged perceptually indistinguishable from the reference, then we know that the estimated mixing time is not smaller than the perceptual mixing time.

CattGaussian was not judged significantly worse than the other conditions (except for the hidden reference) except in one case, the Gaussian condition in HalteresFurnished. It was judged significantly worse than the reference in half of the acoustic scenes, i.e. in HalteresFurnished and Snail. In the other rooms, ShoeboxIsoRefl and Vienne, the difference was not significant. We can conclude that the estimated mixing time is not smaller than the perceptual mixing time in those rooms.

The possibility of perceptual discrimination does not refute the (physical) mixing time estimation algorithm. Perceptual differences may be caused by inaccurate estimation of the reverberation times, especially in rooms presenting double-decay slopes such as HalteresFurnished and Snail. Thus the late reverberation tails of Catt and CattGaussian may have audible differences in spectral coloration. Moreover, since mixing is not perfect, the late reverberation field will never be perfectly isotropic and homogeneous. Table 6.1 shows that the two rooms where perceptual discrimination is possible are also the rooms with the smallest spatial incoherence value, less than 0.5 for both rooms. This could be an indicator that the late reverberation fields are anisotropic, since the spatial incoherence analysis performed in the spherical harmonic domain interprets anisotropy as coherence [32]. An in-depth analysis of the properties of the late reverberation and their perceptual importance will be conducted in future work, for instance with a perceptual test focusing only on the late reverberation. This will be further discussed in Chapter 7.

6.4.2 Partial validation of the reverberation adjustment algorithm

The reverberation adjustment algorithm described in Chapter 4 corresponded to condition DIFFUSENESS, while BASIC was a baseline reverberation delay and gain adjustment algorithm where the delay what set to $\sqrt{V}$ milliseconds and the gain to half of the direct sound gain per frequency band. It was expected that DIFFUSENESS would be closer to the reference than IMAGES (only early reflections) and GAUSSIAN (only late reverberation). This is significantly the case in ShoeboxIsoRefl ($p \approx 5e^{-13}$). In HalteresFurnished, DIFFUSENESS is significantly better than GAUSSIAN ($p = 1e^{-4}$) but significantly worse than IMAGES ($p = 8e^{-13}$). In Snail, DIFFUSENESS is significantly better than IMAGES ($p = 3e^{-3}$) but not than GAUSSIAN. In Vienne, there is no significant difference between DIFFUSENESS and IMAGES or GAUSSIAN.

The results can be explained by the different reverberation times. For HalteresFurnished, with a very low reverberation time of 0.76 second, the IMAGES condition is very close to the CATT reference because it sounds dry and enables very precise localization, while the Gaussian reverberation of the DIFFUSENESS condition is too audible. In Snail, with a reverberation time of 1.78 second, the IMAGES condition is rated poorly because the lack of late reverberation is clearly audible.

The various reverberation times relate to the difficulty of defining an efficient anchor for this experiment. In fact, although IMAGES and GAUSSIAN can be expected to be poorly rated (the former because of a lack of authenticy, the latter because of a lack of localization), they are not evenly so in all rooms, because small reverberation times tend to favor the IMAGES condition while large reverberation times tend to favor the GAUSSIAN. An efficient anchor that would be systematically rated as the worst condition remains to be found.

The DIFFUSENESS condition is never rated significantly better than the baseline BASIC algorithm. It is in fact rated significantly worse in Snail. We can thus conclude that the proposed algorithm, although usually closer to the CATT reference than only early reflections or only late reverberation, does not bring a significant improvement compared to a baseline algorithm in terms of proximity to the ray tracing reference signal.

We can discuss the choice of the ray tracing simulations as reference. In fact, it is known that ray tracing simulations tend to miss some acoustic paths and thus underestimate the amount of the sound energy of the response [9]. This can explain that the late reverberation tails of the CATT conditions tend to have a low gain compared to the DIFFUSENESS, BASIC and CATTGAUSSIAN conditions. Thus we suppose that the DIFFUSENESS condition would show a better success if the reference signals were measured RIRs.

6.5 Dynamic perceptual test

The relevance of the comparison between ray tracing simulations and image source simulations with artificial reverberation can be discussed. In fact, in the context of Virtual and Augmented Reality applications, the auralization scheme must be evaluated in terms of localization, authenticity and immersivity. Therefore, its performance should be evaluated in a dynamic context.

In further works, the auralization scheme will be integrated in Max into an AR/VR framework giving the user the possibility to move inside an acoustic scene. A discrimination test may be designed to evaluate authenticity, where the subject is asked to discriminate between real and virtual sources. The new transition scheme is expected to improve the localization performances. In fact, an accurate mixing time estimation guarantees that the image sources containing the spatial information are heared, and adequately lowering the FDN gain enables to avoid some masking effects.

To conclude, the webMUSHRA test has provided a partial validation of the reverberation adjustment algorithm, which needs to be further confirmed by a dynamic perceptual test designed to assess the authenticity of the auralizations. The perceptual comparison between various reverberation effects has raised theoretical questions on the nature of late reverberation and on the impact of different simulation methods, which will be addressed in the next chapter.

Chapter 7

Discussion

The simple question of adjusting the delay and gain of the artifical reverberator has raised a number of complex questions which must be studied more thoroughly in the future. The main fields of inquiry and ideas for follow-up studies will be briefly presented in this chapter, including the generalization of our method to complex room geometries, the reverberation time estimation methods, the influence of scattering on simulations and on diffuseness metrics, and the anisotropy of the late reverberation field.

7.1 Generalization to arbitrary room geometries and non-exponential decays

While the mixing time estimation method shows good performance on a dataset comprising various room geometries, the shoebox rooms show the best performances. This has been confirmed by the results of the perceptual evaluation presented in Chapter 6, where the reverberation adjustment algorithm performs the best on the shoebox room. In fact, shoebox rooms have a predictable behavior which can be shown on their diffuseness profiles, and most analytic formulas have been derived for shoebox rooms.

It remains to be shown whether this method generalizes well to any geometry, and in particular if the room surface area is indeed a good predictor for the spatial incoherence value. The reverberation gain adjustment algorithm must be adapted to handle cases of nonexponential decays.

The reverberation time estimation, based on Eyring's formula, must also be generalized to complex rooms and nonexponential decays. It is known that Eyring's formula does not perform well when the late reverberation field deviates from homogeneous and isotropic conditions [8]. Methods for a more accurate estimation of the reverberation time per frequency band should be studied, including estimation of the reverberation time using fast ray tracing as a preprocessing step on the room independently of the source and receiver positions, or relying on radiosity [40].

Rooms with non-uniform distribution of absorption fall in the limits of Sabine and Eyring's formulas. They tend to have larger reverberation times than those predicted by these formulas and to present a double-decay slope. Zhou et al. [60] proposed a method for predicting the decay curve based on the room geometry, the distribution of the sound absorbing material and the scattering properties of the walls and furniture.

7.2 Influence of scattering

In its basic implementation, the ISM does not enable scattering, which is caused by the reflection of sound on rough surfaces. This phenomenon is not easily compatible with the ISM, as it violates the fundamental principle of the ISM that one image source corresponds to one acoustic path. On the contrary, ray tracing can easily account for scattering by introducing a random reflection law at boundaries. The absence of scattering on RIR simulations results in temporally sparse RIRs and an overestimation of the energy of early reflections [39].

Methods for introducing scattering in the ISM have been proposed, for example by Siltanen et al. [57] and more recently by Ewert et al. [14]. The goal is to introduce a temporal spreading of the response. In future work, the echo density and spatial incoherence metrics will be applied on RIRs simulated with an ISM implementing scattering, and the resulting diffuseness profiles will be compared to the diffuseness profiles obtained with the classical ISM.

7.3 Anisotropy of the late reverberation field

Mixing time estimation algorithms and artificial reverberators rely on the assumption that the late reverberation field is homogeneous and isotropic. As we have seen, this is rarely the case in practice. Anistropy has been shown to be perceptually noticeable [47], but the threshold for perception of anisotropy remains to be determined [3].

Massé et al. [32] showed that when performed in the spherical harmonic domain, the spatial incoherence metric confuses anisotropy with spatial coherence. In fact, a set of incoherent plane waves might be interpreted as coherent by this metric because of the anisotropy of the sound field. To solve that issue, he proposes to perform the spatial incoherence analysis on the plane-wave decomposition of the signals, so that the measure still manages to capture the incoherence of the sound field in the presence of an anisotropic sound field.

Thus, practical details of the implementation of artificial reverberation expose fundamental questions about its physical properties, which must be addressed together with perceptual investigations.

Chapter 8

Conclusion

While the combination of the Image Source Method and a Feedback Delay Network is widely used in auralization systems, implementation choices lack theoretical backup and perceptual validation. In particular, the delay and gain of the FDN relative to the image sources were the focus of this study. The mixing time is a physically and perceptually motivated choice for the late reverberation delay, since it is defined as the starting time of the validity of the stochastic model and of the impossibility of perceptual discrimination between different source and receiver configurations. In this work, a mixing time estimation algorithm was developed to operate on the early reflections. This has raised many questions regarding the quantification of diffuseness in the early part of a RIR simulated with the ISM. Although ISM simulations lack density due to the absence of scattering, it was shown that spatially encoding the image sources with a sufficient spatial resolution enables to quantify the growth of diffuseness in the first milliseconds of ISM simulations and thus to predict a mixing time. Building on this estimation, the late reverberation gain was computed as a function of the image source gain around the mixing time in order to preserve the continuity of the decay envelope.

The objective and perceptual evaluations of this algorithm suggest that although it shows good performance on simple rectangular rooms, it has limitations in certain room models that are characterized by non-exponential decays, e.g. coupled volumes or rooms with a non-uniform distribution of absorption. The late reverberation sound field may even deviate from the classical properties of homogeneity and isotropy. Those cases give poor results that may be due to incorrect mixing time estimation, unprecise reverberation time estimation or bad late reverberation synthesis. Further work must be devoted to tackle them. The perceptual significance of double-decay slopes and of anisotropic late reverberation must be thoroughly assessed. Future work will include implementing extensions of the simulations methods, for example by adding wave phenomena such as diffraction and scattering in the ISM or by synthesizing anisotropic late reverberation with directional FDNs, and observing the behavior of the diffuseness metrics on the resulting simulations.

Bibliography

[1] ANSI S1. 11. Specification for octave-band and fractional-octave-band analog and digital filters, 2004.

[2] Jonathan S Abel and Patty Huang. A simple, robust measure of reverberation echo density. In *Audio Engineering Society Convention 121*. Audio Engineering Society, 2006.

[3] Benoit Alary, Archontis Politis, Sebastian J Schlecht, and Vesa Välimäki. Directional feedback delay network. *AES: Journal of the Audio Engineering Society*, 67(10):752–762, 2019.

[4] Jont B Allen and David A Berkley. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950, 1979.

[5] Rémi Audfray and Jean-Marc Jot. Reverberation loudness model for mixed-reality audio. In *Audio Eng. Soc. Conf. Headphone Tech*, 2019.

[6] Samuel Bellows and Brian FG Katz. Calibrating geometric room acoustic models using a gradient descent algorithm. In *Audio Eng Soc Conv 156*, 2024.

[7] Stefan Bilbao, Brian Hamilton, Jonathan Botts, and Lauri Savioja. Finite volume time domain room acoustics simulation under general impedance boundary conditions. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(1):161–173, 2015.

[8] Sylvio R Bistafa and John S Bradley. Predicting reverberation times in a simulated classroom. *The Journal of the Acoustical Society of America*, 108(4):1721–1731, 2000.

[9] Jeffrey Borish. Extension of the image model to arbitrary polyhedra. *The Journal of the Acoustical Society of America*, 75(6):1827–1836, 1984.

[10] Thibaut Carpentier. Spat˜: a comprehensive toolbox for sound spatialization in max. *Ideas Sónicas*, 13(24): 12–23, 2021.

[11] Bengt-Inge L Dalenbäck. Room acoustic prediction based on a unified treatment of diffuse and specular reflection. *The journal of the Acoustical Society of America*, 100(2):899–909, 1996.

[12] Guillaume Defrance, Laurent Daudet, and J-D Polack. Using matching pursuit for estimating mixing time within room impulse responses. *Acta Acustica united with Acustica*, 95(6):1071–1081, 2009.

[13] Nicolas Epain and Craig T Jin. Spherical harmonic signal covariance and sound field diffuseness. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(10):1796–1807, 2016.

[14] Stephan D Ewert, Nico Gößling, Oliver Buttler, Steven van de Par, and Hongmei Hu. Computationally-efficient rendering of diffuse reflections for geometrical acoustics based room simulation. *Acta Acustica*, 9:9, 2025.

[15] Carl F Eyring. Reverberation time in "dead" rooms. *The Journal of the Acoustical Society of America*, 1 (2A_Supplement):168–168, 1930.

[16] Ilja Frissen, Brian FG Katz, and Catherine Guastavino. Effect of sound source stimuli on the perception of reverberation in large volumes. In *International Symposium on Computer Music Modeling and Retrieval*, pages 358–376. Springer, 2009.

[17] William G Gardner. A realtime multichannel room simulator. *J. Acoust. Soc. Am*, 92(4):2395, 1992.

[18] Christian Giguère and Sharon M Abel. Sound localization: Effects of reverberation time, speaker array, stimulus frequency, and stimulus rise/decay. *The Journal of the Acoustical Society of America*, 94(2):769–776, 1993.

[19] Georg Götz, Sebastian J Schlecht, and Ville Pulkki. A dataset of higher-order ambisonic room impulse responses and 3d models measured in a room with varying furniture. In *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, pages 1–8. IEEE, 2021.

[20] Takayuki Hidaka, Yoshinari Yamada, and Takehiko Nakagawa. A new definition of boundary point between early reflections and late reverberation in room impulse responses. *The Journal of the Acoustical Society of America*, 122(1):326–332, 2007.

[21] ISO. Iso 3382-1:2009, acoustics–measurement of room acoustic parameters–part 1: Performance spaces. *International Standards Organization*, 2009.

[22] Jean-Marc Jot and Antoine Chaigne. Digital delay networks for designing artificial reverberators. In *Audio Engineering Society Convention 90*. Audio Engineering Society, 1991.

[23] Jean-Marc Jot, Laurent Cerveau, and Olivier Warusfel. Analysis and synthesis of room reverberation based on a statistical time-frequency model. *PREPRINTS-Audio Engineering Society*, 1997.

[24] Jean-Marc Jot, Rémi Audfray, Mark Hertensteiner, and Brian Schmidt. Rendering spatial sound for interoperable experiences in the audio metaverse. In *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, pages 1–15. IEEE, 2021.

[25] Michael Kohnen, Florian Denk, Josep Llorca-Bofí, Birger Kollmeier, and Michael Vorländer. Coat - cross-site oldenburg-aachener transfer- functions, November 2020. URL `https://doi.org/10.5281/zenodo.4556707`.

[26] Heinrich Kuttruff and Michael Vorländer. *Room acoustics*. Crc Press, 2024.

[27] Winfried Lachenmayr. *Perception and quantification of reverberation in concert venues*. PhD thesis, Hochschule für Musik Detmold, 2017.

[28] Song Li and Jianyuan Feng. The influence of acoustic cues in early reflections on source localization. In *Audio Engineering Society Conference: 2022 AES International Conference on Audio for Virtual and Augmented Reality*. Audio Engineering Society, 2022.

[29] Alexander Lindau, Linda Kosanke, and Stefan Weinzierl. Perceptual evaluation of model-and signal-based predictors of the mixing time in binaural room impulse responses. *Journal of the Audio Engineering Society*, 60(11): 887–898, 2012.

[30] Paul Luizard and Brian FG Katz. Coupled volume multi-slope room impulse responses: a quantitative analysis method. In *Intl. Conf. on Auditorium Acoustics*, volume 33, pages 169–176, 2011.

[31] Pierre Massé. *Analysis, treatment, and manipulation methods for spatial room impulse responses measured with spherical microphone arrays*. PhD thesis, Sorbonne Université, 2022.

[32] Pierre Massé, Thibaut Carpentier, Olivier Warusfel, and Markus Noisternig. Denoising directional room impulse responses with spatially anisotropic late reverberation tails. *Applied Sciences*, 10(3):1033, 2020.

[33] Pierre Massé, Thibaut Carpentier, Olivier Warusfel, and Markus Noisternig. A robust denoising process for spatial room impulse responses with diffuse reverberation tails. *The Journal of the Acoustical Society of America*, 147(4):2250–2260, 2020.

[34] Thomas McKenzie, Leo McCormack, and Christoph Hold. Dataset of spatial room impulse responses in a variable acoustics room for six degrees-of-freedom rendering and analysis. *arXiv preprint arXiv:2111.11882*, 2021.

[35] Donald H Mershon, William L Ballenger, Alex D Little, Patrick L McMurtry, and Judith L Buchanan. Effects of room reflectance and background noise on perceived auditory distance. *Perception*, 18(3):403–416, 1989.

[36] James A Moorer. About this reverberation business. *Computer music journal*, pages 13–28, 1979.

[37] Gerhard Müller and Michael Möser. *Handbook of engineering acoustics*. Springer Science & Business Media, 2012.

[38] Markus Noisternig, Alois Sontacchi, Thomas Musil, and Robert Holdrich. A 3d ambisonic based binaural sound reproduction system. In *Audio Engineering Society Conference: 24th International Conference: Multichannel Audio, The New Reality*. Audio Engineering Society, 2003.

[39] Markus Noisternig, Brian FG Katz, Samuel Siltanen, and Lauri Savioja. Framework for real-time auralization in architectural acoustics. *Acta Acustica United with Acustica*, 94(6):1000–1015, 2008.

[40] Eva-Marie Nosal, Murray Hodgson, and Ian Ashdown. Investigation of the validity of radiosity for sound-field prediction in cubic rooms. *The Journal of the Acoustical Society of America*, 116(6):3505–3514, 2004.

[41] Alejandro Osses Vecchi, Glen McLachlan, and Armin Kohlrausch. Assessing the perceived reverberation in different rooms for a set of musical instrument sounds. *The Journal of the Acoustical Society of America*, 148(1): EL93–EL98, 2020.

[42] Stephanie Peichert, Julien De Muynke, and Brian FG Katz. Notre-dame whispers: A geolocated immersive walk through the sonic memory of notre-dame de paris cathedral. In *Résonances Gothiques*, pages 105–128. Collegium Musicae, 2023.

[43] David Poirier-Quinot, Brian FG Katz, and Markus Noisternig. Evertims: Open source framework for real-time auralization in architectural acoustics and virtual reality. In *20th International Conference on Digital Audio Effects (DAFx-17)*, 2017.

[44] Jean-Dominique Polack. *La transmission de l'énergie sonore dans les salles*. PhD thesis, Le Mans, 1988.

[45] Jean-Dominique Polack. Playing billiards in the concert hall: The mathematical foundations of geometrical room acoustics. *Applied Acoustics*, 38(2-4):235–244, 1993.

[46] Lorenzo Rizzi and Gabriele Ghelfi. Measuring mixing time in non-sabinian rooms: how scattering influences small-room responses. In *132nd AES Convention, Budapest*, 2012.

[47] David Romblom, Catherine Guastavino, and Philippe Depalle. Perceptual thresholds for non-ideal diffuse field reverberation. *The Journal of the Acoustical Society of America*, 140(5):3908–3916, 2016.

[48] Per Rubak and Lars G Johansen. Artificial reverberation based on a pseudo-random impulse response. In *Artificial Reverberation Based on a Pseudo-Random Impulse Response*, 1998.

[49] Lauri Savioja and U Peter Svensson. Overview of geometrical room acoustic modeling techniques. *The Journal of the Acoustical Society of America*, 138(2):708–730, 2015.

[50] Sebastian Schlecht. Fdntb: The feedback delay network toolbox. In *International Conference on Digital Audio Effects*, pages 211–218. DAFx, 2020.

[51] Michael Schoeffler, Fabian-Robert Stöter, Harald Bayerlein, Bernd Edler, and Jürgen Herre. An experiment about estimating the number of instruments in polyphonic music: A comparison between internet and laboratory results. In *ISMIR*, pages 389–394, 2013.

[52] Michael Schoeffler, Sarah Bartoschek, Fabian-Robert Stöter, Marlene Roess, Susanne Westphal, Bernd Edler, and Jürgen Herre. webmushra—a comprehensive framework for web-based listening tests. *Journal of Open Research Software*, 6(1), 2018.

[53] M Schröder. Eigenfrequenzstatistik und anregungsstatistik in räumen. *Acta Acustica united with Acustica*, 4(4): 456–468, 1954.

[54] Manfred R Schroeder. New method of measuring reverberation time. *The Journal of the Acoustical Society of America*, 37(3):409–412, 1965.

[55] Manfred R Schroeder. Digital simulation of sound transmission in reverberant spaces. *The Journal of the acoustical society of america*, 47(2A):424–431, 1970.

[56] Manfred R Schroeder. The "schroeder frequency"revisited. *The Journal of the Acoustical Society of America*, 99 (5):3240–3241, 1996.

[57] Samu Siltanen, Tapio Lokki, Sakari Tervo, and Lauri Savioja. Modeling incoherent reflections from rough room surfaces with image sources. *The Journal of the Acoustical Society of America*, 131(6):4606–4614, 2012.

[58] Rebecca Stewart and Mark Sandler. Statistical measures of early reflections of room impulse responses. In *Proc. of the 10th int. conference on digital audio effects (DAFx-07), Bordeaux, France*, pages 59–62, 2007.

[59] International Telecommunication Union. Method for the subjective assessment of intermediate quality level of audio systems. *International Telecommunication Union, Tech. Rep. ITU-R BS*, pages 1534–3, 2015.

[60] Xiaoru Zhou, Moritz Späh, Klaudius Hengst, and Ting Zhang. Predicting the reverberation time in rectangular rooms with non-uniform absorption distribution. *Applied Acoustics*, 171:107539, 2021.