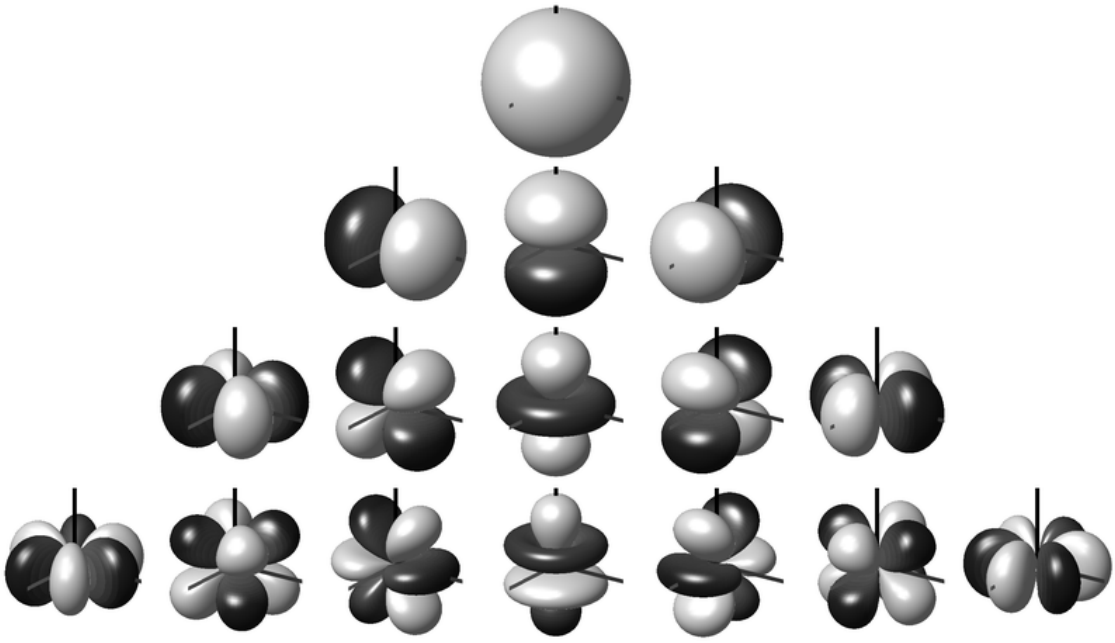

Covariance and Coherence Matching for Binaural Room Impulse Responses

Implementation and Evaluation of Rendering Methods Using Different-Order Ambisonics

Master's thesis in Master Programme Sound and Vibration

ELIN HEDLUND

DEPARTMENT OF ARCHITECTURE AND CIVIL ENGINEERING

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2023
www.chalmers.se

Master's thesis 2023

Covariance and Coherence Matching for Binaural Room Impulse Responses

Implementation and Evaluation of Rendering Methods Using Different-Order Ambisonics

ELIN HEDLUND

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Architecture and Civil Engineering
Division of Applied Acoustics
Chalmers University of Technology
Gothenburg, Sweden 2023

Covariance and Coherence Matching for Binaural Room Impulse Responses
Implementation and Evaluation of Rendering Methods Using Different-Order Ambisonics
ELIN HEDLUND

Supervisor: Jens Ahrens, Division of Applied Acoustics
Examiner: Jens Ahrens, Division of Applied Acoustics

Master's Thesis 2023
Department of Architecture and Civil Engineering
Division of Applied Acoustics
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Spherical harmonics up to third degree Obtained from [1].

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2023

Covariance and Coherence Matching for Binaural Room Impulse Responses
Implementation and Evaluation of Rendering Methods Using Different-Order Ambisonics
ELIN HEDLUND
Department of Architecture and Civil Engineering
Chalmers University of Technology

Abstract

Rendering binaural room impulse responses with the use of spherical microphone arrays and sets of head-related transfer functions minimize required measurements but will however exhibit a difference in perception due to, among other things, spherical harmonic truncation. By utilizing the spatial perceptually relevant inter-aural cues, the similarity to a directly measured BRIR could be enhanced.

In this thesis, two different methods for rendering binaural room impulse responses with different-order Ambisonics are employed to evaluate and compare the similarity between the generated BRIRs and directly measured counterparts using an artificial head and torso recording. Both methods utilize inter-aural cues within the inter-aural coherence and covariance matrix, respectively, in order to process the diffuse part of the BRIR.

The results show an improvement for the covariance matrix framework compared to the coherence matching method, even for rendering employing first-order Ambisonics. A pilot study in the form of an informal listening test was conducted to investigate the perceived similarity between the generated and measured BRIRs. The results of the listening test indicate that an increased spatial resolution due to higher-order Ambisonics will enhance the perceived similarity, while also displaying varying results depending on rendering method. However, a more extensive listening test would be beneficial. The findings indicate that the covariance framework with even higher-order Ambisonics would further optimize the quality and perceived similarity.

Keywords: Ambisonics, Binaural Rendering, Covariance, Coherence, Spherical Harmonics.

Acknowledgements

I would like to thank my supervisor Jens Ahrens for providing me with feedback on my work and Thomas Deppish for guidance in the theoretical understanding of the Ambisonics domain and general support in the thesis work. Additionally, thank you Leon Müller for helping with measurements and spontaneous questions.

Elin Hedlund, Gothenburg, March 2023

List of Acronyms

Below is the list of acronyms that have been used throughout this thesis listed in alphabetical order:

ACN	Ambisonic Channel Number
BRIR	Binaural Room Impulse Response
DOA	Direction of Arrival
FOE	Figure-of-Eight
FuMa	Furse-Malham
HOA	Higher Order Ambisonics
HRIR	Head Related Impulse Response
HRTF	Head Related Transfer Function
IC	Interaural Coherence
ILD	Interaural Level Difference
ITD	Interaural Time Difference
MagLs	Magnitude Least-Squares
MUSHRA	Multi Stimulus Test with Hidden Reference and Anchor
PIV	Pseudo-intensity Vector
RIR	Room Impulse Response
SMA	Spherical Microphone Array
SH	Spherical Harmonic
SHT	Spherical Harmonic Transform
STFT	Short Time Fourier Transform

Contents

1
Introduction

When generating a headphone-based virtual representation of a room, i.e., a binaural room impulse response (BRIR), several parameters affect the degree of similarity between said BRIR and what a listener would perceive when occupying the physical space.

Firstly, a room's acoustical behaviour, represented by the room impulse response (RIR), is roughly comprised of three parts: direct sound, early reflections, and diffuse sound. The direct sound and early reflections will typically be temporally separated and in extension hold distinct directions of arrival (DOA), as opposed to the reverberant sound which ideally emanates from all directions with the same energy simultaneously, making it time-invariant. Level differences and temporal relationships between the RIR components, the DOAs, and the length of the reverberant tail are all giving information about the sound field. By using spherical microphone arrays (SMA) consisting of several microphone capsules, and utilizing the full-sphere audio format Ambisonics, it is possible to represent a sound field in terms of the spherical harmonics (SH), which are defined on the surface of a sphere, centered around a listener. The amount of information that is attained when measuring RIRs is of course highly dependent on the spatial resolution of the obtained data which is closely linked to the order N Ambisonics. There is, however, a physical limitation on the number of microphone capsules that fit an SMA, which will cause spherical harmonic truncation.

Secondly, it is not only how the room behaves that includes vital information for the BRIR, but the characteristics of the listener's auditory system, which has the ability to localize and distinguish sound sources as well as get an impression of the space. This can be explained by different auditory cues. The interaural time difference (ITD) appears as the time between sound waves reaching the left and right ear, varying with incident angle. The interaural level difference (ILD), expressed in dB, is the intensity level difference between the left and right ear which also depends on the incidence angle where the sound waves diffract around the head and creates an intensity level difference. Interaural coherence (IC) is a spatial cue for the similarity between the two ears and is given on a scale between 0 and 1. The auditory cues are understandably dependent on frequency as well as the shapes and sizes of the listeners' heads, torsos, and ears. Such an impulse response that defines the auditory response and includes these auditory cues is denoted head-related impulse response (HRIR), or its spectral equivalent head-related transfer function (HRTF), and describes how sound arrives at a listener's ears, at a certain position and angle of incidence.

Measuring binaural room impulse responses for a large number of rooms, posi-

tions and angles will become tedious, even more so if separate measurements are performed for each room and listener. Therefore, less time-consuming approaches are needed to compute BRIRs efficiently while maintaining the characteristics of a measured counterpart. Using a set of HRTFs, which contain numerous HRTFs distributed over multiple incident angles, together with Ambisonic signals, binaural signals can be rendered, minimizing the equipment and measurements required.

The overall purpose of this thesis is to evaluate the processes of generating reverberant BRIRs with the utilization of interaural coherence and covariance. The processing will be performed using spherical harmonic expansion, for both first- and fourth-order Ambisonics, with the use of SMA recordings and HRTFs.

1.1 Related Work and Overview

Several studies explore ways to generate BRIRs from HRTF sets in combination with RIRs using different-order spherical harmonics. This thesis will base the work on two of these methods and make evaluations and comparisons between them. Both methods base their processing of the reverberant parts of the BRIR with a focus on the interaural coherence or covariance of an HRTF set.

The first method for this thesis is presented by Menzer et al. in the article titled *Obtaining Binaural Room Impulse Responses From B-Format Impulse Responses Using Frequency-Dependent Coherence Matching* [2]. Here, the authors suggest a method for generating the reverberant part of a first-order Ambisonic (FOA) BRIR with the objective of obtaining the same power spectra and interaural coherence as for a BRIR in an ideally diffuse sound field.

The second method is presented by Vilkamo et al. in the paper titled *Optimized Covariance Domain Framework for Time-Frequency Processing of Spatial Audio* [3]. The method comprises a spatial audio processing framework for obtaining reverberant BRIRs of any order Ambisonics, utilizing the covariance matrix and binaural rendering filters. The binaural rendering filters are generated using a magnitude least-squares solution [4].

Chapter 2 contains a summary of the relevant theory on which this thesis is based. This includes the full-sphere audio format Ambisonics with descriptions of the spherical harmonics, head-related transfer functions, spherical microphone arrays with spherical microphone encoding, binaural rendering using magnitude least-squares, coherence and covariance which extends to the interaural coherence and interaural covariance matrix, and lastly an overview of the short time Fourier transform.

In chapter 3, the implementation of the thesis work is explained which includes the processing steps of generating BRIRs with the use of the two methods, and descriptions of the spherical microphone array and binaural room impulse response measurements, using an artificial head and torso, that were performed in three separate rooms.

Chapter 4 includes the evaluation of the thesis work which comprise the results for the three different rooms following the implemented methods, a description of the pilot study which was conducted including its results, and a discussion on the thesis results.

Lastly, a summary conclusion on the outcome of the thesis' methods and results is presented in chapter 5, along with an outlook for future developments.

2
Theory

2.1 Ambisonics

To accurately portray a three-dimensional sound field, it is desirable to construct a non-interrupted spherical representation of it, which is the concept of Ambisonics. Unlike traditional channel-based stereo or surround sound which directs sound to a certain number of sources, the scene-based Ambisonics full-sphere depiction of a sound field utilizes spherical harmonics which are defined continuously on a sphere centered around the listener.

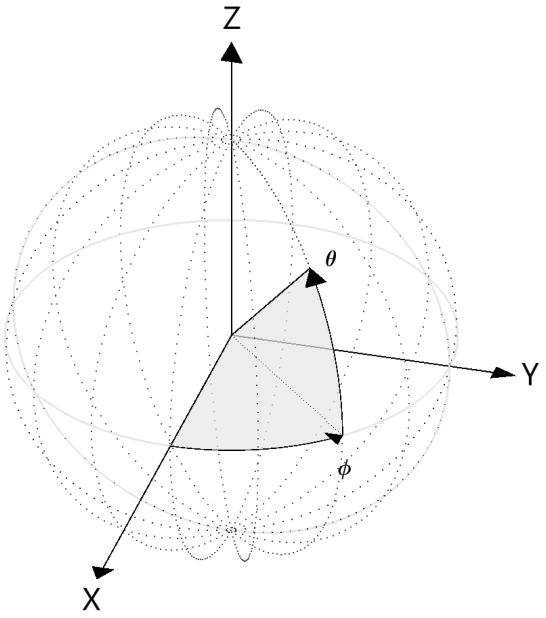

Figure 2.1: Spherical coordinate system, with azimuth, ϕ, and elevation, θ.

The spherical coordinates in which Ambisonics operates are defined by (r, ϕ, θ), where r is the distance to the origin of the spherical coordinate system. The coordinates azimuth, $\phi \in [0, 2\pi]$, and elevation, $\theta \in [-\pi, \pi]$, can be obtained from the cartesian three-dimensional coordinate system, (X, Y, Z), by using the relationships

$$\phi = \arctan\left(\frac{Y}{X}\right) \tag{2.1}$$

$$\theta = \arctan\left(\frac{Z}{\sqrt{X^2 + Y^2}}\right) , \tag{2.2}$$

i.e., the angle of the orthogonal projection to the XY-plane is the azimuth and the polar angle from the orthogonal projection towards the Z-axis is the elevation. Note that other conventions of defining the polar angle can be used, such as colatitude, which is the polar angle originating from the Z-axis. For this thesis, elevation is used. An illustration of the spherical coordinate system is shown in Figure 2.1.

2.1.1 Spherical Harmonics

The spherical harmonics which are defined on the surface of a sphere and utilized when constructing a representation of a three-dimensional sound field are expressed for order $n \geq 0$ and degree m $(-n \leq m \geq n)$ as

$$Y_n^m(\theta, \phi) = N_n^{|m|} P_n^{|m|}(\cos\theta)\ \mathrm{e}^{im\phi}\ , \tag{2.3}$$

where $N_n^{|m|}$ is a normalization term, the exponential expression $\mathrm{e}^{im\phi}$ correlates to the dependency on azimuth, and $P_n^{|m|}$ are the Legendre polynomials which together with $\cos\theta$ form the Legendre functions and thus an orthonormal basis that defines the dependence of elevation. The order and degree determine the spatial resolution of the sound field representation. For order $n = 0$, only one SH exists (degree $m = 0$) with omnidirectional directivity, hence poor spatial resolution. For order $n = 1$, three additional channels are included which have a forward, leftward, and upward figure-of-eight (FOE) directional response. In Figure 2.2, spherical harmonics up to order $n = 3$ are visualized. Here it is clear that the spherical harmonic degree indicates the number of waves experienced if one travels along the sphere in a horizontal direction, and the order indicates $n - |m|$ waves, traveling in the vertical

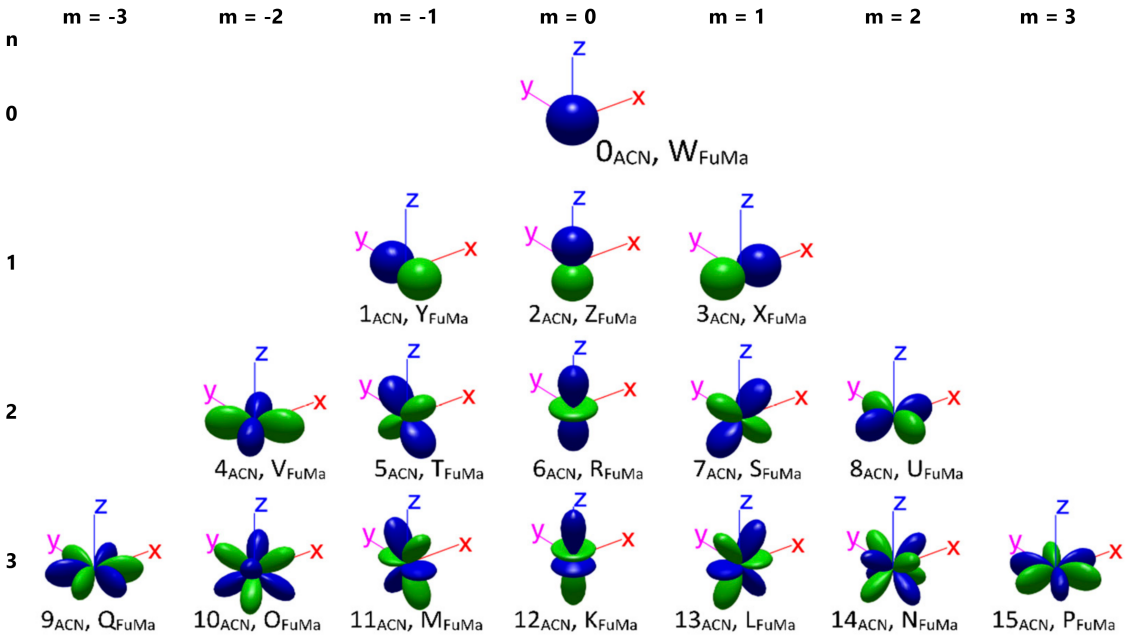

Figure 2.2: Visualization of the spherical harmonics with FuMa and ACN channel ordering for the ambisonic domain. [5], edited.

direction, i.e, a higher order enhances the resolution and more sources are distinguishable. The normalization term $N_n^{|m|}$ in equation 2.3 is defined by

$$N_n^{|m|} = \sqrt{\frac{2n+1}{4\pi}\frac{(n-m)!}{(n+m)!}} \ , \tag{2.4}$$

which is labeled N3D (full 3D normalization) and adjusts the signals so that no channel will exceed the level of the spherical harmonic of zeroth order. Note that this normalization term may have a variation depending on the convention and area of use, which is also true for the spherical harmonic notation. Figure 2.2 includes the SH ordering notation according to two common SH ordering standards; Ambisonic Channel Number (ACN) and Furse-Malham (FuMa). ACN ordering is preferably used for higher order spherical harmonics which is discussed further in section 2.1.3.

The space-domain sound pressure defined in ambisonics at a certain location with radius, r and direction $\Omega = (\theta, \phi)$ is described by

$$p(k, r, \Omega) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} p_n^m(k, r, \Omega) Y_n^m(\Omega_0)^* \ , \tag{2.5}$$

where $Y_n^m(\Omega_0)$ are the spherical harmonics, $(\cdot)^*$ denotes the complex conjugate and $k = \omega/c$ is the wave number with angular frequency ω and speed of sound c. The term $p_n^m(k, r, \Omega)$ is the sound pressure from a plane wave at aforesaid location defined by

$$p_n^m(k, r, \Omega) = b_n(k, r) Y_n^m(\Omega) \ , \tag{2.6}$$

where $b_n(k, r)$ includes the radial dependency according to

$$b_n(k, r) = 4\pi i^n j_n(k, r) \ , \tag{2.7}$$

where $j_n(k, r)$ is the spherical Bessel function which is a mathematical function describing in- and outward wave propagation.

For more detailed information on the math and components of Ambisonics, see [6].

2.1.2 Head Related Impulse Responses

The head related impulse response, HRIR, or head related transfer function, HRTF, in frequency domain describes how a sound signal arrives at a listeners ears, at a certain position and angle of incidence. A HRIR can be recorded in a room by inserting microphones in a listeners ears which will generate the binaural room impulse response for that individual and includes the acoustical effects arising as a result of the size and shape of the torso, head and ears. This will understandably only give an ideal response for that listener which is why an artificial head and torso can be used, which is designed to represent the head and torso of an average listener. Varying deviations will certainly still occur for listeners, but the artificial head HRTFs are more widely applicable.

Measuring HRIRs in anechoic conditions generates filters that include the sound experienced by a listener without the influence of a specific physical space. If these

measurements are performed for evenly spaced directions in terms of azimuth and elevation over a sphere, a three-dimensional representation of the auditory response is obtained, denoted HRTF set. In Figure 2.3a, an example of an artificial head and torso is shown.

2.1.3 Spherical Microphone Arrays

There are different types of spherical microphone arrays (SMA) that are used to record three-dimensional representations of a sound field, with the most common representing FOA signals. The FOA microphone arrays, commonly referred to as B-format microphones, comprise four microphone capsules arranged in a tetrahedral array with a cardioid or sub-cardioid directivity which is used to form the channels in the spherical harmonic domain. This includes a channel with an omnidirectional response and three channels with a forward, leftward, and upward figure-of-eight (FOE) directional response, i.e., directivity as for the four spherical harmonic channels described in section 2.1.1. In Figure 2.3b, an example of a B-format microphone is shown.

Higher-order Ambisonic (HOA) microphone arrays comprise a larger number of microphone capsules than that of the B-format microphone. The microphone capsules can be positioned on the surface of a rigid sphere as shown in Figure 2.3c. However, spatial limitations of the array introduce spatial aliasing for higher frequencies, following the microphones not being mounted sufficiently close together. For lower frequencies, larger arrays are sought after but could be cumbersome to handle and increase the scattering from the body of the rigid sphere.

Encoding the signals from a microphone array to equivalent spherical harmonic signals is performed using spherical harmonic transform (SHT). It is essentially the process of breaking down the pressure signals from the SMA and observing to what extent they compose directivity patterns that are equivalent to those of the spherical harmonics, i.e., omnidirectional, FOE, etc. This will, of course, be limited in relation to the number of microphone capsules in the array. The SHT can be performed using a least-squares solution [4] in order to find the coefficients ψ which are expanded over the spherical harmonics that most closely model the pressure signals, p_{SMA}, according to

$$\psi_N = \arg\min_{\psi_N} \left[\left\| Y_N \psi_N - p_{\text{SMA}} \right\|^2 \right] \ , \tag{2.8}$$

where $\|\cdot\|$ is the norm. For the discrete samples from the SMA, this is performed as

$$\psi_N = (Y_N)^\dagger \ p \ , \tag{2.9}$$

where $(\cdot)^\dagger$ denotes the pseudo-inverse. See [6] for further derivation of the SMA encoding.

For B-format microphone signals decoded to Ambisonic signals, the channels are commonly labeled w, x, y, z according to FuMa. Spherical harmonic channel ordering in accordance with FuMa is an expansion of the B-format with lettered notation, which is intuitive for first-order Ambisonics where the harmonics have a

clear directionality towards the X-, Y- and Z-axis. This becomes inconvenient for higher orders, which is why a component ordering with symmetry around $m = 0$, i.e., Z-rotational harmonics, is preferable. As shown in Figure 2.2, this is the convention of notation according to ACN.

(a) **(b)** **(c)**

Figure 2.3: (a) GRAS 45BB KEMAR Head & Torso [7] (b) RØDE NT-SF1 B-format microphone array [8] (c) mh acoustics' em32 Eigenmike fourth-order microphone array [9], Edited.

2.1.4 Binaural Rendering Using Magnitude Least-Squares

An ideal sound field is not order-limited, as opposed to the result of the SHT. When truncating higher-order components to lower-order Ambisonics, aliasing and spectral roll-off for higher frequencies will occur. One explanation is the off-center location of the ears in the coordinate system of the spherical harmonics. Since the ITDs are less relevant for higher frequencies [10], the linear phases for higher order modes can be removed and the energy loss due to the truncation will be reduced. One method of avoiding these unwanted effects while performing binaural decoding is by the means of magnitude least-squares (MagLs). The magnitude least-squares rendering filter [4], Q, is obtained similarly to the SHT according to

$$\begin{aligned} Q(\omega_k) = \arg\min_{Q} \Big[& \lambda(\omega_k) \| Y_M Q - h_M \|^2 \dots \\ & + (1 - \lambda(\omega_k)) \| |Y_M Q| - |h_M| \|^2 \Big] , \end{aligned} \tag{2.10}$$

where Y_M are the SHs for the set of directions M, h_M are the HRTFs, ω_k is the center angular frequency of the k-th bin and

$$\lambda(\omega_k) = \begin{cases} 1 \text{ for } & \omega \leq \omega_c \\ 0 \text{ for } & \omega > \omega_c \end{cases} , \tag{2.11}$$

where ω_c is the cut-on frequency. This leads to an approximation of the magnitudes for frequencies above ω_c while ignoring the phase error. In practice, as in equation 2.9, the solution for the binaural rendering filter is obtained by multiplying the pseudoinverse of Y_M with the HRTF set and adding an exponential term with a phase shift corresponding to the previous bin for $\omega > \omega_c$.

2.2 Coherence and Covariance

The coherence function estimates the linear relationship between two signals according to

$$\Phi_{xy}(\omega) = \frac{\left|S_{xy}(\omega)\right|}{\sqrt{S_{xx}(\omega)S_{yy}(\omega)}}\,, \tag{2.12}$$

where ω is the angular frequency, S_{xy} is the cross-spectrum and S_{xx} and S_{yy} are the autospectra of signal x and y respectively [11]. The coherence function will generate values between zero and one where one indicates two signals with a perfect linear relationship.

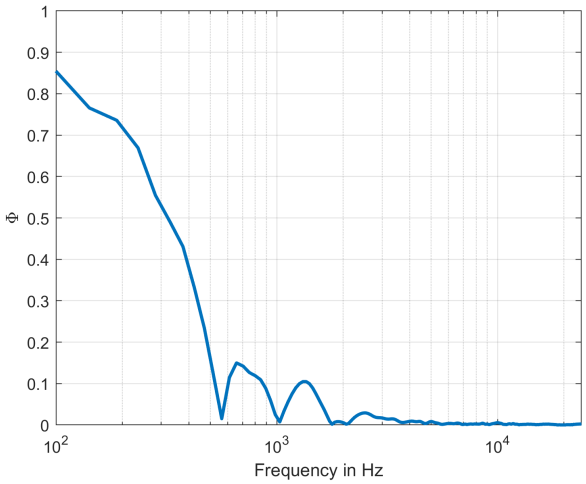

Figure 2.4: Interaural coherence for an HRTF set which represents the similarity between left and right ear signals in a diffuse sound field.

The interaural coherence, IC, is a measure of the similarity of the sound between the left and right ear of a listener which in extension includes spatial information of the sound field. In Figure 2.4, the interaural coherence for a set of HRTFs, measured in anechoic conditions, is shown, where it is visible that the IC reaches towards zero for the greater part of the frequency spectrum. With IC close to zero, the left and right ear will perceive two independent sounds. For lower frequencies with wavelengths that are greater than the distance between left and right ear, the coherence will start to reach towards one, since these sound waves are able to reach both ears at the same time. With IC close to one, i.e. the left and right ear signals are identical, the perceived sound in the left and right ear will be combined

to a single sound located in the center of the head of the listener, in other words more focused [12]. The interaural coherence for an HRTF set, as in Figure 2.4, is comparable to the ideal diffuse interaural coherence.

The covariance matrix, $\mathbf{C}_\mathbf{x}$ of an audio signal $\mathbf{x}$ is obtained according to

$$\mathbf{C}_\mathbf{x} = \mathrm{E}\left[\mathbf{x}\mathbf{x}^\mathrm{H}\right] , \tag{2.13}$$

where H marks the conjugate transpose and E $[\cdot]$ is the expected value. This creates a $N_x \times N_x$ matrix containing the correlation between each pair of signal elements. For a binaural signal, $\mathbf{y}$, the channel energies are included in the covariance matrix and would accordingly contain the auto- and cross-correlation according to

$$\mathbf{C}_\mathbf{y} = \begin{bmatrix} y_L y_L^\mathrm{H} & y_L y_R^\mathrm{H} \\ y_R y_L^\mathrm{H} & y_R y_R^\mathrm{H} \end{bmatrix} , \tag{2.14}$$

where L denotes the left ear signal and R denotes the right ear signal. Hence, the coherence can be determined from the covariance matrix. Note that for a finite number of data points, the expectancy operator is replaced by an average operator which is performed with the use of short-time Fourier transform (STFT), explained in Section 2.3.

Moreover, since the covariance matrix fulfills the requirement of being hermitian (equal to its own conjugate transpose) and positive-semi-definite, it can be decomposed as

$$\mathbf{C} = \mathbf{K}\mathbf{K}^H , \tag{2.15}$$

where $\mathbf{K}$ is obtained through eigendecomposition of $\mathbf{C}$ which generates matrices $\mathbf{S}$ and $\mathbf{U}$ which are the eigenvalues and right eigenvectors respectively with the relationship

$$\mathbf{K} = \mathbf{U}\sqrt{\mathbf{S}} . \tag{2.16}$$

Since the decomposition condition is valid for any unitary matrix $\mathbf{P}$, the covariance matrix can be decomposed according to

$$\mathbf{C} = \mathbf{K}\mathbf{P}\mathbf{P}^H\mathbf{K}^H . \tag{2.17}$$

2.3 Short Time Fourier Transform

If the frequency content of a signal is obtained by Fourier transform, the result includes the total frequency information averaged over the whole signal. This is a bad solution if there are significant changes in frequencies over time. In order to monitor the behaviour of a signal over time, a short time Fourier transform (STFT) can be calculated. The procedure of performing the STFT is to divide a signal in time segments of equal length and performing the Fourier transform for each time-block. The segmenting is executed by sliding a window in time which is applied to the signal, usually including an overlap in order to avoid or reduce

unwanted artifacts at the edges. However, if the signal is to be recreated, care needs to be taken to be able to reconstruct the signal accurately when performing the inverse short time Fourier transform (ISTFT). There is a trade-off between time and frequency resolution when choosing the window length and the overlap needs to be chosen so that the signal is reconstructed perfectly. In Figure 2.5, an illustration of the STFT steps is shown, where an overlap length of 50 % is applied.

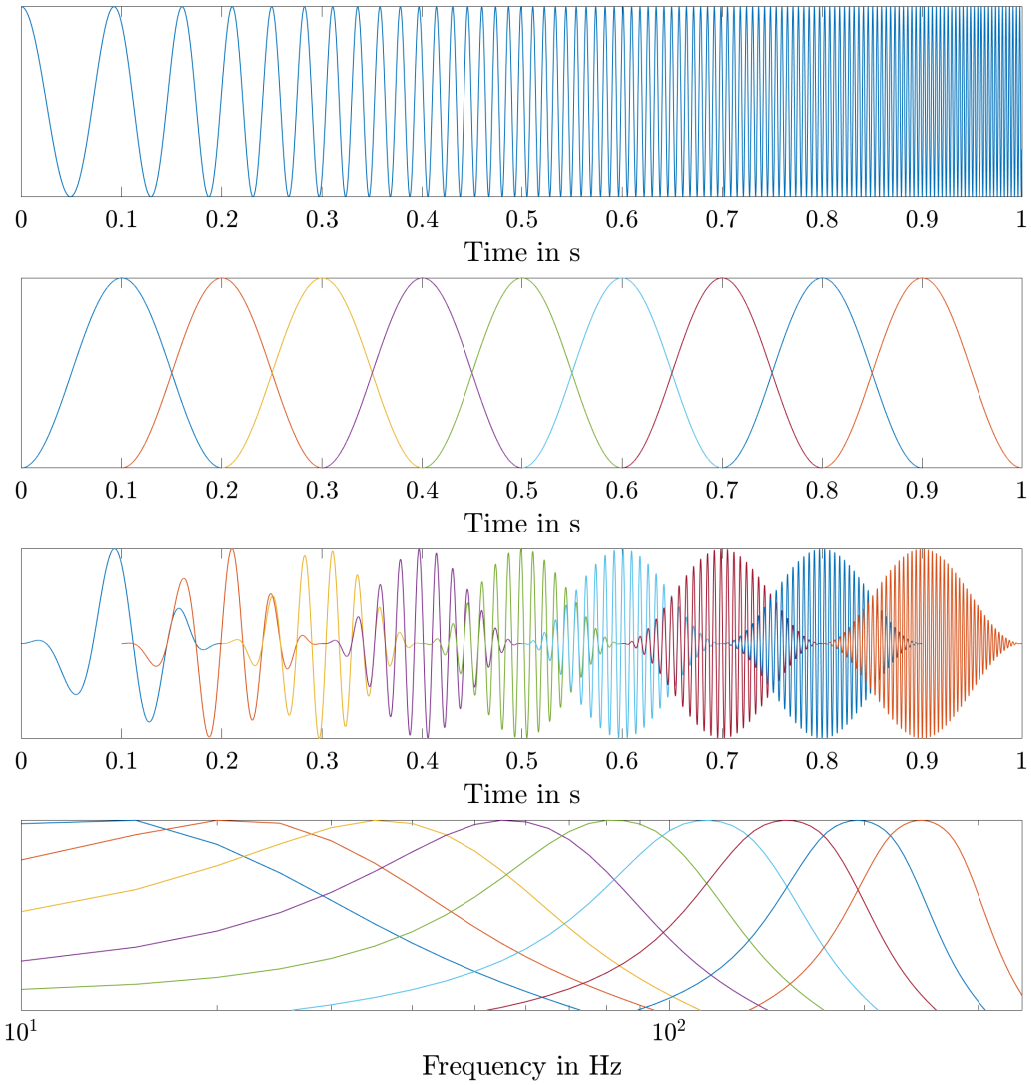

Figure 2.5: The main steps of the short-time Fourier Transform. The graphs from top to bottom include the original time signal, the windowing function in time, the time segments filtered by the window, and the time segments in the frequency domain.

3

Implementation

The processing performed for this thesis was executed in MATLAB with the use of short-time Fourier transform. However, the following derivations in this section will be expressed for each time block, for the sake of brevity. In the processing, the block length was set to 128 samples with an overlap of 50 % for perfect reconstruction and the FFT length was set to 1024 samples, which was found to be a good compromise between desired result and computational load. Where HRTFs were required, a set from [13] was used which contain a full sphere $1^\circ \times 1^\circ$ resolution HRTF data set obtained using the KEMAR head and torso simulator.

3.1 Obtaining BRIRs From FOA RIRs Using Coherence Matching

This section includes the method of generating binaural room impulse responses using head-related transfer functions and spherical microphone array recordings rendered to first-order ambisonics, which enables separate measurements of RIRs and sets of HRTFs. The approach follows the methods presented by Menzer et al. [2], where BRIRs are generated by processing direct and reverberant sounds separately. The objective of this method is to generate a left and right BRIR which have a reverberant part modeled from the power spectra and coherence of an ideally diffuse left and right BRIR. For this method, the Ambisonics ACN channels 0, 1, 2 and 3 will be denoted according to FuMa with w, y, z and x respectively.

3.1.1 Generating Direct BRIR

The first step in this method is to separate the direct and reverberant part, where the reverberant part contains both early and late reflections. The separation was obtained by localizing the absolute minimum within 10 ms after the absolute maximum of the energy envelope of the omnidirectional spherical harmonic channel, $w(t)$. To improve on the results, the 10 ms interval was reduced to exclude eventual early reflections which would alter the DOA of the BRIR. An example of the recommended and applied split between the direct and reverberant parts of $w(t)$ is illustrated in Figure 3.1.

The direct part of $w(t)$, $w_{direct}(t)$, was filtered with the HRIR which most closely corresponded with the estimated direction of arrival, DOA, of $w_{direct}(t)$. The DOA of the direct BRIR, in terms of azimuth and elevation, was estimated according to

equations 2.1 and 2.2, where the magnitudes of channels x, y and z were computed by utilizing the pseudo-intensity vector PIV, $\vec{I}$ according to

$$\vec{I} = \begin{cases} I_x = \sum\limits_{n \in T_D} x(t)w(t) \\ I_y = \sum\limits_{n \in T_D} y(t)w(t) \\ I_z = \sum\limits_{n \in T_D} z(t)w(t) \end{cases} , \tag{3.1}$$

where T_D is the time interval of the direct sound.

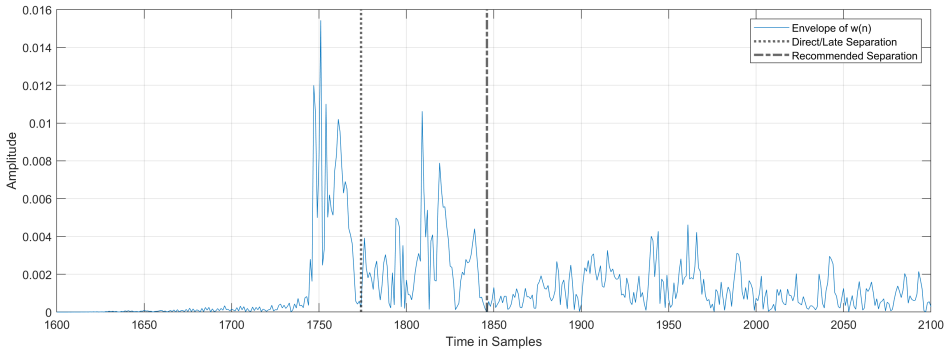

Figure 3.1: Example of separation between the direct and diffuse part of the omnidirectional spherical harmonic channel, where both the, by the model, recommended separation, and the adjusted separation are shown.

3.1.2 Generating Late BRIR

An ideally diffuse sound field will have sound arriving independently from all directions with the same power. With D_i as the diffuse sound with a certain incident angle $i \in \{1, 2, ..., A\}$, an ideal late omnidirectional response, W_{late}, is obtained as

$$W_{\text{late}} = \sum_{i=1}^{A} D_i(\omega) \ , \tag{3.2}$$

where W_{late} is the frequency spectrum of w_{late}, ω is the angular frequency and A is the number of incident angles. From this expectation, the ideal left and right late BRIR are computed as

$$B_{L,\text{late}} = \sum_{i=1}^{A} L_i(\omega)D_i(\omega) \tag{3.3}$$

$$B_{R,\text{late}} = \sum_{i=1}^{A} R_i(\omega)D_i(\omega) \ , \tag{3.4}$$

where $L_i(\omega)$ and $R_i(\omega)$ are the left and right HRTFs. Consequently, the power spectra of the left and right ideal BRIRs are calculated according to

$$P_L(\omega) = \frac{|W_{\text{late}}(\omega)|^2}{A} \sum_{i=1}^{A} |L_i(\omega)|^2 \tag{3.5}$$

$$P_R(\omega) = \frac{|W_{\text{late}}(\omega)|^2}{A} \sum_{i=1}^{A} |R_i(\omega)|^2 \ . \tag{3.6}$$

From equation 2.12, the interaural coherence between the left and right BRIR is expressed as

$$\Phi(\omega) = \frac{\left|\langle B_{L,\text{late}}(\omega) B^*_{R,\text{late}}(\omega)\rangle\right|}{\sqrt{\langle |B_{L,\text{late}}(\omega)|^2\rangle \langle |B_{R,\text{late}}(\omega)|^2\rangle}} \ , \tag{3.7}$$

where $\langle\rangle$ denotes the expected value and * denotes the complex conjugate. To obtain the target coherence which is applied to the generated left and right BRIR, the IC is calculated in terms of the left and right HRTFs according to

$$\Phi(\omega) = \frac{\left|\sum_{i=1}^{A} L_i(\omega) R_i^*(\omega)\right|}{\sqrt{\sum_{i=1}^{A} |L_i(\omega)|^2 \sum_{i=1}^{A} |R_i(\omega)|^2}} \ . \tag{3.8}$$

The azimuthal directional responses of the left and right late BRIR can be modeled according to

$$D_L(\omega, \phi) = H_L(\omega)(v(\omega) + (1 - v(\omega))cos\phi) \tag{3.9}$$

$$D_R(\omega, \phi) = H_R(\omega)(v(\omega) - (1 - v(\omega))cos\phi) \ , \tag{3.10}$$

where $H_L(\omega)$ and $H_R(\omega)$ are filters that adjust the power spectrum. By utilizing the directional responses in the magnitude of the coherence, it can be shown that the result from equation 3.8 can be expressed in terms of a frequency dependant constant $v(\omega)$ according to

$$v(\omega) = \frac{\Phi(\omega) + 1}{3\Phi(\omega) - 1} - \frac{\sqrt{4(\Phi(\omega) + 1)^2 - 4(3\Phi(\omega) - 1)(\Phi(\omega) + 1)}}{6\Phi(\omega) - 2} \ . \tag{3.11}$$

The generated left and right late BRIRs are then calculated as

$$B_{L,\text{late,gen}}(\omega) = H_L(w) \left(v(\omega) W_{\text{late}}(\omega) + \frac{Y_{\text{late}}(\omega)}{\sqrt{2}} (1 - v(\omega)) \right) \tag{3.12}$$

$$B_{R,\text{late,gen}}(\omega) = H_R(w) \left(v(\omega) W_{\text{late}}(\omega) - \frac{Y_{\text{late}}(\omega)}{\sqrt{2}} (1 - v(\omega)) \right) \ . \tag{3.13}$$

With the assumption that the the power spectra of the ideal left and right BRIRs are equal to the generated left and right BRIRs, the filters $H_L(\omega)$ and $H_R(\omega)$ are

obtained by equating equations 3.5 and 3.6 with the power spectra of equations 3.12 and 3.13 according to

$$H_L(\omega) = \frac{\sqrt{P_L(\omega)}}{\left|v(\omega)W_{\text{late}}(\omega) + \frac{1}{\sqrt{2}}(1 - v(\omega))Y_{\text{late}}(\omega)\right|} \tag{3.14}$$

$$H_R(\omega) = \frac{\sqrt{P_R(\omega)}}{\left|v(\omega)W_{\text{late}}(\omega) - \frac{1}{\sqrt{2}}(1 - v(\omega))Y_{\text{late}}(\omega)\right|} \ . \tag{3.15}$$

3.2 Obtaining BRIRs Using Covariance Domain Framework

This section includes the method of generating BRIRs using head related transfer functions and spherical microphone array recordings rendered to both first and fourth-order ambisonics. The approach follows the methods presented by Vilkamo et al. [3], where the diffuse part of the BRIR is constructed using a covariance domain framework. This method will only be used for processing the late BRIR, i.e., the direct part of the BRIR is obtained the same way as described in section 3.1.1. Since higher order Ambisonics is utilized for this method, ACN notation will be adopted. Note that $\mathbf{x}$ and $\mathbf{y}$ will be used as notation for the input and output signal matrices respectively and should not be confused with channel x and y of the FuMa notation.

The objective for this method is to find a mixing matrix, $\mathbf{M}$, which modifies the N_x number of input signals in $\mathbf{x}$ to the binaural output signal $\mathbf{y}$, which has the target covariance matrix, $\mathbf{C_y}$. The target covariance is established by determining the interaural coherence for the HRTF set according to equation 2.14, which contributes to the auditory response of an ideally diffuse sound field. However, to be able to give an accurate depiction of the actual magnitude response and eventual early directional responses originating from the input signals, a prototype signal is constructed according to

$$\hat{\mathbf{y}} = \mathbf{Q}\mathbf{x} \ , \tag{3.16}$$

which includes the input signals and a binaural magnitude least-squares rendering filter, $\mathbf{Q}$, which is generated from a MATLAB script obtained from [14]. Following equation 2.13, the prototype signal has the covariance matrix

$$\mathbf{C}_{\hat{\mathbf{y}}} = \mathrm{E}\left[\mathbf{Q}\mathbf{x}\mathbf{x}^H\mathbf{Q}^H\right] = \mathbf{Q}\mathbf{C}_\mathbf{x}\mathbf{Q}^H \ . \tag{3.17}$$

The prototype signal is used to modify the target covariance matrix which is constructed according to

$$\mathbf{C_y} = \begin{bmatrix} c_{\hat{y}_{LL}} & \gamma c_{y_{LR}} \\ \gamma c_{y_{RL}} & c_{\hat{y}_{RR}} \end{bmatrix} \ , \tag{3.18}$$

where γ is a ratio scalar which adapts the magnitude of the HRTF set IC, $\mathbf{C_y}$, to the IC of the input signals and binaural rendering filter, $\mathbf{C_{\hat{y}}}$, according to

$$\gamma = \sqrt{\frac{c_{\hat{y}_{LL}} c_{\hat{y}_{RR}}}{c_{y_{LL}} c_{y_{RR}}}} \ . \tag{3.19}$$

The output signal is formulated as

$$\mathbf{y} = \mathbf{M}\mathbf{x} \ , \tag{3.20}$$

where for a binaural output, $N_y = 2$. Following the same approach as for equation 3.17, the target covariance is expressed as

$$\mathbf{C_y} = \mathrm{E}\left[\mathbf{M}\mathbf{x}\mathbf{x}^H\mathbf{M}^H\right] = \mathbf{M}\mathbf{C_x}\mathbf{M}^H \ . \tag{3.21}$$

Together with the unitary matrices, $\mathbf{P_x}$ and $\mathbf{P_y}$, and the decomposition conditions according to equation 2.17, equation 3.21 can further be expanded as

$$\mathbf{K_y}\mathbf{P_y}\mathbf{P_y}^H\mathbf{K_y}^H = \mathbf{M}\mathbf{K_x}\mathbf{P_x}\mathbf{P_x}^H\mathbf{K_x}^H\mathbf{M}^H \ . \tag{3.22}$$

The final mixing matrix, $\mathbf{M}$, is then simplified as

$$\mathbf{M} = \mathbf{K_y}\mathbf{P}\mathbf{K_x^{-1}} \ , \tag{3.23}$$

where $\mathbf{P} = \mathbf{P_y}\mathbf{P_x}^H$ is a unitary matrix which is constructed to minimize the error measure e which is expressed according to

$$e = \mathrm{E}\left[\left\|\mathbf{G_{\hat{y}}}\hat{\mathbf{y}} - \mathbf{y}\right\|^2\right] \ , \tag{3.24}$$

where $\|\cdot\|$ is the norm operator and $\mathbf{G_{\hat{y}}}$ is constructed to normalize the energies of $\hat{\mathbf{y}}$ to those of $\mathbf{y}$ with the diagonal elements

$$g_{\hat{y}_{ii}} = \sqrt{\frac{c_{y_{ii}}}{c_{\hat{y}_{ii}}}}, \ i = 1, 2 \ . \tag{3.25}$$

For a derivation of the minimization of e, see [3], in which it is determined that the optimal $\mathbf{P}$ is obtained for

$$\mathbf{P} = \mathbf{V}\mathbf{\Lambda}\mathbf{U}^H \ , \tag{3.26}$$

where Λ is an zero-padded identity matrix with $N_y \times N_x$ elements and the unitary matrices $\mathbf{V}$ and $\mathbf{U}$ are obtained trough the single value decomposition

$$\mathbf{U}\mathbf{S}\mathbf{V}^H = \mathbf{K_x}^H\mathbf{Q}^H\mathbf{G_{\hat{y}}}^H\mathbf{K_y} \ . \tag{3.27}$$

If the inverse $\mathbf{K_x^{-1}}$ in equation 3.23 include very large components following small components in $\mathbf{x}$, a regularization factor can be applied using $\mathbf{S_x}$ from the single value decomposition of $\mathbf{K_x}$ according to

$$\mathbf{K_x} = \mathbf{U_x}\mathbf{S_x}\mathbf{V_x}^H \ , \tag{3.28}$$

where the regularization factor, α, which was set to 0.2 in reference to [3], scales the maximum component in $\mathbf{S_x}$ and creates a scaled diagonal matrix, $\mathbf{S_x'}$, according to

$$s'_{x_{ii}} = \max[s_{x_{ii}}, \alpha s_{x_{\max}}],\ i = 1, 2\ . \tag{3.29}$$

The diagonal matrix is used for calculating the regularized inverse

$$\mathbf{K_x'^{-1}} = \mathbf{V_x S_x'^{-1} U_x^H}\ , \tag{3.30}$$

which is then used as substitution for $\mathbf{K_x^{-1}}$ in equation 3.23.

3.3 Measurements

In order to generate, evaluate and compare the generated binaural room impulse responses, corresponding spherical microphone array and BRIR measurements were performed in three rooms. The SMA recordings were performed using mh acoustics' em32 Eigenmike [9], which contain 32 microphone capsules and is capable of capturing up to fourth-order Ambisonics. The reference BRIRs were recorded using the same model artificial head as for the used HRTF set, at the same positions as for the SMA. The data was obtained via sweep deconvolution in the range of 20 Hz - 20 kHz, played back by a loudspeaker. The three rooms in which the measurements were performed are

- An entrance hall with dimensions of approximately 8.2 m × 14.0 m × 8.5 m, (WLH), where the receiver was placed facing the loudspeaker at a distance of approximately 5 m between them. Figure 3.2, shows an image of the entrance hall.
- A reverberation chamber with dimensions of approximately 5.3 m × 6.0 m × 3.7 m, (WLH), where the receiver was placed facing the loudspeaker at a distance of approximately 3 m between them. Figure 3.3, shows an image of the reverberation chamber.
- An open office space with dimensions of approximately 5.8 m × 32.0 m × 3.1 m, (WLH), where the receiver was placed facing left to the loudspeaker at a distance of approximately 1.7 m between them. Figure 3.4 shows an image of the open office space.

In Table 3.1, the reverberation times in third octave bands for the three rooms are shown. It can be noted that the reverberation chamber has the longest reverberation times and the open office space has the shortest reverberation times.

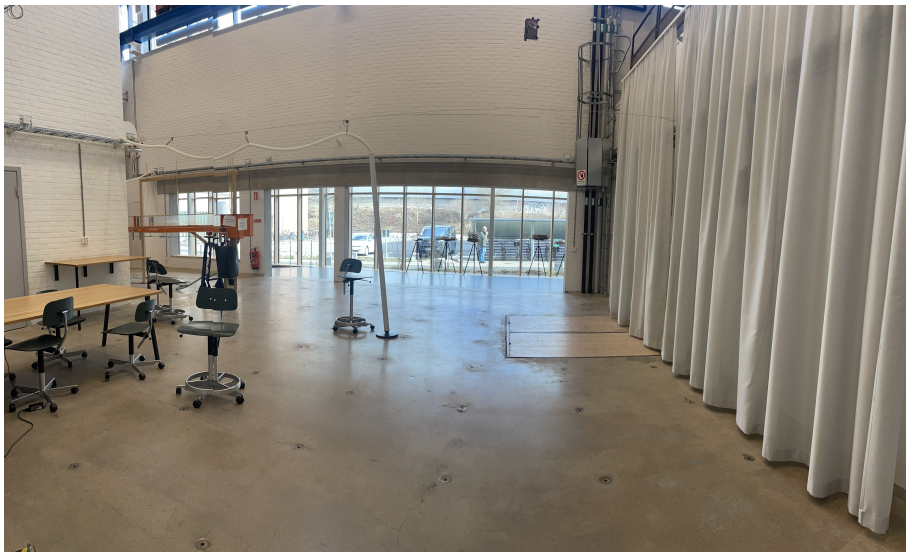

Figure 3.2: The measurement site of the entrance hall.

Figure 3.3: The measurement site of the reverberation chamber.

Figure 3.4: The measurement site of the open office space.

Table 3.1: Reverberation times in third octave bands with center frequencies, $\boldsymbol{F_c}$, in the range of 125 Hz to 16 kHz for the three rooms.

$\boldsymbol{F_c}$ (Hz)	Entrance Hall (s)	Reverberation Chamber (s)	Office Space (s)
125	1.09	2.02	0.67
160	1.04	2.00	0.52
200	1.10	1.43	0.39
250	1.05	1.79	0.37
315	0.98	1.90	0.39
400	1.08	1.94	0.32
500	1.17	1.96	0.31
630	1.16	1.95	0.30
800	0.99	1.96	0.29
1000	1.00	1.89	0.33
1250	1.06	1.91	0.30
1600	0.99	1.77	0.35
2000	0.96	1.61	0.32
2500	0.87	1.52	0.31
3150	0.78	1.26	0.31
4000	0.72	1.13	0.30
5000	0.63	0.87	0.29
6300	0.55	0.73	0.27
8000	0.46	0.62	0.25
10000	0.45	0.52	0.18
12500	0.38	0.45	0.19
16000	0.36	0.34	0.17

4

Evaluation

This chapter, although with the thesis focused on the coherence matching and covariance framework, will include separate results of binaural decoding using the magnitude least-squares solution. The results will therefore include MagLs and covariance framework rendered with first and fourth-order Ambisonics, coherence matching rendered with first-order Ambisonics, and the results for the measured BRIRs as a reference.

4.1 Entrance Hall

The coherence for the entrance hall generated BRIRs, compared to the measured BRIR and HRTF set coherence, are shown in Figure 4.1, with separate illustrations for the total late BRIRs and the late BRIRs, starting approximately 150 ms after the direct/late split. The coherence of both covariance framework BRIRs resembles the measured BRIRs coherence. This is also true for fourth-order MagLs, although slightly higher. However, the coherence for the first-order MagLs and coherence matching do not have the same tendency. Around 3 kHz, the coherence for the first-order MagLs has a peak that is significantly higher than for the measured.

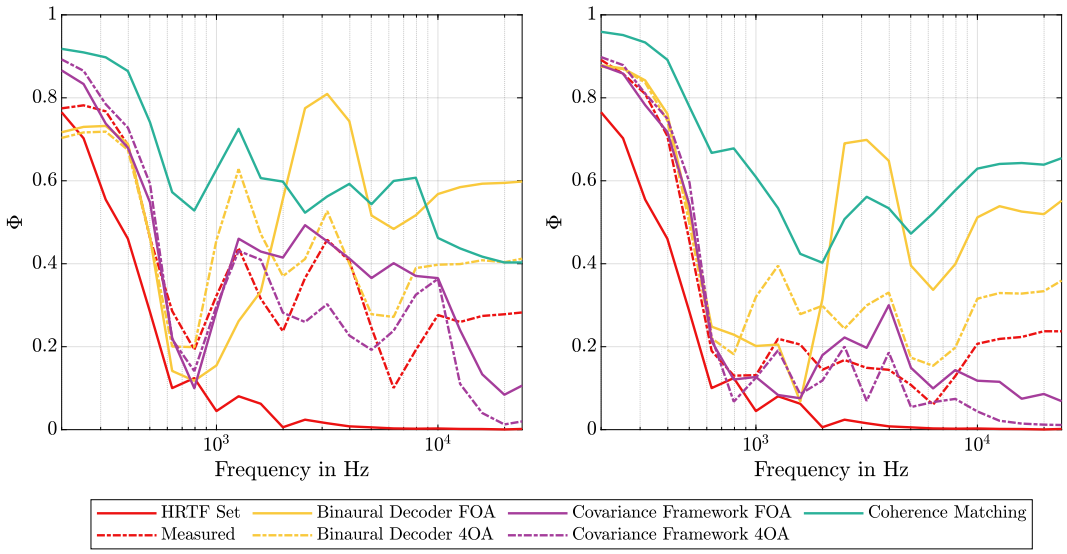

Figure 4.1: Interaural coherence of the entrance hall generated BRIRs, compared to the IC of the measured artificial head and HRTF set. Left: IC of total late BRIRs. Right: IC of late BRIRs, starting approximately 150 ms after the direct/late split.

Even for frequencies above 3 kHz, the coherence is higher. For the coherence matching, the same peaks as for the measured are distinguishable, although higher for the whole frequency spectrum. For the right graph which includes the later part of the BRIR, all curves drop down towards the HRTF coherence. The coherence matching and first-order MagLs coherence, however, do not drop down to the same degree.

To illustrate the deviations of the generated BRIRs coherence curves for the two main methods of the thesis more clearly, Figure 4.2 shows the differences in coherence, $\Delta\Phi$, compared to the measured, where $\Delta\Phi = 0$ indicates complete similarity. For the covariance framework, both first and fourth-order Ambisonics, the curves fluctuate around 0 but show a greater difference toward higher frequencies. The deviation for the coherence matching is clearly higher than for the covariance framework and even increases for the later part of the BRIR, where the covariance framework deviations decrease.

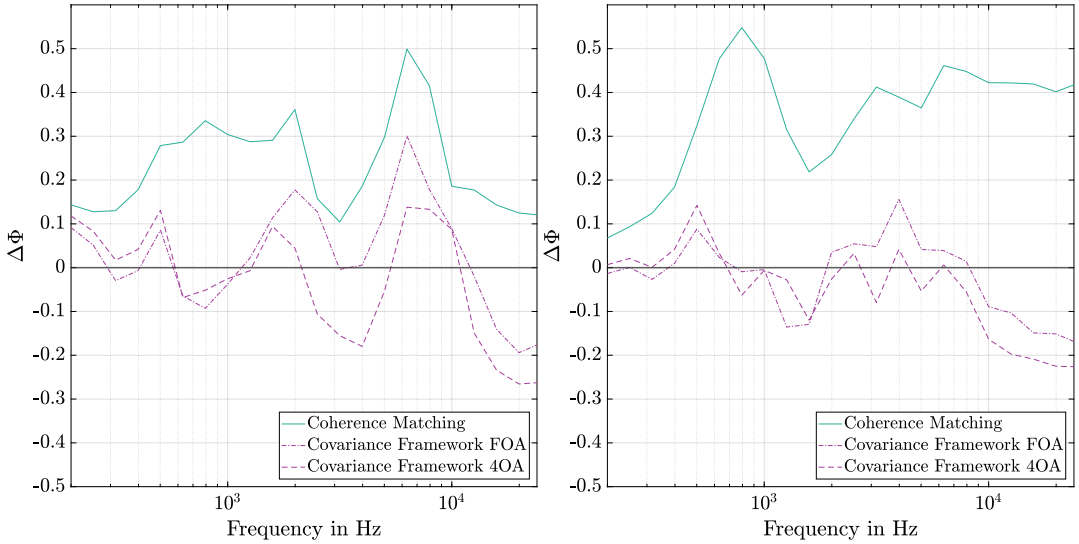

Figure 4.2: Difference in interaural coherence, $\Delta\Phi$, for the covariance framework and the coherence matching, compared to the measured artificial head in the entrance hall. Left: Total late BRIRs. Right: Late BRIRs, starting approximately 150 ms after the direct/late split.

Figures 4.3-4.6, show the frequency spectra for the measured artificial head and generated BRIRs in the entrance hall, including separate illustrations for the total and late parts. For the MagLs generated BRIRs in Figure 4.3, the curves fit fairly well to the measured for both the total and late spectra. However, for lower frequencies, up to around 700 Hz, there is a noticeable divergence where the generated curves are higher than the measured, though with a marginal improvement from first to fourth-order. The coherence matching spectra in Figure 4.4 show more of a deviation than for MagLs for both total and late BRIR but follow the general shape of the measured spectra. Towards 2 kHz and up, the fit improves, however. This is comparable to the first-order covariance framework spectra in Figure 4.5, whereas the fourth-order covariance framework spectra in Figure 4.6 show a closer fit to the

measured compared to the other BRIRs. There is a general improvement for all the BRIRs when removing the direct part and looking at the total late BRIRs, as well as an improvement when removing the first 150 ms.

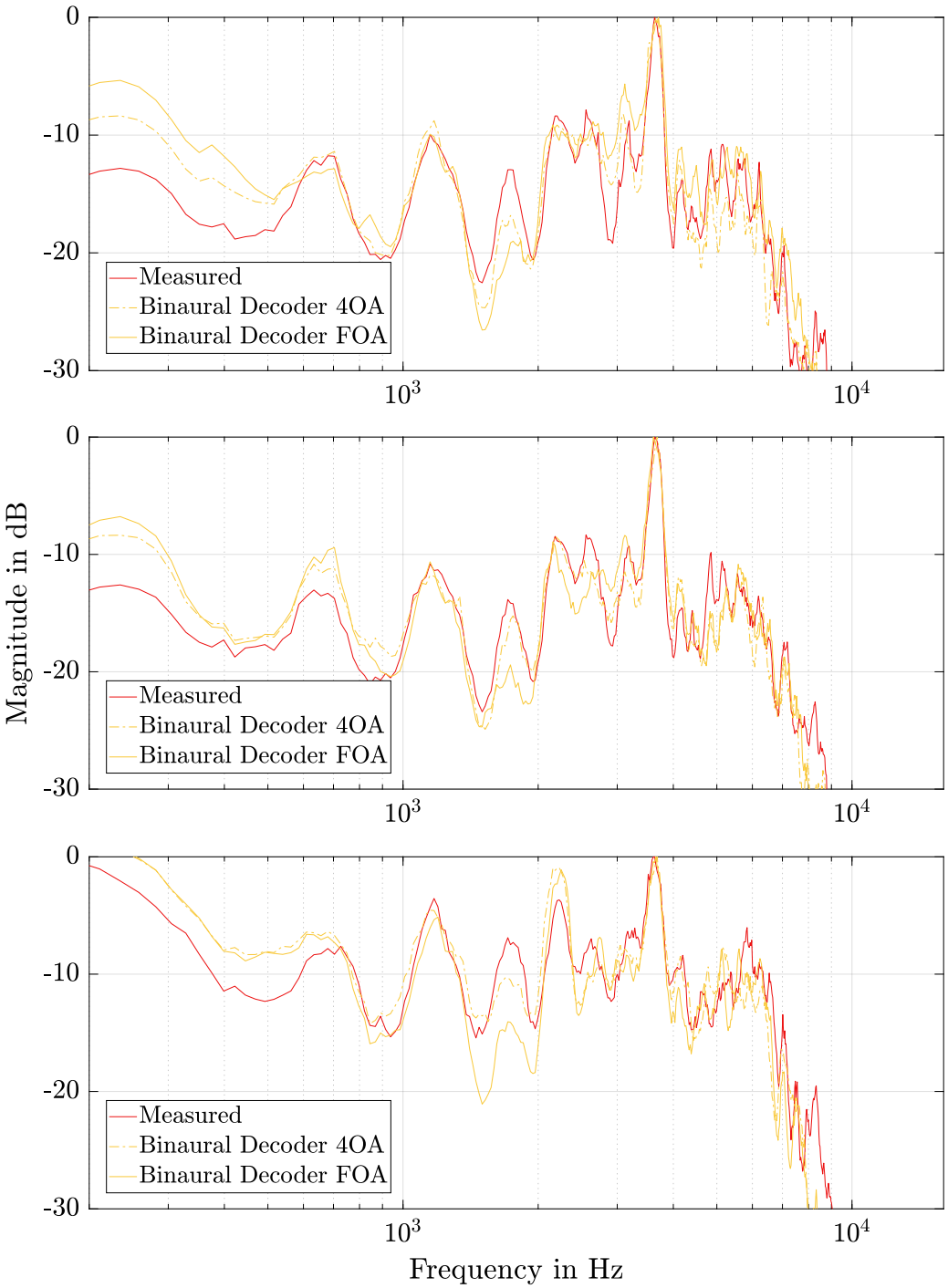

Figure 4.3: Frequency spectra of the generated left BRIRs in the entrance hall for first- and fourth-order MagLs, compared to the left BRIR measured with the artificial head. Top: Total BRIRs. Middle: Total late BRIRs. Bottom: Late BRIRs, starting approximately 150 ms after the direct/late split.

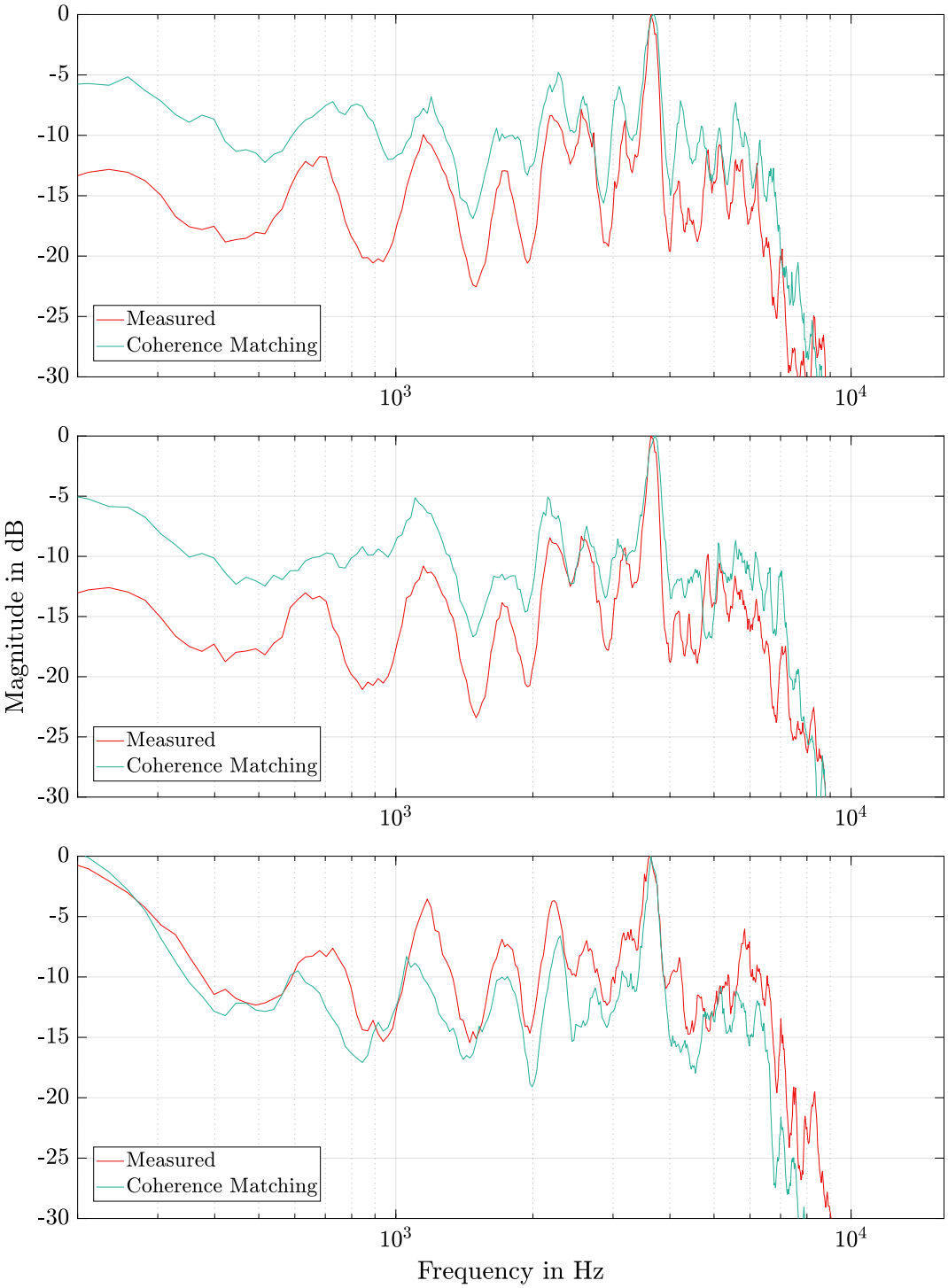

Figure 4.4: Frequency spectra of the generated left BRIRs in the entrance hall for the coherence matching, compared to the left BRIR measured with the artificial head. Top: Total BRIRs. Middle: Total late BRIRs. Bottom: Late BRIRs, starting approximately 150 ms after the direct/late split.

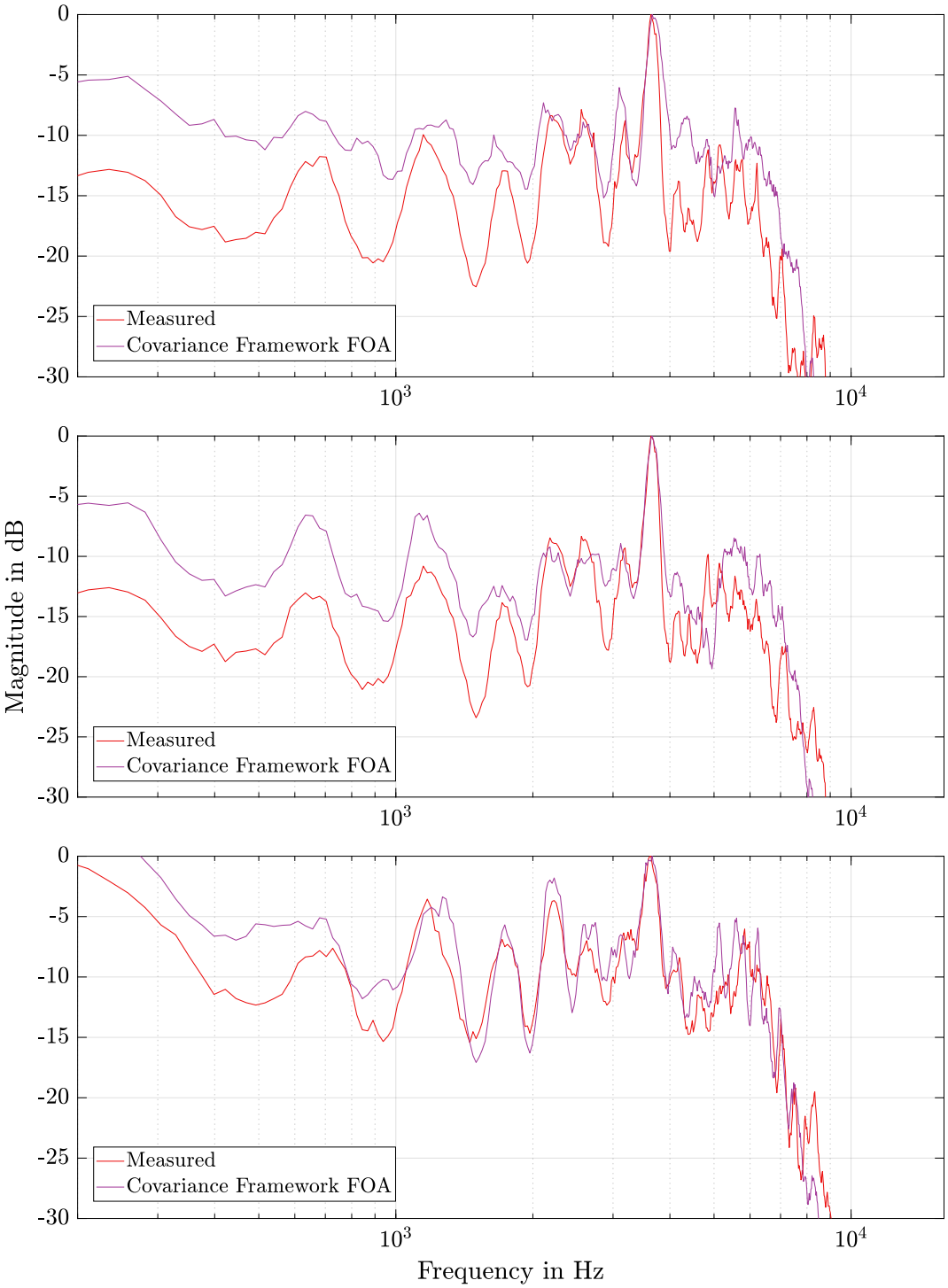

Figure 4.5: Frequency spectra of the generated left BRIRs in the entrance hall for the first-order covariance framework, compared to the left BRIR measured with the artificial head. Top: Total BRIRs. Middle: Total late BRIRs. Bottom: Late BRIRs, starting approximately 150 ms after the direct/late split.

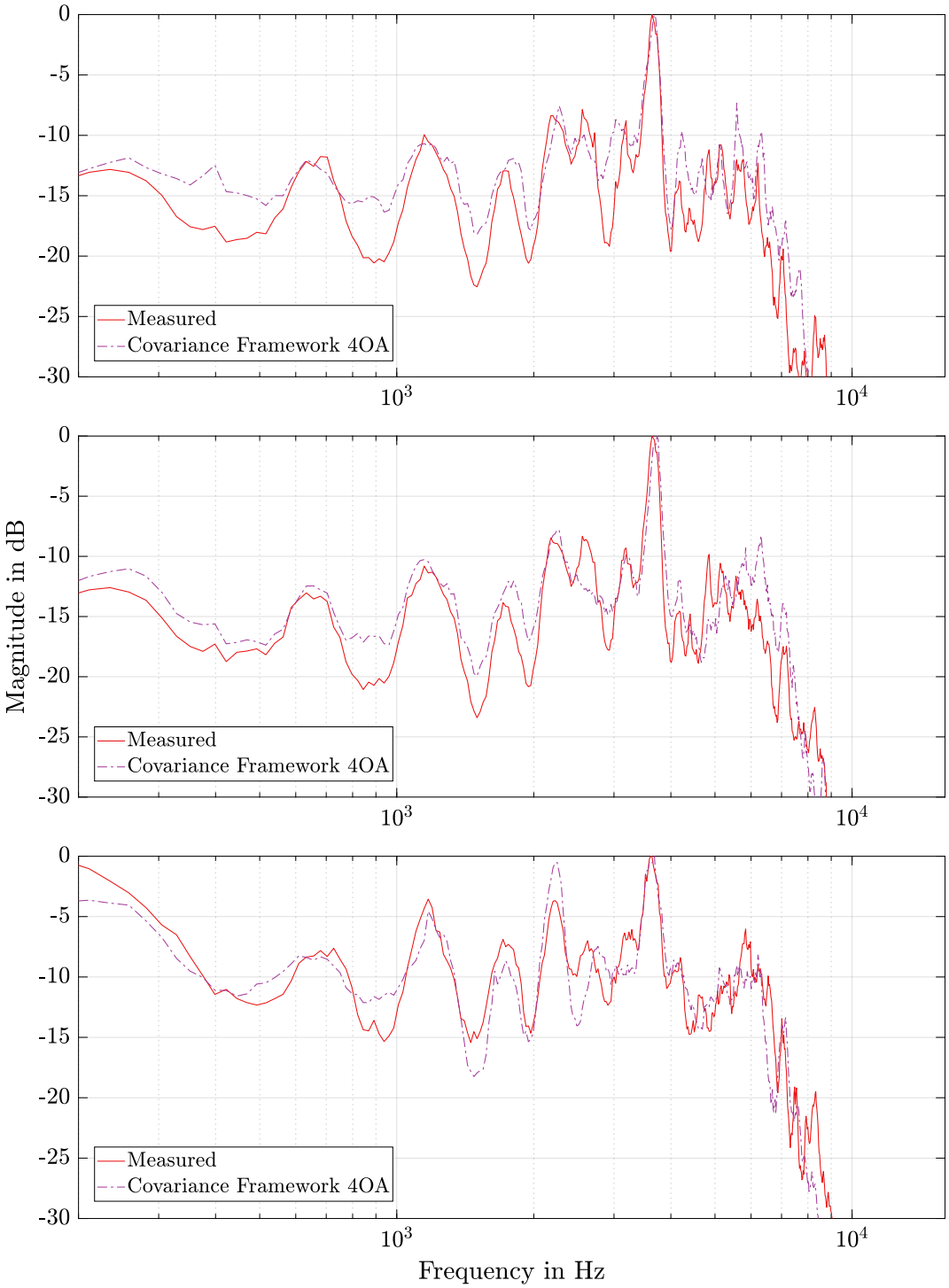

Figure 4.6: Frequency spectra of the generated left BRIRs in the entrance hall for the fourth-order, compared to the left BRIR measured with the artificial head. Top: Total BRIRs. Middle: Total late BRIRs. Bottom: Late BRIRs, starting approximately 150 ms after the direct/late split.

4.2 Reverberation Chamber

The coherence for the reverberation chamber generated BRIRs, compared to the measured BRIR and HRTF set coherence, are shown in Figure 4.7, with separate illustrations for the total late BRIRs and the late BRIRs, starting approximately 150 ms after the direct/late split. Here, the same tendencies as for the entrance hall coherence curves in Figure 4.7 can be observed, where both covariance framework and fourth-order MagLs BRIRs are similar to the measured. The peak slightly above 1 kHz for fourth-order MagLs is, however, notably higher. Although not as high as for the coherence matching, which again reaches far above the coherence of the measured BRIR for the whole frequency spectrum in both graphs. The 3 kHz peak for first-order MagLs is also present here.

Worth noting is that there is no considerable general decrease in coherence for the later BRIR for any of the curves, which is further apparent by looking at Figure 4.8, where the differences in coherence between the measured BRIR and the coherence matching and both covariance framework are shown. Here, the coherence matching experience a considerable difference which stays quite flat over the whole spectra, with a slight decrease for the later BRIR. Both covariance framework curves fluctuate around 0 indicating a close fit, with the fourth-order being the closest.

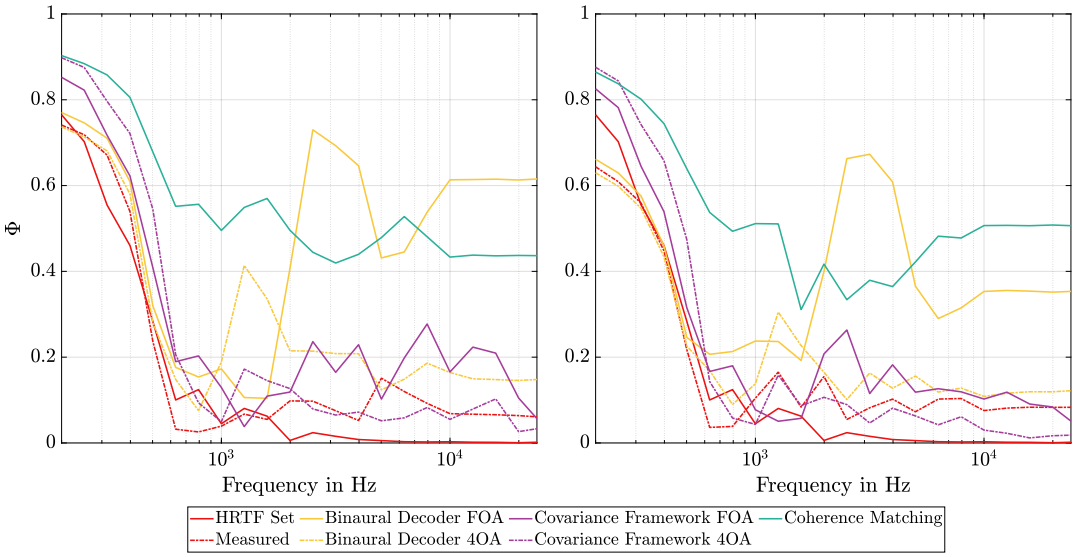

Figure 4.7: Interaural coherence of the reverberation chamber generated BRIRs, compared to the IC of the measured artificial head and HRTF set. Left: IC of total late BRIRs. Right: IC of late BRIRs, starting approximately 150 ms after the direct/late split.

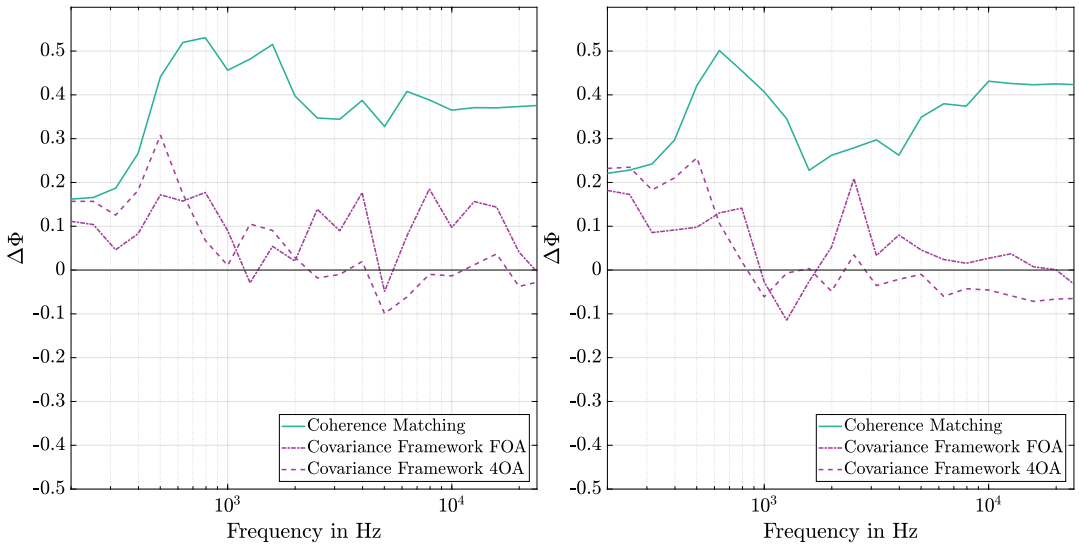

Figure 4.8: Difference in interaural coherence, $\Delta\Phi$, for the covariance framework and the coherence matching, compared to the measured artificial head in the reverberation chamber. Left: Total late BRIRs. Right: Late BRIRs, starting approximately 150 ms after the direct/late split.

The frequency spectra for the measured and generated BRIRs are shown in Figures 4.9-4.12, for both total and late parts. All generated BRIRs spectra fit quite well with the measured spectra. As for the entrance hall BRIRs, the lower frequencies experience a divergence from the measured BRIR. The covariance fourth-order covariance framework, however, is a better fit even for the lower frequencies. Between the total, late and 150 ms later BRIRs, there is no significant improvement for any of the methods. An additional similarity for the reverberation chamber spectra is that all methods experience a roll-off towards higher frequencies, which seem to be most prominent when looking at the total BRIRs.

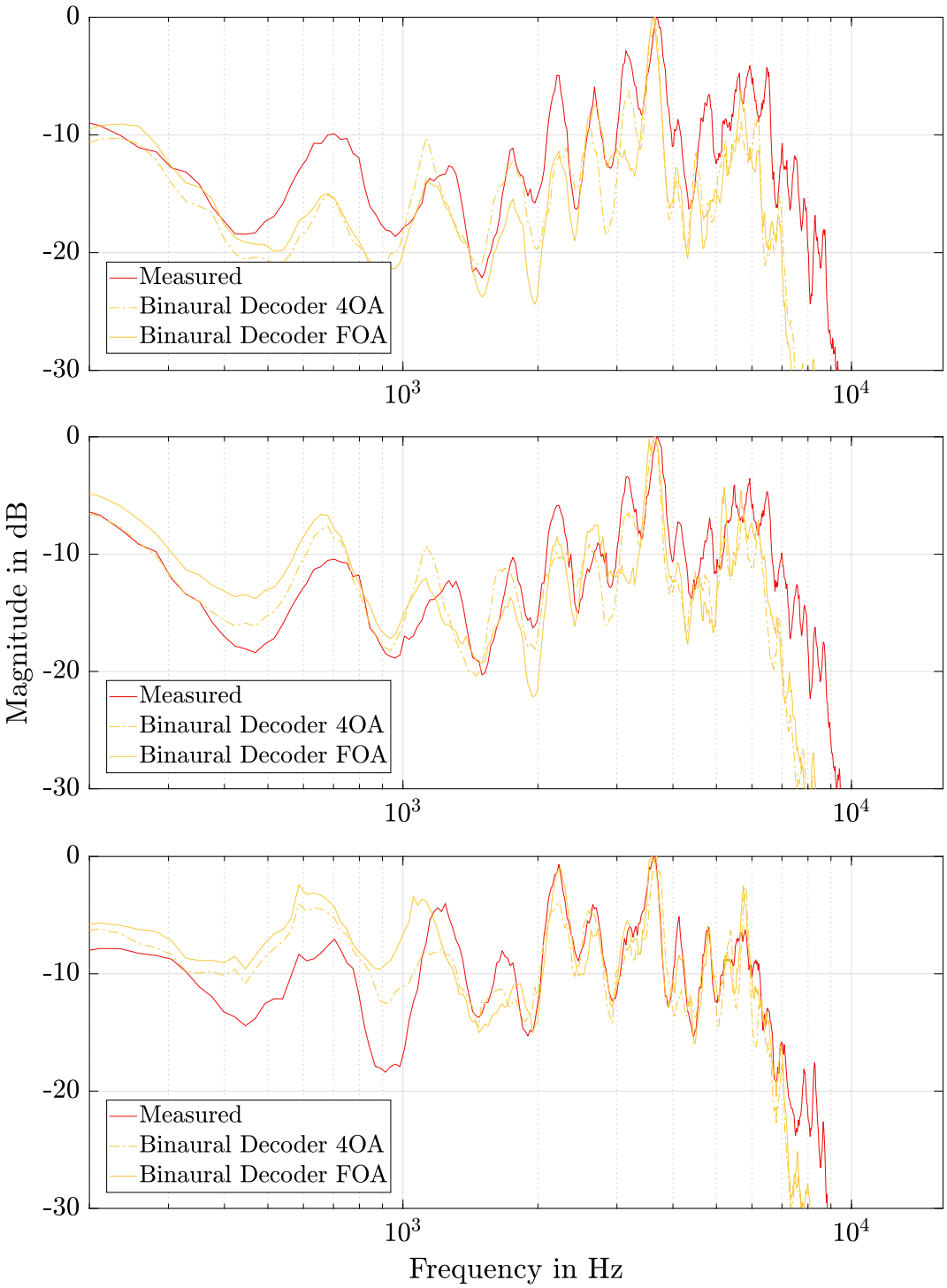

Figure 4.9: Frequency spectra of the generated left BRIRs in the reverberation chamber for first- and fourth-order MagLs, compared to the left BRIR measured with the artificial head. Top: Total BRIRs. Middle: Total late BRIRs. Bottom: Late BRIRs, starting approximately 150 ms after the direct/late split.

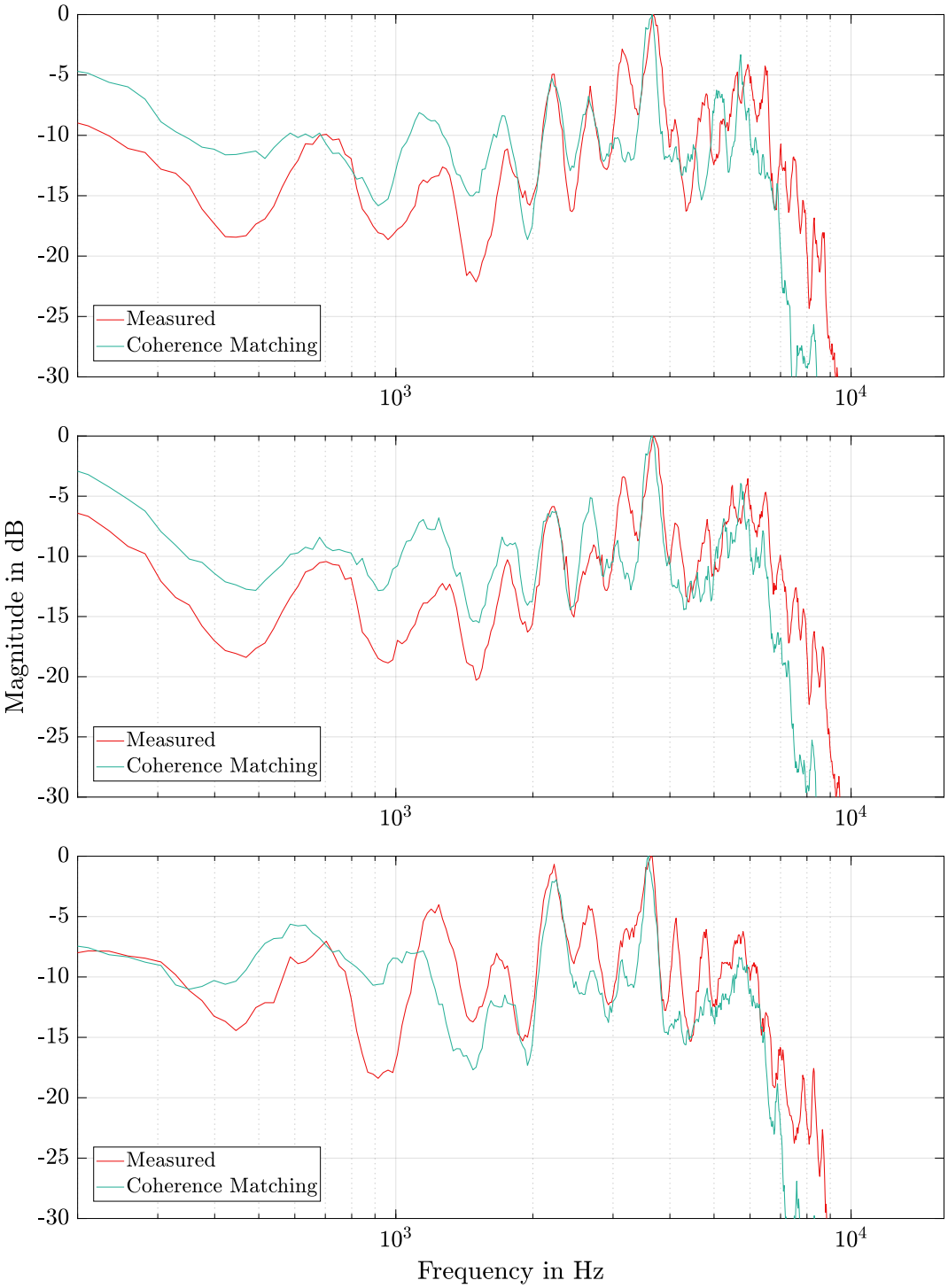

Figure 4.10: Frequency spectra of the generated left BRIRs in the reverberation chamber for the coherence matching, compared to the left BRIR measured with the artificial head. Top: Total BRIRs. Middle: Total late BRIRs. Bottom: Late BRIRs, starting approximately 150 ms after the direct/late split.

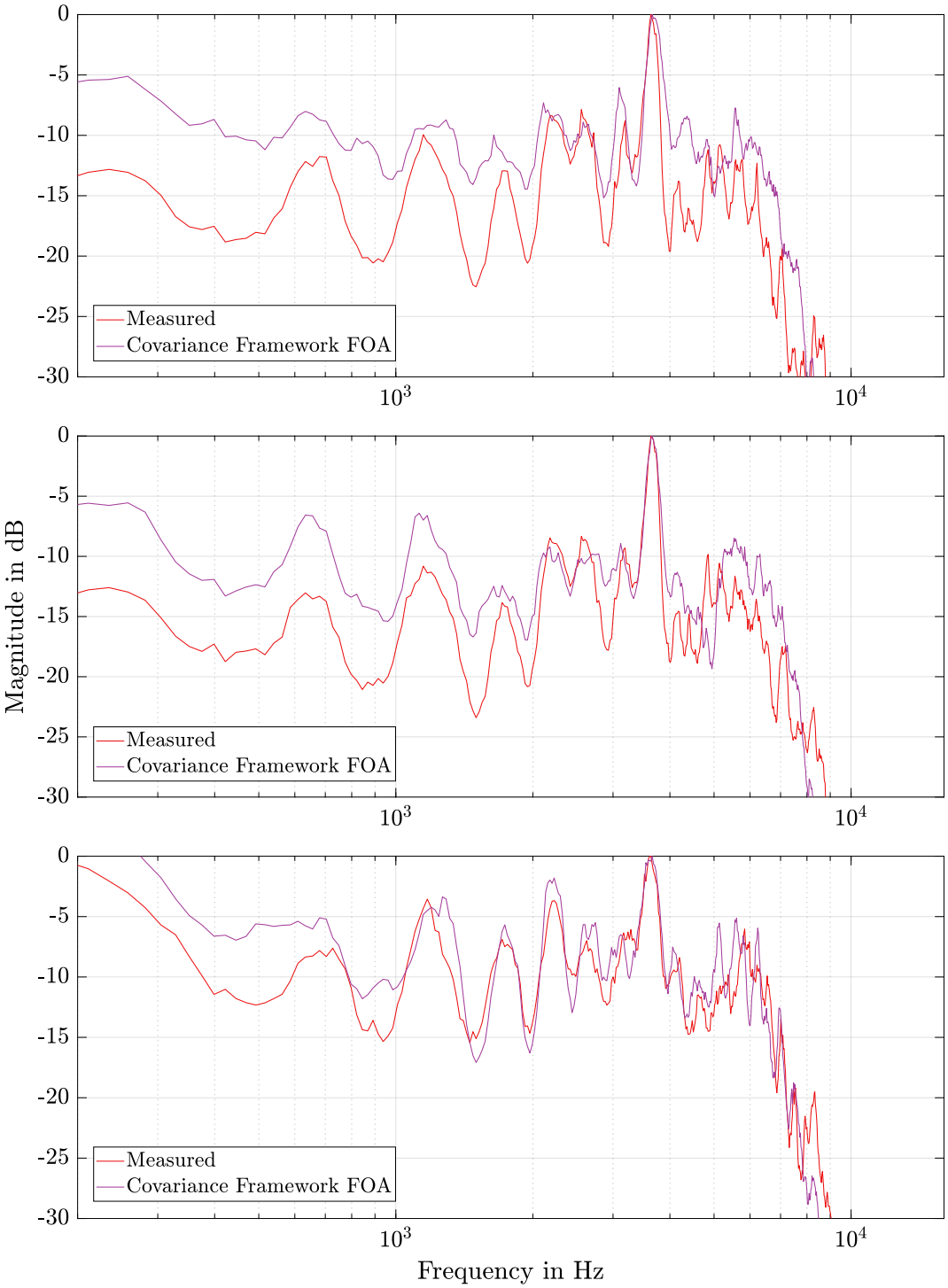

Figure 4.11: Frequency spectra of the generated left BRIRs in the reverberation chamber for the first-order covariance framework, compared to the left BRIR measured with the artificial head. Top: Total BRIRs. Middle: Total late BRIRs. Bottom: Late BRIRs, starting approximately 150 ms after the direct/late split.

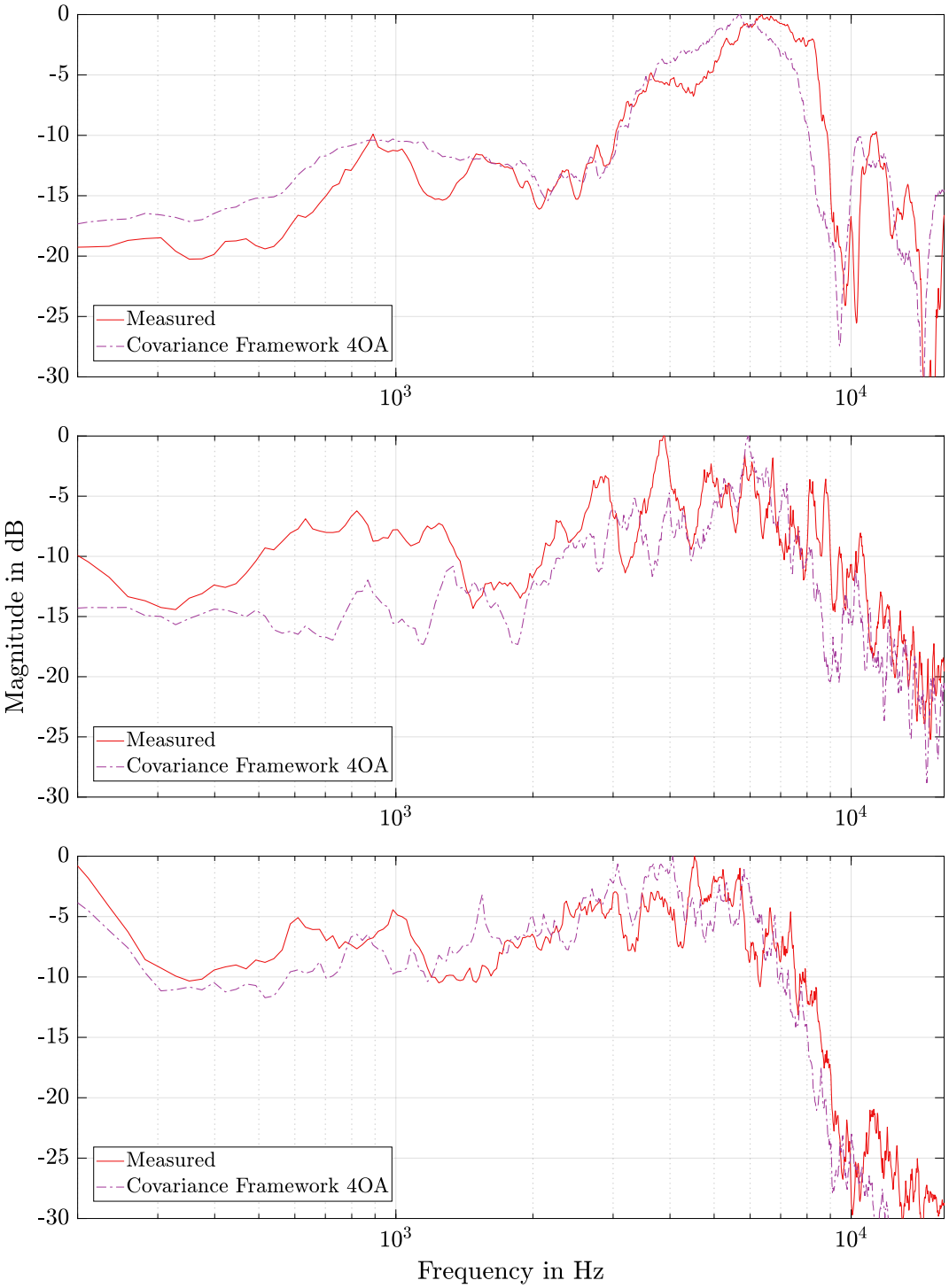

Figure 4.12: Frequency spectra of the generated left BRIRs in the reverberation chamber for the fourth-order, compared to the left BRIR measured with the artificial head. Top: Total BRIRs. Middle: Total late BRIRs. Bottom: Late BRIRs, starting approximately 150 ms after the direct/late split.

4.3 Open Office Space

Figure 4.13 shows the coherence for the BRIRs measured and generated from the open office space recordings, as well as the HRTF set coherence, with separate illustrations for the total late BRIRs and the late BRIRs, starting approximately 150 ms after the direct/late split. The first-order MagLs coherence is similar to the two previous rooms, where a prominent peak occurs at 3 kHz. For higher frequencies, however, it is closer to the measured, especially for the later part in the right graph. The coherence matching coherence also resembles the curves for the entrance hall and reverberation chamber being significantly higher than the measured, especially for the late reverberant part. This is visible from looking at Figure 4.14, where the deviations for the coherence matching can be seen to reach towards 0 for several parts over the spectrum for the total late BRIR, whereas, for the later reverberant part, these dips are not as prominent.

Fourth-order MagLs and both covariance framework curves in Figure 4.13 are closer to the measured coherence for the total late BRIR, where the covariance framework curves adjust to the measured coherence to a higher degree for the late reverberant part. Here, it can be noticed that for lower frequencies, the covariance framework curves are a closer fit to the HRTF set coherence than that of the measured BRIR. Looking at Figure 4.14, it is clear that for lower frequencies, both covariance framework curves deviate from the measured, but stay close to 0 after about 1 kHz.

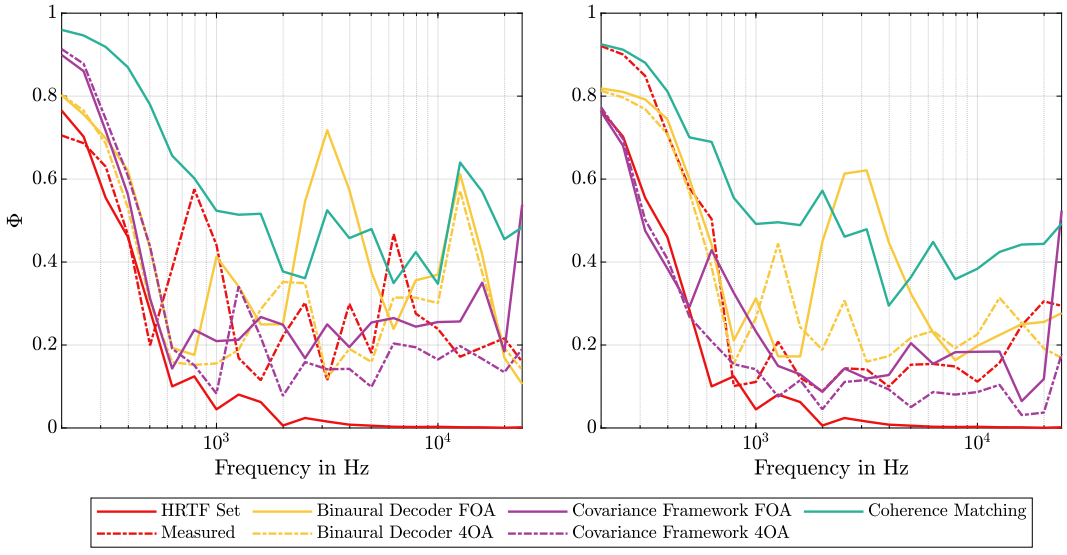

Figure 4.13: Interaural coherence of the open office space generated BRIRs, compared to the IC of the measured artificial head and HRTF set. Left: IC of total late BRIRs. Right: IC of late BRIRs, starting approximately 150 ms after the direct/late split.

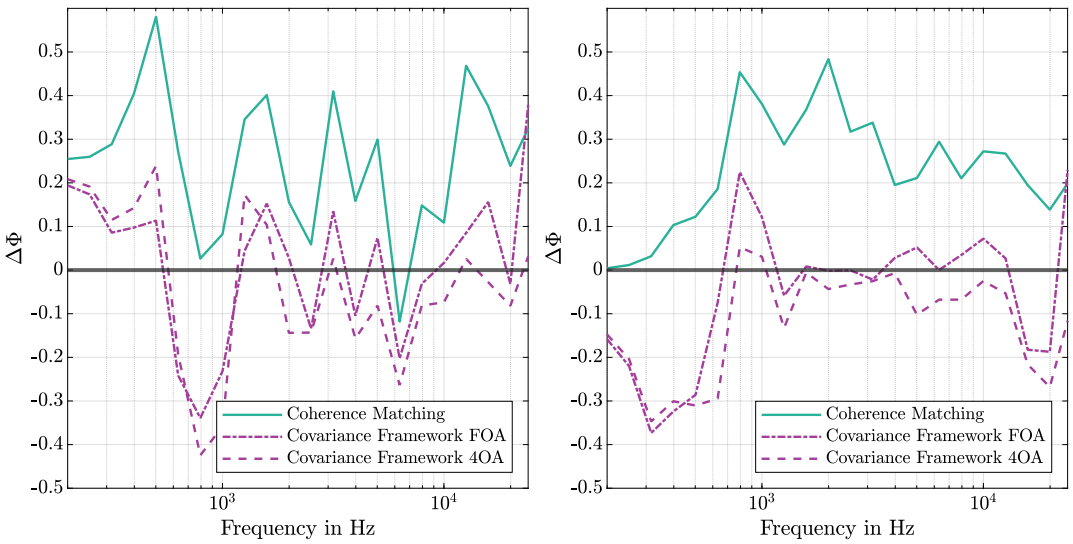

Figure 4.14: Difference in interaural coherence, $\Delta\Phi$, for the covariance framework and the coherence matching, compared to the measured artificial head in the open office space. Left: Total late BRIRs. Right: Late BRIRs, starting approximately 150 ms after the direct/late split.

The frequency spectra in Figures 4.15-4.18 for the generated BRIRs show that there are considerable deviations from the spectra of the measured artificial head, for both the total, late, and later BRIRs. The spectra for the total BRIR first-order MagLs in Figure 4.15 deviate from the measured over the whole spectrum but follow the general shape of the peaks, which improves for the late and later parts. Here, fourth-order MagLs deviate for lower frequencies but adjusts to a higher degree above around 1 kHz. For both order MagLs, there is a roll-off that occurs earlier than for the measured. The coherence matching and both order covariance framework in Figure 4.16, Figure 4.17 and Figure 4.18, respectively, are comparable to each other, with a generally better fit, compared to MagLs. For the later reverberant part, the fourth-order covariance framework seems to experience more of an improvement than the other BRIRs.

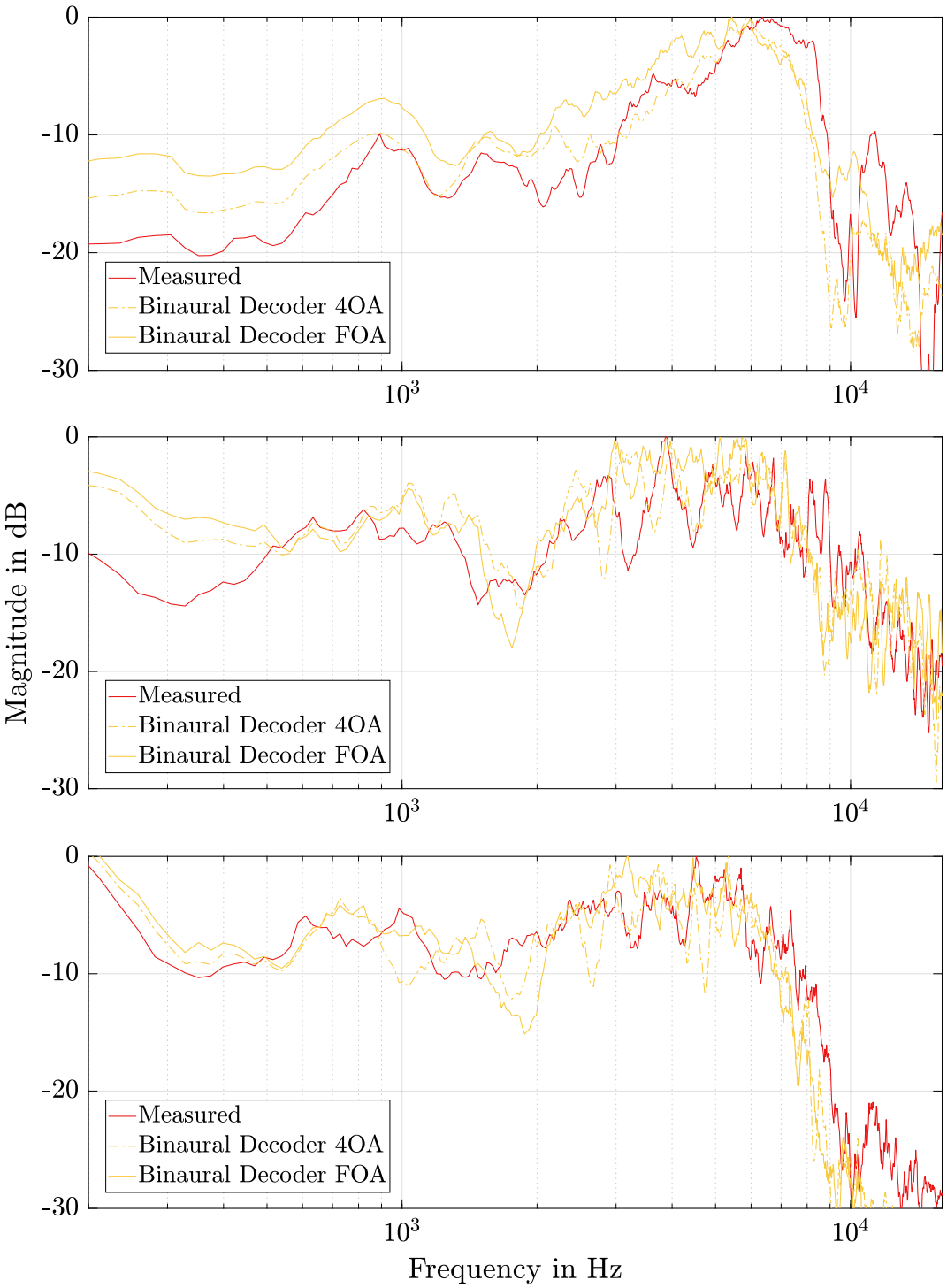

Figure 4.15: Frequency spectra of the generated left BRIRs in the open office space for first- and fourth-order MagLs, compared to the left BRIR measured with the artificial head. Top: Total BRIRs. Middle: Total late BRIRs. Bottom: Late BRIRs, starting approximately 150 ms after the direct/late split.

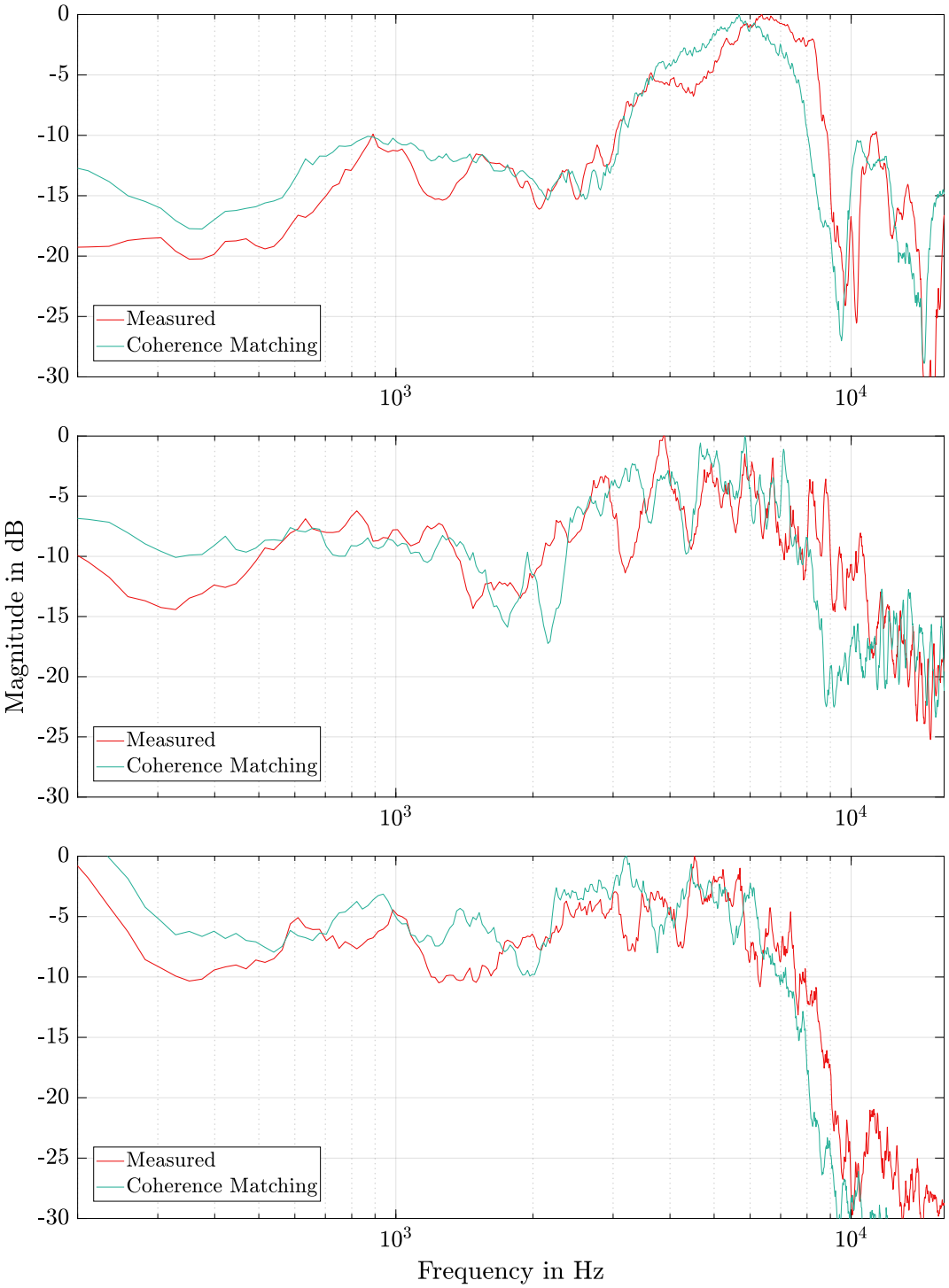

Figure 4.16: Frequency spectra of the generated left BRIRs in the open office space for the coherence matching, compared to the left BRIR measured with the artificial head. Top: Total BRIRs. Middle: Total late BRIRs. Bottom: Late BRIRs, starting approximately 150 ms after the direct/late split.

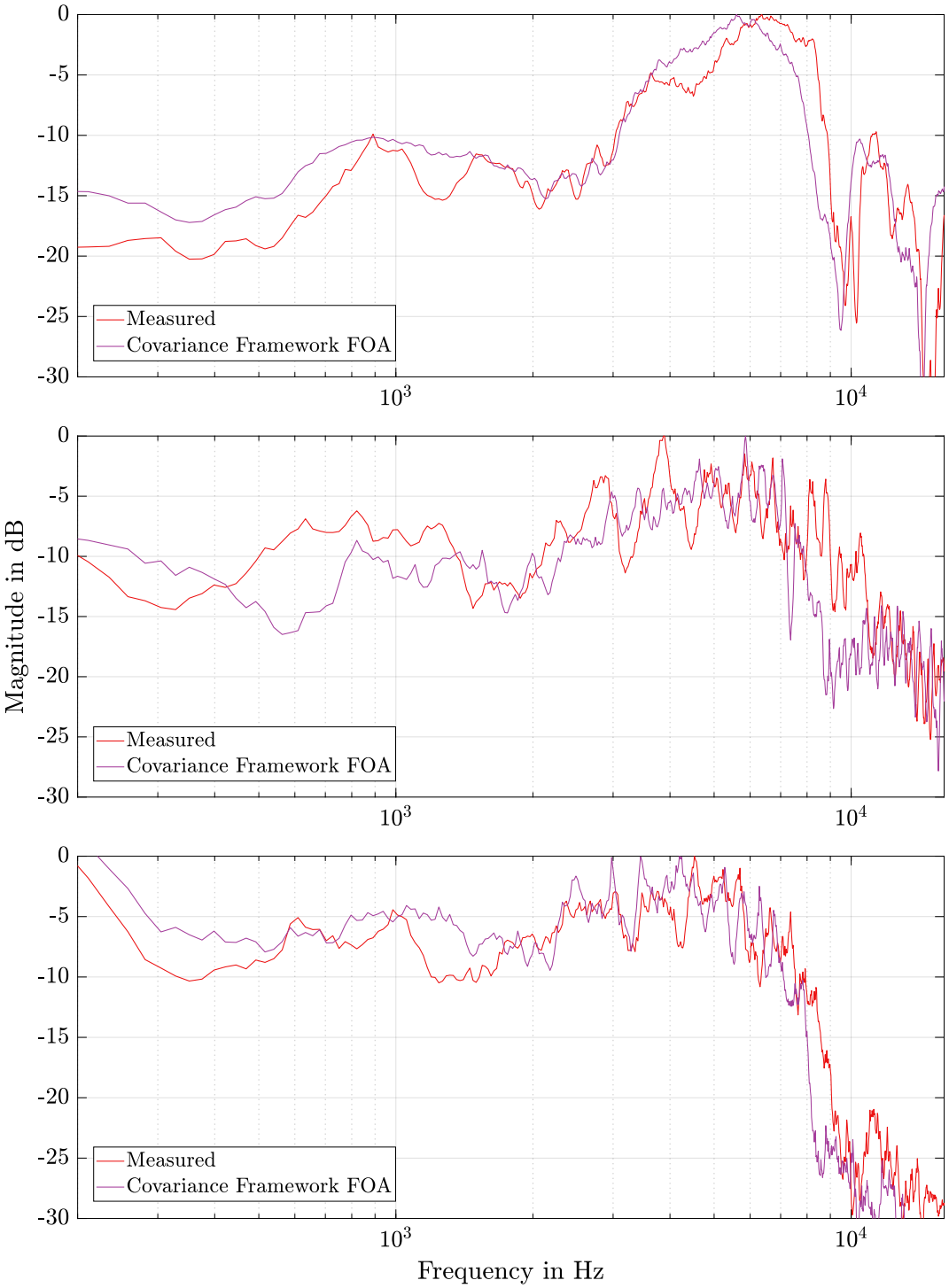

Figure 4.17: Frequency spectra of the generated left BRIRs in the open office space for the first-order covariance framework, compared to the left BRIR measured with the artificial head. Top: Total BRIRs. Middle: Total late BRIRs. Bottom: Late BRIRs, starting approximately 150 ms after the direct/late split.

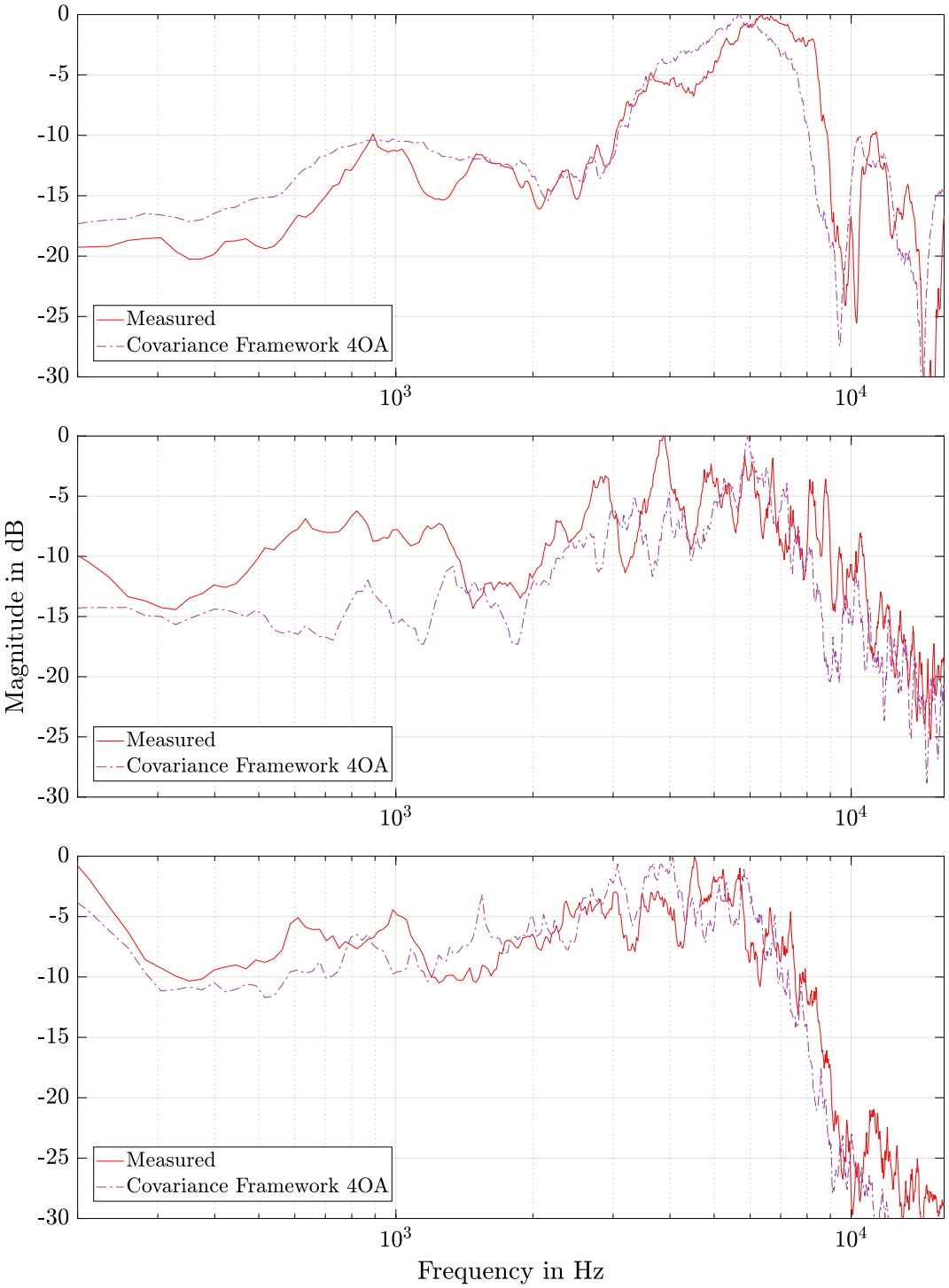

Figure 4.18: Frequency spectra of the generated left BRIRs in the open office space for the fourth-order, compared to the left BRIR measured with the artificial head. Top: Total BRIRs. Middle: Total late BRIRs. Bottom: Late BRIRs, starting approximately 150 ms after the direct/late split.

4.4 Pilot Study

A pilot study was conducted to evaluate the similarity between the artificial head BRIRs and the different generated BRIRs, using a drum-loop and a speech signal in an informal listening test. The listening test was conducted according to the general MUSHRA (Multi Stimulus Test with Hidden Reference and Anchor) guidelines where a listener determines similarity of a stimuli compared to a reference, and score the similarity on a sliding scale between "*very different*", "*different*", "*similar*", "*very similar*" and "*identical*". The comparable stimuli includes a hidden reference to validate the results and is expected to obtain the highest rating. No anchor was included in this listening test, however. The listening test consisted of twelve steps, six drum-loop steps and six speech steps, i.e, two iterations per room and sound, where the order of the steps were randomized for each participant. For each step, 6 stimuli were compared which included the hidden reference, MagLs (first- and fourth-order), covariance framework (first- and fourth-order) and coherence matching.

The test subjects for the pilot study were aged 26-44, four male and one female. One subject was excluded following that they rated the hidden reference with a score less than 90 for more than 15 % of the steps [15]. Hence, the results are presented for four expert listeners.

In Figures 4.19, 4.20 and 4.21, the results from the informal listening tests are shown. For all rooms, it can be observed that the hidden reference was specified as identical to the measured BRIR by all participants.

For the entrance hall BRIRs in Figure 4.19, it is clearly shown that the fourth-order covariance framework was scored as closest to the reference, both for speech and drums. Worth noting is that these stimuli also have the shortest span of individual scores, with the lowest being between *similar* and *very similar* for drums, and not scoring below *similar* for speech. Here, the answers from the separate participants were highly consistent, although slightly less for the drums. For the other BRIRs, the consistency was varied, where some sets of participant answers were quite contrasting. One participant, for example, scored the coherence matching drum stimuli as being both *very different* and *similar*. The covariance framework first-order has the biggest difference of median and percentile between speech and drums, where individual scores for speech can be found above *very similar* and for drums below *different*. For MagLs, slightly higher scores are given for fourth-order compared to first-order, although the individual scores for fourth-order speech span a greater distance than for first-order. However, first-order MagLs has one individual scoring of *very different*. The coherence matching has the majority of scores around the midpoint, *similar*, with one individual score slightly above *very different*.

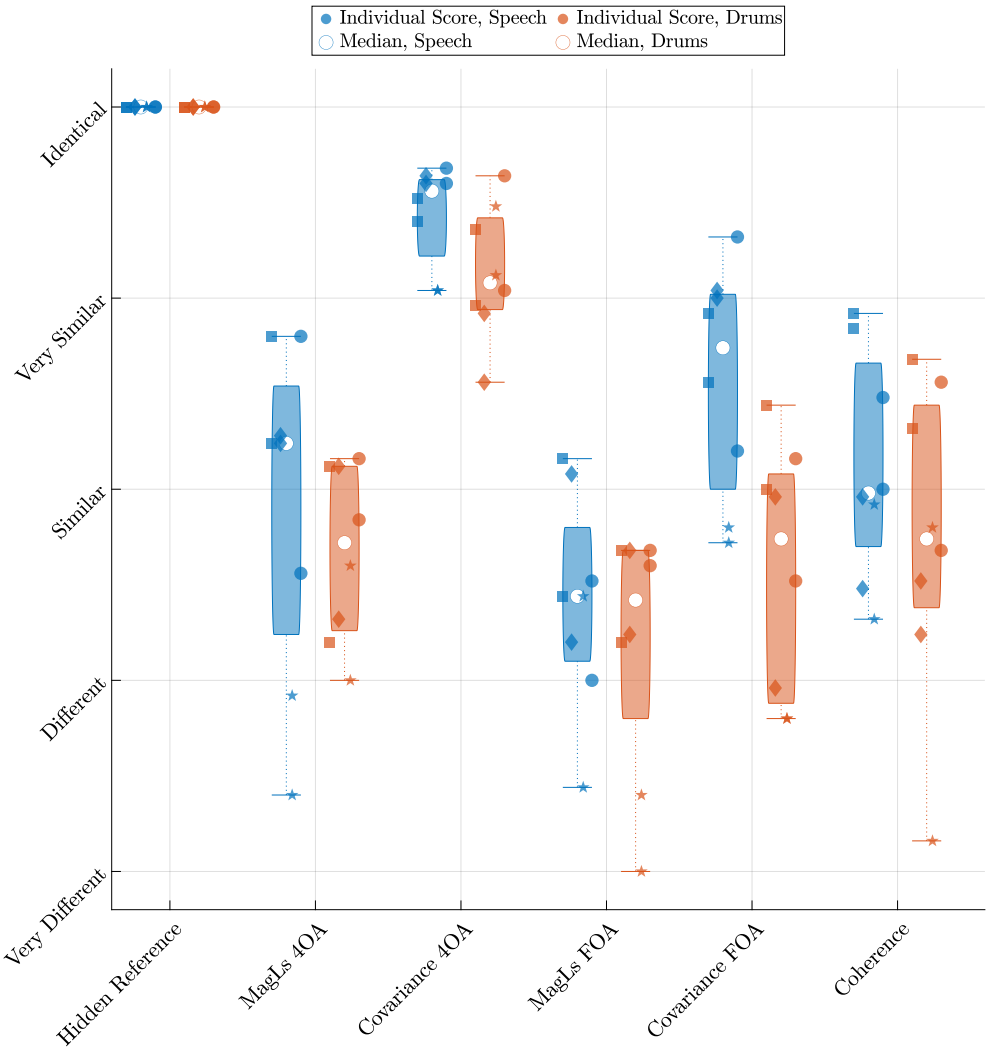

Figure 4.19: Results from the pilot study listening test of the entrance hall BRIRs, with indications of 25$^{\text{th}}$ and 75$^{\text{th}}$ percentiles, median and individual scores. For the individual scores, each symbol indicates answers from a single participant. The whiskers extend to the maximum and minimum values, excluding outliers.

The results from the reverberation chamber listening tests, which are seen in Figure 4.20, coincide quite well with the entrance hall results. The covariance framework first-order has the biggest difference between speech and drums regarding median and percentile values. For speech, the spread of answers is also high, spanning below *different* to above *very similar*. The coherence matching is as previously centered around *similar*, with no scoring falling below *different*. The highest scoring is again given for the fourth-order covariance framework, but reaches further down in scoring, especially for drums. The individual sets of answers also experience a drop in consistency for the fourth-order covariance framework compared to the entrance hall results. Both first and fourth-order MagLs received similar scoring as for the entrance hall, although with higher minimum scores for first-order.

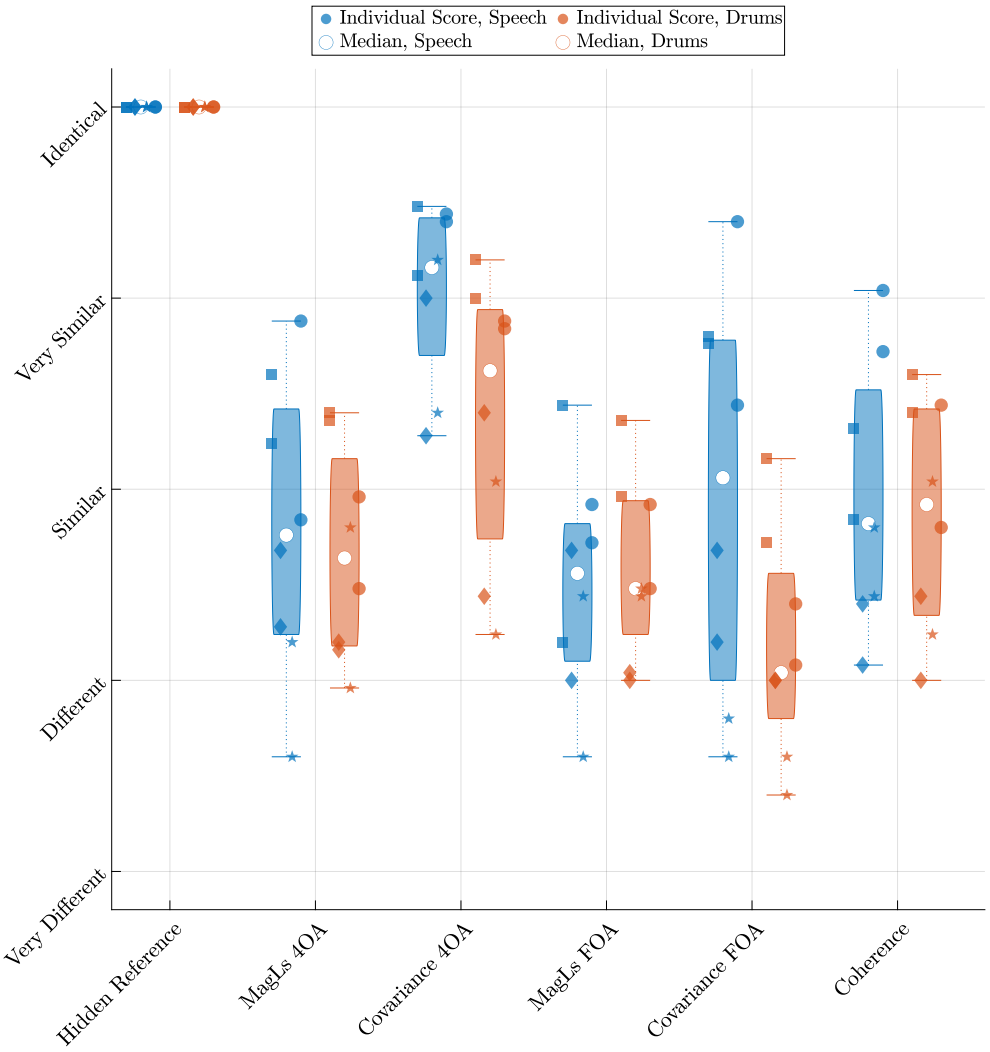

Figure 4.20: Results from the pilot study listening test of the reverberation chamber BRIRs, with indications of 25$^{\text{th}}$ and 75$^{\text{th}}$ percentiles, median and individual scores. For the individual scores, each symbol indicates answers from a single participant. The whiskers extend to the maximum and minimum values, excluding outliers.

In Figure 4.21, where the results from the open office space listening test are shown, it is apparent that these BRIRs were generally ranked the highest out of the three rooms. With the exception of MagLs first-order and one individual score for MagLs fourth-order, all answers fall above *different*, with the majority above *similar* even. MagLs first-order speech is also the one stimulus with the highest spread in answers, even when comparing the results from the other two rooms. For the coherence matching, most of the scoring fall above *similar*. Not only does the covariance framework first-order have higher scoring than the other two rooms for both speech and drums, but the difference between median and percentile values of the speech and drum stimuli are not as prominent as for the other two rooms.

However, the answers from each participant are not as consistent for the first-order covariance drum stimulus. Again, the fourth-order covariance framework is ranked as the most similar compared to the reference, although there are individual answers for the other generated BRIRs that are similar in scoring.

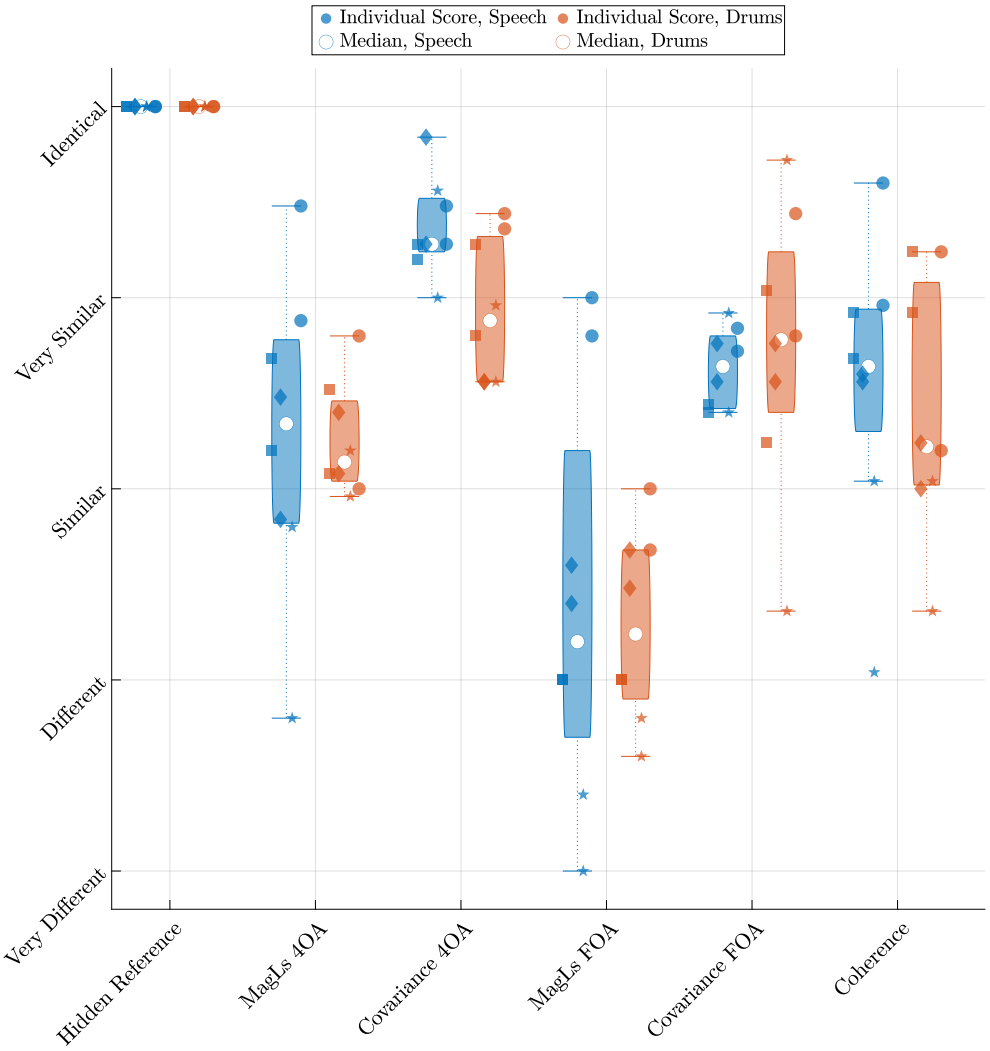

Figure 4.21: Results from the pilot study listening test of the open office space BRIRs, with indications of 25$^{\text{th}}$ and 75$^{\text{th}}$ percentiles, median and individual scores. For the individual scores, each symbol indicates answers from a single participant. The whiskers extend to the maximum and minimum values, excluding outliers.

4.5 Discussion

For all three rooms, there are clear differences in the behavior of the resulting coherence, where the coherence matching and first-order MagLs see the greatest deviations from the coherence of the measured BRIR. The fact that the coherence for most of the BRIRs drops towards the ideal HRTF coherence for the later part is not unexpected. The impulse responses become more diffuse in time, which in turn will lead to the left and right ear becoming more uncorrelated, making the coherence reach towards zero. Since the reverberation chamber is intended to create a diffuse field, it is not surprising that there is no significant change in the coherence between the total late BRIR and the late BRIR moving forward 150 ms. Nor is it surprising that the coherence of the measured BRIR, along with the coherence for three of the generated BRIRs, is close to the ideal HRTF coherence even for the total late part of the reverberation chamber. For the other two rooms, there is a drop in coherence between the late and later parts, since the more prominent direct and early reflection part subsides.

The peak for the MagLs first-order coherence at 3 kHz is prominent for all three rooms, which could be due to artifacts arising from spherical harmonic truncation. That, however, does not explain why the peak is not present for the first-order covariance framework coherence, although it utilizes the first-order MagLs filter. The time-adaptive processing for the covariance framework could be a possible explanation for the absence of these artifacts, despite being first-order. Since the 3 kHz peak does not occur for the fourth-order MagLs, it is clear that both the spherical harmonic order and adaptive implementation of the interaural cues seem to play a part in replicating the interaural coherence of the measured BRIR.

Considering that both the coherence matching and covariance framework use similar approaches and that both methods are dependent on the ideal response of the interaural cues of the HRTF set to a varying degree, it would be expected that the interaural coherence for the coherence matching method would be closer to that of the measured and more aligned with the first-order covariance framework coherence. The fact that this is not the case suggests that the process of adapting the magnitude of the HRTF set interaural coherence to that of the input signals, as in the covariance framework method as opposed to utilizing the ideal interaural coherence of the HRTF set as for the coherence matching, is an important element to obtain a similar coherence to that of the measured BRIR. The covariance framework not only adapts the output signal to the input signal, but it includes the ability to fine-tune the degree of regularization and adaptation to the input signal, contrary to the coherence matching method which has the same processing regardless of the input signal. It is possible to adjust the parameters of the STFT, and there is a choice of HRTF set, but for the comparison, these were the same for both methods.

Spectral roll-off for higher frequencies, which is especially prominent for the reverberation chamber, could be explained by the off-center position of the head in the Ambisonic coordinate system which would affect the timbre of the signals. This concurs with the listening test results for the fourth-order covariance framework where the entrance hall spectrum is a better fit for higher frequencies than for the reverberation chamber. This could thus explain why the entrance hall fourth-order

covariance framework is ranked as more similar to the measured with less spread in answers. This could also explain why there generally is a greater spread in answers for the reverberation chamber evaluations than for the entrance hall.

Seeing that the frequency response for higher frequencies is a worse fit to the measured response, one could expect that the drum signal which has more high-frequency content would be perceived as less similar to the speech signal. This does, however, not seem to be the general case for the different BRIRs, except for the covariance framework both first- and fourth-order where the drum signal was ranked as being both less similar and with a generally wider spread for all three rooms compared to the speech signal. I.e., from the pilot study results, the frequency content of the stimuli used does not seem to be of very high relevance when it comes to perceiving the generated BRIRs as close to the measured.

Since the spatial complexity increases with frequency, and that the spatial resolution for first-order spherical harmonics is worse than for fourth-order, the spectral roll-off for the first-order BRIRs would be expected to be greater. In reference to the spatial resolution of a head-related impulse response, a fourth-order BRIR is still not a high-order representation of a sound field, and clearly not a high enough order to avoid eventual spectral roll-off that the first-order BRIRs experience.

For the BRIRs that are modeled using first-order spherical harmonics, the sound image is perceived as having less externalization when compared to the measured BRIR. This is not unexpected since the first-order spherical harmonics comprise broader directivity patterns and will therefore not be able to distinguish the information to the same extent as higher-order spherical harmonics. Hence, the first-order BRIRs would be expected to have fewer localization cues and less sense of spaciousness, which seem to be in line with the listening test results when comparing the individual listening test answers for different orders for the same method. However, comparing the answers for different orders between methods, there is not a clear improvement. For example, first-order coherence matching and covariance framework were generally ranked higher than MagLs fourth-order for all rooms, which shows that the order of the BRIR is not solely responsible for the similarity to the corresponding HRIR.

Since the direct part is modeled the same way for both first and fourth-order BRIRs, the perceived more separated sound image for the fourth-order BRIRs is most likely stemming from the fact that the higher-order spherical harmonic is more accurate, which will lead to better separation of the early reflections that are included in the processing of the BRIRs reverberant part. Comparing the stimuli of the coherence matching to those of the first-order MagLs, the timbre is similar. The similarity to the directional information of the stimuli filtered with the measured BRIR, however, is closer for the coherence matching. This is again most likely due to the processing of the direct part which is separately performed for the coherence matching, while the MagLs filter processes direct and diffuse sound in the same way.

From the pilot study results, it is reasonable to conclude that the fourth-order covariance framework method was perceived as being closest to the artificial head measurements, for all three rooms. However, not with as much certainty for the drum signal, which for both covariance framework methods was generally ranked

lower, or with a larger spread in answers, compared to the speech signal. This could suggest that deviations from the measured signal are more noticeable for transient signals using the covariance framework. In the listening test results from the reverberation chamber, both MagLs and coherence matching show similar responses between speech and drum signals, unlike the responses for the covariance framework methods which highlights that there is a difference in perception depending on stimulus. Another explanation for the difference in perception between the two stimuli is the varying degree of similarity to the measured BRIRs over the frequency spectra. As the fit is worse for lower frequencies, as well as a roll-off for higher frequencies, it is not unexpected that the drum signal, which includes both to a higher degree than the speech signal, is perceived as less similar.

Comparing the listening test results for the different rooms, there seems to be an increase in confidence interval for an increasing reverberation time. This could be explained simply by the content relationship between the direct and diffuse parts of the BRIR since a room with a low reverberation time will be dominated by direct sound. Modeling the direct sound for the two main methods within this thesis is a relatively straightforward process since it is filtered with a single HRTF in a specific direction. This in return makes it easier to replicate a BRIR where the direct sound constitutes the majority of the information, contrary to rooms with longer reverberation times where the diffuse part dominates and requires more complex processing which introduces greater a margin of error.

5
Conclusion

Judging from the results for both the processing and pilot study, it is fair to say that the covariance framework is the method from this thesis work that to the highest degree possesses the ability to replicate the measured BRIR. However, even with the BRIRs generated using fourth-order covariance framework, there are differences when compared to the measured BRIR. With the covariance framework offering the ability to generate BRIRs with the use of any Ambisonic order, there is the possibility of increasing the spherical harmonic order and in turn the spatial resolution. The generated BRIRs using higher-order Ambisonics would most likely experience improvement in similarity. However, measurements for higher orders using spherical microphone arrays could introduce additional problems, as mentioned in the theory section, which might not be possible to counteract using the processing method.

Since the pilot study was performed by only four participants, conclusions based on the listening test responses have to be seen as mere indications rather than clear results. A palpable development for this thesis work would thus be to carry out a more comprehensive listening test that included more participants. In that case, the hope would be that the 25^{th} and 75^{th} percentiles and median of the evaluations would become narrower, thus creating a more distinct result. Similarly, it would be beneficial to conduct additional room measurements of which to generate the BRIRs, so that rooms with both similar and a wider range of parameters would be possible to evaluate. This would make it more discernible what parameters are important to the perceived similarity.

Bibliography

[1] F. Zotter, Wikimedia Commons, Dec. 2013, [Online]. Available: https://commons.wikimedia.org/wiki/File:Spherical_Harmonics_deg3.png

[2] F. Menzer, C. Faller and H. Lissek, "Obtaining Binaural Room Impulse Responses From B-Format Impulse Responses Using Frequency-Dependent Coherence Matching," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 396-405, Feb. 2011. doi: https://10.1109/TASL.2010.2049410.

[3] J. Vilkamo, T. Bäckström, and A. Kuntz, "Optimized Covariance Domain Framework for Time–Frequency Processing of Spatial Audio," in *J. Audio Eng. Soc.*, vol. 61, no. 6, pp. 403-411, June 2013.

[4] M. Zaunschirm, C. Schörkhuber, R. Höldrich, "Binaural rendering of Ambisonic signals by head-related impulse response alignment and a diffuseness constraint," in *J. Acoust. Soc. Am.*, vol. 143, no. 6, pp. 3616-3627, June 2018.

[5] B. Mróz, P. Odya, B. Kostek, "Creating a Remote Choir Performance Recording Based on an Ambisonic Approach," in *Appl. Sci.*, vol. 12, no. 7, pp. 3316, Dec., doi: https://doi.org/10.3390/app12073316

[6] F. Zotter, M. Frank, "Ambisonics, A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality," in *Springer Topics in Signal Processing*, vol. 19, 2019. doi: https://doi.org/10.1007/978-3-030-17207-7

[7] GRAS "45BB-1 KEMAR Head & Torso for Hearing Aid Test, 1-Ch LEMO", [Online]. Available: https://www.grasacoustics.com/products/head-torso-simulators-kemar/kemar-for-hearing-aid-test-1-ch/product/499-45bb-1

[8] RØDE, "NT-SF1 Ambisonic Microphone," [Online]. Available: https://rode.com/en/microphones/360-ambisonic/nt-sf1

[9] mh acoustics, "em32 Eigenmike," [Online]. Available: https://mhacoustics.com/

[10] W. M. Hartmann, B. Rakerd, Z. D. Crawford, P. X. Zhang, "Transaural experiments and a revised duplex theory for the localization of low-frequency tones," in *J. Acoust. Soc. Am*, vol. 139, no. 2, pp. 968-985, 2016. doi: https://doi.org/10.1121/1.4941915

[11] J. S. Bendat and A. G. Piersol. Random data: Analysis and measurement procedures. John Wiley & Sons, New York, 1971.

[12] J. Blauert, W. Lindemann, "Spatial mapping of intracranial auditory events for various degrees of interaural coherence", in *J. Acoust. Soc. Am.*, vol. 79, no. 3, pp. 806–813, 1986. doi: https://doi.org/10.1121/1.393471

[13] H. Braren, J. Fels, "A High-Resolution Head-Related Transfer Function Data Set and 3D-Scan of KEMAR," Institute for Hearing Technology andAcoustics, RWTH Aachen University, 2020. doi: https://doi.org/10.18154/RWTH-2020-11307

[14] T. Deppisch, H. Helmholz, J. Ahrens, "End-to-End Magnitude Least Squares Binaural Rendering of Spherical Microphone Array Signals," in *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, pp. 1-7, 2021. doi: 10.1109/I3DA48870.2021.9610864.

[15] International Telecommunication Union Radiocomunnication Sector, "Recommendation ITU-R BS.1534-3: Method for the subjective assessment of intermediate quality level of audio systems," 2015.

[16] Gustaver, M. (2020) A Chalmers University of Technology Master's thesis template for LaTeX. Unpublished.

DEPARTMENT OF SOME SUBJECT OR TECHNOLOGY
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY