Binaural Auralization of Microphone Array Room Impulse Responses Using Causal Wiener Filtering

Viktor Gunnarsson, *Student Member, IEEE*, and Mikael Sternad, *Senior Member, IEEE*

***Abstract*—Binaural room auralization involves Binaural Room Impulse Responses (BRIRs). Dynamic binaural synthesis (i.e., head-tracked presentation) requires BRIRs for multiple head poses. Artificial heads can be used to measure BRIRs, but BRIR modeling from microphone array room impulse responses (RIRs) is becoming popular since personalized BRIRs can be obtained for any head pose with low extra effort. We present a novel framework for estimating a binaural signal from microphone array signals, using causal Wiener filtering and polynomial matrix formalism. The formulation places no explicit constraints on the geometry of the microphone array and enables directional weighting of the estimation error. A microphone noise model is used for regularization and to balance filter performance and noise gain. A complete procedure for BRIR modeling from microphone array RIRs is also presented, employing the proposed Wiener filtering framework. An application example illustrates the modeling procedure using a 19-channel spherical microphone array. Direct and reflected sound segments are modeled separately. The modeled BRIRs are compared to measured BRIRs and are shown to be waveform-accurate up to at least 1.5 kHz. At higher frequencies, correct statistical properties of diffuse sound field components are aimed for. A listening test indicates small perceptual differences to measured BRIRs. The presented method facilitates fast BRIR data set acquisition for use in dynamic binaural synthesis and is a viable alternative to Ambisonics-based binaural room auralization.**

***Index Terms*—Beamforming, binaural recording, binaural room impulse response (BRIR), head-related transfer function (HRTF), interaural coherence, MIMO, virtual acoustic environment, virtual artificial head (VAH).**

I. Introduction

Auditory experiences are defined by the sound that enters the ear canals. By reproducing the ear signals corresponding to a real or simulated acoustic event using headphones or loudspeakers, the auditory sensation of the original event can be replicated [1]–[3]. This is referred to e.g. as binaural synthesis or creating a Virtual Acoustic Environment (VAE).

The sound pressure at the ears of a listener, in response to a sound source in a specific direction, is described by Head-Related Impulse Responses (HRIRs) or their frequency-domain counterpart, Head-Related Transfer Functions (HRTFs).

Binaural Room Impulse Responses (BRIRs) are used to create a VAE corresponding to listening to loudspeakers in a reverberant acoustic environment, also referred to as auralization of the acoustic environment [4]. A BRIR can loosely be defined as the two impulse responses from a sound source in a room to the two ears of a listener. Typically it contains both direct sound and room reflections. By convolving an audio signal with a BRIR, ear signals are created.

BRIRs can be used, for example, to create VAEs for virtual or augmented reality, to auralize sound systems using headphones for research or commercial product purposes, to auralize recording studios for remote work, to tune sound systems remotely using auralization, or to auralize simulated sound system prototypes in virtual product development.

The traditional method to measure a BRIR is to put microphones in the ears of a real subject or artificial head and measure impulse responses to the ears from a loudspeaker. When using the resulting BRIR for binaural synthesis, the perceived virtual sound source direction is tied to the frame of reference defined by the listener's head. Thus when the listener rotates the head, the perceived physical sound source location changes.

Natural listening experiences can be realized if BRIRs are available for a large range of head poses. This enables dynamic binaural synthesis [3], [5], where the BRIR processing is updated in real-time, taking the listener head pose into account using data from a head-tracking sensor. The intended result is that perceived virtual sound source directions remain fixed with reference to the physical environment the listener is in, even as the listener "looks around" in the VAE.

Acquiring BRIRs for different head poses can be facilitated by placing an artificial head on a turntable controlled with a step-motor or use an apparatus that can move the artificial head into any orientation, not limited to the horizontal plane, as in [6]. A drawback of using an artificial head is that it does not produce individualized BRIRs. BRIRs differ significantly between individuals due to anatomical differences, and using non-personalized BRIRs can lead to perceived localization errors and spectral coloration [7]. As another drawback, it can be expected to take quite a long time to step through many head orientations, and the necessary equipment may be bulky.

Modeling of BRIRs using microphone array Room Impulse Response (RIR) measurements, the topic of this article, has

Manuscript received January 26, 2021; revised July 6, 2021; accepted August 31, 2021. Date of publication September 13, 2021; date of current version September 17, 2021. This work was supported by the Swedish Governmental Agency for Innovation Systems. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Stefan Bilbao. *(Corresponding author: Viktor Gunnarsson.)*

Viktor Gunnarsson is with the Signals and Systems Group, Uppsala University, SE-75103 Uppsala, Sweden, and also with Dirac Research AB, SE-753 40 Uppsala, Sweden (e-mail: viktor.gunnarsson@dirac.com).

Mikael Sternad is with the Signals and Systems Group, Uppsala University, SE-75103 Uppsala, Sweden (e-mail: mikael.sternad@signal.uu.se).

This article has supplementary downloadable material available at https://dx.doi.org/10.21227/p336-3h94., provided by the authors.

Digital Object Identifier 10.1109/TASLP.2021.3110340

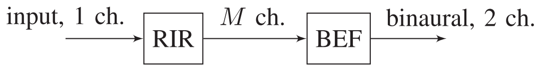


Fig. 1. Principle of BRIR modeling using a microphone array RIR representing M microphone channels and a binaural estimation filter (BEF) outputting a binaural signal.

several benefits – the measurement procedure is fast making it practical to acquire BRIRs for many rooms; BRIRs can be obtained with low effort for any given head pose, and it is possible to model personalized BRIRs. The term modeling is used here to imply that there may be perceptual and/or waveform-level differences to BRIRs measured directly, and it is desirable to minimize these differences. The block diagram in Fig. 1 illustrates a conceptual BRIR modeling procedure. The input signal represents one input channel to a sound system that is to be auralized, and the output is a binaural signal. The RIR-block represents impulse responses from the sound system input to each microphone on a microphone array placed in the desired listening position. The output of the RIR-block represents M simulated microphone array signals. In the second block, the microphone signals are filtered to estimate the binaural signal that would occur for a listener in the position of the microphone array. This filter is referred to as a *binaural estimation filter* in the following. When the input to the system in Fig. 1 is an impulse, the output becomes a BRIR.

Binaural estimation filter design is an active area of research, e.g. [8]–[13]. The effective function of the binaural estimation filter is to synthesize a microphone array directivity pattern which is similar to that of an ear on a human head, referred to here as *HRTF beamforming*. There are two main approaches found in the current literature. The first is direct estimation of binaural signals from microphone signals. In [8], the combination of a microphone array and a filter for direct estimation of binaural signals is referred to as a Virtual Artificial Head (VAH). The other approach is to use the Ambisonics framework and includes two steps, wherein the first (encoding) step, the microphone signals are combined to yield a number of intermediate signals representing a spherical harmonic (SH) decomposition of the sound field [14]. In the second (decoding) step, binaural signals are estimated from the SH-signals with a filter [9], [15].

The main contribution of this work is a proposed framework for binaural estimation filter design, for direct estimation of a binaural signal, which uses a Wiener filtering formulation. The waveform-level error of the binaural signal estimate is minimized under the conditions of a specified sound field spatial energy distribution and a specified microphone self-noise spectrum. No explicit constraints are placed on the microphone array geometry. The response of the microphone array is modeled using anechoic measurements, and a database of HRIRs is used to define the target response in different directions.

The Wiener filtering formulation is new in the context of binaural estimation filter design. It is more general with regard to the flexible sound field and microphone noise models used compared to the problem formulations for direct binaural estimation filter design presented in the references cited above. The sound field model is used to specify a frequency-dependent spatial energy distribution and can be used to weight the estimation error spatially. Filter regularization is controlled by adjusting the frequency-dependent signal-to-noise ratio (SNR) of the modeled microphone array signals.

A time-domain, polynomial matrix formalism is used. The notation, problem formulation and closed-form solution to the filter design problem have been adapted from previous publications on MIMO feed-forward sound field control [16], [17] which is a dual [18] problem. The polynomial methodology has its origins in control theory and is not widespread in audio signal processing research, but has proven to be versatile also in audio, c.f. discussion in [16]. In contrast to popular frequency domain methods, e.g. [19], the filter is constrained to be causal. This facilitates low-latency filter design, as the optimal filter is calculated for a given target latency. To simplify the notation somewhat, the current work uses FIR models for all transfer functions in the filter design problem formulation instead of general IIR models used in prior publications using the polynomial framework. The framework has been extended in the feed-forward case for robust filter design with respect to model errors [16], this has also not been considered in the present work and is a possibility for future research.

By applying our proposed filter design framework to BRIR modeling, we demonstrate its flexibility and suitability to the application of binaural estimation filter design. It is natural to compare the results using our method to prior research on (non-parametric) Ambisonics-based auralization since both approaches in practice implement HRTF beamforming (as defined above), and a majority of previous research on the HRTF beamforming approach to binaural auralization uses the Ambisonics framework. We also use a Spherical Microphone Array (SMA) in our application example, typically used with Ambisonics [20]. A comprehensive comparison of the direct (VAH) and Ambisonics approach to binaural signal estimation is outside the scope of this article. In the present work, we use direct estimation (although the framework can be used to design filters for use with Ambisonics as well, see Section V). We motivate this choice partly by that we make use of directional error weighting in the application of the framework to BRIR modeling (which is not straightforward with Ambisonics), and partly by that from an information-theoretic perspective (and using our problem formulation), we argue that two linear & time-invariant (LTI) filters in series (for calculation of intermediate SH-signals) cannot produce a better estimate of the binaural signal than a single LTI filter (in a mean-square error sense). We also do not make use of features of the Ambisonics approach, such as providing a format for distribution of a SH-based sound field representation and sound field transformations in the SH-domain [15].

In the following, we review some results from research on Ambisonics-based auralization using SMAs. The quality of the binaural signal that can be obtained from a microphone array recording depends on the limitations of the microphone array – its geometry, number of microphones, and microphone electrical noise level. The finite number of microphones in the array dictates the maximum SH-order that can be employed, with a higher SH-order enabling estimation of a binaural signal with lower error. In [21] it was evaluated which SH-order is necessary for a low perceptual error in binaural room auralization, compared to a dummy head reference. Significant differences to the reference were found below 8th order for lateral sources, in [21] corresponding to 110 microphone channels. Since commercially available microphone arrays are currently limited to a SH-order of four and lower, it is of interest to find improved methods for microphone array based room auralization that perform well with fewer microphone channels.

Approaches have been developed to mitigate the effects of limited SH-order on binaural rendering [22]. One approach mentioned in [22] is pre-processing of the HRTF target responses, which can reduce the SH-order needed to represent the HRTFs without giving perceptual consequences [23], [24] (this idea could also be applied to our case of direct binaural signal estimation and is a possibility for future research). However, there is no evidence presented in the cited articles that HRTFs can be simplified in a perceptually transparent way, for all directions and all frequencies, to enable modeling of direct sound at a SH-order down to three, which is the maximum of the SMA employed in our application example.

In [25], it is argued that the minimum SH-order required for Ambisonics-based binaural auralization is mostly dictated by the direct sound path. They suggest a "hybrid Ambisonics" approach where the direct sound part of the BRIR is rendered separately using a spatially dense HRIR data set. In a listening test, the perceived quality of hybrid Ambisonics auralizations ceased to improve beyond an SH-order of three. They used a fourth-order rendering as a reference, the limit of the microphone array they used.

Similarly to e.g. the "hybrid Ambisonics" approach of [25], we model the direct and reflected BRIR parts separately, but here within the same filter design framework. Compared to [25] we present a complete filter design process for BRIR modeling that includes DoA-estimation of the direct sound and inversion of the microphone array dynamics. In contrast to [25] we also compare the modeled BRIRs to reference measured BRIRs using both objective performance metrics and a listening test. The results support the feasibility of our method and show a small perceptual difference between measured and modeled BRIRs.

As an alternative to the HRTF beamforming approach to BRIR modeling that is the focus of this article, a class of perceptually motivated methods that typically rely on a simplified parametric description of the sound field aim at reproducing the most important perceptual features of BRIRs, and usually only require a small number of microphone channels for RIR measurements [26]–[33]. Our work is inspired by these methods in that we consider perceptually important parameters and employ DoA-estimation, but distinct from these methods in that accurate waveform-level modeling is accomplished up to a frequency that is dependent on the capability of the microphone array used. Accurate waveform-level modeling can be necessary e.g. when auralizing arrays of speakers and the in-room phase relationship between speakers must be modeled correctly.

The article is organized as follows: first, the polynomial matrix notation is introduced. Section II then summarizes the proposed BRIR modeling procedure. Section III presents the binaural estimation filter design problem formulation and its solution, as well as definitions of performance metrics. Section IV presents a practical BRIR modeling example employing a commercially available 19-channel microphone array. Section V discusses the results and Section VI gives conclusions.

A. Polynomial Matrix Notation

The notation used is as follows. The discrete time index is denoted by t. The time delay operator q^{-n} has the effect of a delay of n samples, so that $q^{-n}y(t) = y(t-n)$. A polynomial in q^{-1} thus represents a difference equation and a scalar polynomial $c(q^{-1}) = (c_0 + c_1q^{-1} + c_2q^{-2} + \cdots + c_Nq^{-N})$ represents a difference equation describing an FIR filter. Multiplication of a polynomial in q^{-1} with a time signal or another polynomial in q^{-1} results in a convolution operation.

A polynomial matrix has polynomials as elements, each element representing a finite impulse response, and is denoted by bold capital letters, e.g. $\boldsymbol{A}(q^{-1})$. Writing $\boldsymbol{A}_{(i,j)}(q^{-1})$ selects the element at row i and column j, whereas using the colon operator selects an entire row or column, e.g. $\boldsymbol{A}_{(i,:)}(q^{-1})$.

A polynomial matrix subject to the conjugate operator, $\boldsymbol{A}_*(q)$, is complex conjugate transposed and the time delay operator q^{-1} is substituted by its reciprocal, the time-advance operator q, i.e. each polynomial is reflected around time 0 and effectively time-reversed. This paper deals only with real-valued polynomial matrices. See e.g. [34], [35] for an introduction to polynomial methods as used in control engineering.

The windowing operator $\mathcal{W}\{\cdot\}$ applies a time window to the coefficients of a polynomial expression. The window properties are described in the context where it is used. A matrix of scalars containing the n^{th} degree coefficients of a polynomial expression is constructed by writing $\{\cdot\}_{deg=n}$.

Expressions are evaluated in the frequency domain at angular frequency ω by the Discrete-Time Fourier Transform operator $\mathscr{F}^{\omega}\{\cdot\}$, which has the effect of substituting the time delay operator q^{-n} by the function $e^{-j\omega nT_s}$, where T_s is the sampling period. Writing $\mathscr{F}^{\omega}\{\cdot\}_{(i,j)}$ selects the element at row i and column j.

Regular matrices and vectors of scalars are denoted by upper case and lower case bold letters respectively, e.g. $\boldsymbol{A}$, $\boldsymbol{a}$. Scalar quantities are written with normal font-weight. In some places, the argument (q^{-1}) to a polynomial matrix has been omitted for brevity, in contexts where the risk is low to confuse it with a regular matrix.

Important Symbols

$\boldsymbol{B}(q^{-1})$	Array model (anechoic measurements)
$\boldsymbol{G}(q^{-1})$	Array RIR measurement
$\boldsymbol{S}(q^{-1})$	Target HRIRs
$\boldsymbol{C}(q^{-1})$	Sound field signal model
$\boldsymbol{M}(q^{-1})$	Microphone noise model
$\boldsymbol{F}(q^{-1})$	Binaural estimation filter
$\boldsymbol{\Gamma}(q^{-1})$	Power spectrum correction filter

II. BRIR Modeling Procedure

To visualize the structural makeup of a BRIR, consider a simplified measurement procedure where a loudspeaker outputs an acoustic impulse and the resulting BRIR is measured at the ears of a human test subject. In a normal room environment, the sound field around the head would consist of a superposition of a direct response from the loudspeaker and delayed room reflections. Structurally, the measured BRIR consists of a superposition of HRIRs corresponding to the direct sound and each delayed reflection, with the direct sound normally making up the first few milliseconds of the BRIR. Exchanging the human subject with a microphone array and measuring a multichannel RIR, the RIR contains corresponding direct and reflected sound time segments.

The proposed BRIR modeling scheme splits the microphone array RIR into two parts using time windowing, one part containing the direct sound and the other part containing reflected sound. The RIR time segments are combined with individually designed binaural estimation filters.

The same filter design framework is used for the direct and reflected sound binaural estimation filters. The idea is to measure the directional distribution of sound power of the direct sound, which is assumed to be highly directional, and use this information when designing the direct sound binaural estimation filter such that the direct sound HRTF response is accurately modeled. To this end, we perform a direction-of-arrival (DoA) analysis on the direct sound RIR segment. The reflected sound field is assumed to have low directionality and the binaural estimation filter for reflected sound is designed assuming that no knowledge is available about the sound field directionality.

The BRIR modeling scheme will be expressed mathematically for one sound system input channel in the following, with the understanding that the procedure is repeated several times for a multichannel sound system.

Let the polynomial matrix $\boldsymbol{G}(q^{-1})$ represent a microphone array RIR measurement taken at a desired listening position, giving $\boldsymbol{G}(q^{-1})$ the dimensions $[M \times 1]$, where M is the number of microphone array channels. The direct part of the RIR, denoted $\boldsymbol{G}_d(q^{-1})$, is extracted by applying a time window function to each polynomial in $\boldsymbol{G}(q^{-1})$. The reflected sound part of the RIR is then given by $\boldsymbol{G}_r(q^{-1}) = \boldsymbol{G}(q^{-1}) - \boldsymbol{G}_d(q^{-1})$.

The modeled BRIRs are obtained by filtering $\boldsymbol{G}_d(q^{-1})$ and $\boldsymbol{G}_r(q^{-1})$ with the binaural signal estimation filters $\boldsymbol{F}_d(q^{-1})$ and $\boldsymbol{F}_r(q^{-1})$, which are to be designed, then summing the result:

$$\boldsymbol{H}_{BRIR}(q^{-1}) = \boldsymbol{F}_d(q^{-1})\boldsymbol{G}_d(q^{-1}) + \boldsymbol{F}_r(q^{-1})\boldsymbol{G}_r(q^{-1}). \quad (1)$$

Here, $\boldsymbol{F}_d(q^{-1})$ and $\boldsymbol{F}_r(q^{-1})$ have dimensions $[2 \times M]$ and $\boldsymbol{H}_{BRIR}(q^{-1})$ has dimensions $[2 \times 1]$, i.e. containing BRIRs for the two ears.

The proposed BRIR modeling topology is illustrated in Fig. 2. The required major steps of the modeling procedure are as follows:

- Measure the anechoic multichannel impulse response model of the microphone array used, for N directions covering a full sphere, see polynomial matrix $\boldsymbol{B}(q^{-1})$, defined in Section III-A.
- Place the microphone array in an acoustic environment for which it is desired to obtain a BRIR and measure $\boldsymbol{G}(q^{-1})$, representing RIRs to M microphones.
- Split $\boldsymbol{G}(q^{-1})$ into time segments $\boldsymbol{G}_d(q^{-1})$, containing early arriving direct sound, and $\boldsymbol{G}_r(q^{-1})$, containing late arriving reflected sound.
- Determine DoA of the direct sound by a DoA-analysis of $\boldsymbol{G}_d(q^{-1})$, details in Section IV-B3.
- Design binaural estimation filters $\boldsymbol{F}_d(q^{-1})$ and $\boldsymbol{F}_r(q^{-1})$ for direct and reflected sound respectively. Use the DoA-information when designing the direct sound filter, details in Section IV-B3 and IV-B2.
- Filter the RIR time segments $\boldsymbol{G}_d(q^{-1})$ and $\boldsymbol{G}_r(q^{-1})$ with filters $\boldsymbol{F}_d(q^{-1})$ and $\boldsymbol{F}_r(q^{-1})$ to obtain direct and reflected sound BRIR segments respectively. Sum the BRIR segments to obtain $\boldsymbol{H}_{BRIR}(q^{-1})$ according to (1).

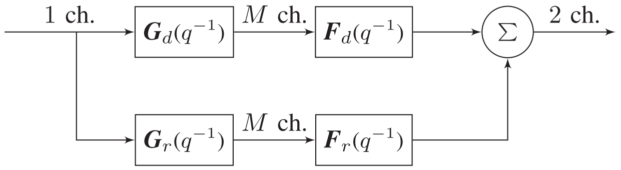

1 ch.
$\boldsymbol{G}_d(q^{-1})$
M ch.
$\boldsymbol{F}_d(q^{-1})$
Σ
2 ch.
$\boldsymbol{G}_r(q^{-1})$
M ch.
$\boldsymbol{F}_r(q^{-1})$

Fig. 2. Proposed topology for BRIR modeling, with separate modeling of direct/reflected sound parts of the BRIR.

III. Binaural Estimation Filter Design

This section covers general filter design theory for binaural signal estimation from microphone array recordings. Section IV then discusses suitable design parameters when designing filters $\boldsymbol{F}_d(q^{-1})$ and $\boldsymbol{F}_r(q^{-1})$ for the application of BRIR modeling.

A. Multichannel Wiener Filter Problem Formulation

The goal is to design a filter $\boldsymbol{F}(q^{-1})$ for a microphone array to estimate the ear signals that would be observed for a head placed in the same position as the array, given a specified sound field model. To formulate this as a Wiener filter design problem, it is necessary to specify known reference binaural ear signals that are to be estimated, in a minimum mean square error (MMSE) sense, by filtering of measured noisy microphone signals. As a first step, let the model sound field consist of N mutually uncorrelated sound sources at spatial locations $\Omega_i, i = 1 \ldots N$. An approximately uniform distribution of source locations over a spherical surface is a suitable choice for applications where the estimation error needs to be controlled for all directions.

Let the sound source outputs be statistically modeled as moving average (MA) processes: $u_i(t) = c_i(q^{-1})e_i(t)$, where $e_i(t)$ is white, zero-mean, unit variance noise and the polynomials $c_i(q^{-1})$ are minimum-phase. The polynomial $c_i(q^{-1})$ represents an MA difference equation acting on $e_i(t)$. The vector-valued source signal model then becomes

$$\boldsymbol{u}(t) = \boldsymbol{C}(q^{-1})\boldsymbol{e}(t) \quad (2)$$

where $\boldsymbol{u}(t)$ and $\boldsymbol{e}(t)$ are both of dimension $[N \times 1]$ and $E[\boldsymbol{e}(t)\boldsymbol{e}(t)^T] = I_N$. The $[N \times N]$ polynomial matrix $\boldsymbol{C}(q^{-1})$

is square and diagonal to model uncorrelated but potentially colored sound sources.

The selection of $\boldsymbol{C}(q^{-1})$ specifies an assumed spatial distribution of sound power in the sound field and can be used to weight the estimation error for different directions. Setting only selected diagonal elements of $\boldsymbol{C}(q^{-1})$ to be non-zero models a directional sound field, whereas setting all diagonal elements of $\boldsymbol{C}(q^{-1})$ to be equal and non-zero can be used to model a diffuse sound field.[1] Both these examples are used in Section IV when designing filters $\boldsymbol{F}_d(q^{-1})$ and $\boldsymbol{F}_r(q^{-1})$, respectively.

The two sound pressures at the ears of a listener in the model sound field constitute reference signals, given by the reference signal vector $\boldsymbol{f}(t)$ of dimension $[2 \times 1]$, which is a concatenation of the two (left and right) ear signals. The ear signals are superpositions of filtered versions of the source signals and can be written as

$$\boldsymbol{f}(t) = \boldsymbol{S}(q^{-1})\boldsymbol{u}(t). \tag{3}$$

Here, $\boldsymbol{S}(q^{-1})$ is a polynomial matrix of dimension $[2 \times N]$ and each polynomial element represents the HRIR FIR-coefficients for a corresponding source direction and ear.

Suppose now that the listener is substituted by a microphone array with M microphones at the listening position. The $[M \times 1]$ measurement signal vector $\boldsymbol{y}(t)$ can then be written as

$$\boldsymbol{y}(t) = \boldsymbol{B}(q^{-1})\boldsymbol{u}(t) + \boldsymbol{M}(q^{-1})\boldsymbol{v}(t) \tag{4}$$

where $\boldsymbol{B}(q^{-1})$ is of dimension $[M \times N]$. Element m, n of $\boldsymbol{B}(q^{-1})$ contains a polynomial representing the impulse response from sound source n (at location Ω_n) to microphone m. The term $\boldsymbol{M}(q^{-1})\boldsymbol{v}(t)$ models additive noise. It can for example model microphone electrical self-noise or can generally be used to regularize the filter design problem, as discussed in Section III-C. The noise signal vector $\boldsymbol{v}(t)$ is of dimensions $[M \times 1]$ and contains M mutually uncorrelated zero-mean unit-variance white noise signals. The polynomial matrix $\boldsymbol{M}(q^{-1})$ models the spectral properties of the additive noise and is square, diagonal, of dimensions $[M \times M]$ and each element is minimum phase.

We can now formulate our estimation problem as follows: we seek a causal and stable filter $\boldsymbol{F}(q^{-1})$ that, when applied to the measured microphone signals $\boldsymbol{y}(t)$, produces an estimate of the reference signal vector $\boldsymbol{f}(t)$. This can be written as

$$\hat{\boldsymbol{f}}(t - t_m|t) = \boldsymbol{F}(q^{-1})\boldsymbol{y}(t) \tag{5}$$

where the vector $\hat{\boldsymbol{f}}(t - t_m|t)$ is an estimate of $\boldsymbol{f}(t - t_m)$ given measurements up to and including time index t. The *smoothing lag* design parameter t_m is normally chosen to be positive and non-zero. A larger t_m lets the filter have a longer pre-response, i.e. a larger look-ahead time, potentially improving the quality of the estimate.

The estimation error signal $\boldsymbol{\varepsilon}(t)$ to be minimized is defined by

$$\boldsymbol{\varepsilon}(t) = \hat{\boldsymbol{f}}(t - t_m|t) - \boldsymbol{f}(t - t_m). \tag{6}$$

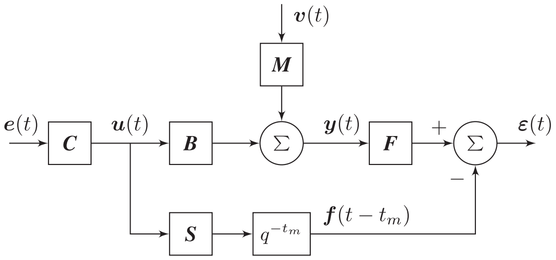


Fig. 3. The multichannel Wiener filter design problem.

The optimization criterion for the filter $\boldsymbol{F}(q^{-1})$ is that it should minimize the variance of $\boldsymbol{\varepsilon}(t)$:

$$\min_{\boldsymbol{F}(q^{-1})}: J = E\left[\boldsymbol{\varepsilon}(t)^T\boldsymbol{\varepsilon}(t)\right] \tag{7}$$

where $E[\cdot]$ represents the expectation with respect to random signals $\boldsymbol{e}(t)$ and $\boldsymbol{v}(t)$ and $\boldsymbol{F}(q^{-1})$ is constrained to be causal and stable. The complete filter design problem is illustrated in Fig. 3.

B. Solution

The final filter is obtained by first calculating the MMSE-optimal Wiener filter according to the above problem formulation, then applying a perceptually motivated power spectrum correction to it.

1) Wiener Solution: The unique solution to (7) can be found using methods of numerical optimization. However, solving this optimization problem can be quite computationally demanding and numerically sensitive for high-dimensional MIMO problems. In the limit $t_m \to \infty$, the solution to (7) is the noncausal Wiener filter. It can be obtained by expressing (2)–(4) in the frequency domain (substituting the delay operator q^{-n} by $e^{-j\omega n T_s}$) and then minimizing the variance of $\boldsymbol{\varepsilon}(t)$ pointwise in the frequency domain. While being a viable solution in some cases, the noncausal Wiener filter needs to have its noncausal part truncated to be realizable, making the filter sub-optimal if the truncated part contains significant energy. Wrap-around effects can also be an issue when going from the frequency domain to the time domain, if the transform size is insufficient.

In our case, with a fixed and finite smoothing lag t_m, we seek to calculate the causal (realizable) Wiener filter (5). This filter can in general not be obtained by pointwise optimization in the frequency domain, as the optimal properties of the filter at one frequency are influenced by the model properties at other frequencies. Here, a time-domain solution, using a polynomial equations approach, offers a versatile alternative. We use this approach. The solution, outlined below, is a special case of that presented and derived in section V.C of [36], which treats a more general case, where the blocks in Fig. 3 are not restricted to finite impulse responses. In the present case, all blocks are modeled with finite impulse responses, except the optimal filter $\boldsymbol{F}(q^{-1})$, which is a matrix of stable IIR filters, as discussed below.

As detailed in [36], two polynomial matrix equations need to be solved to find the optimal filter, a *spectral factorization* and

[1] assuming that a spatial sampling scheme is used for the source locations Ω_i that has equal quadrature weight in all directions. Otherwise, quadrature weights can be added to $\boldsymbol{C}(q^{-1})$ so that directions of more dense sampling are not weighted higher.

a *Diophantine equation*. The spectral factorization equation,

$$\boldsymbol{\beta}(q^{-1})\boldsymbol{\beta}_*(q) = \boldsymbol{B}(q^{-1})\boldsymbol{C}(q^{-1})\boldsymbol{C}_*(q)\boldsymbol{B}_*(q) + \boldsymbol{M}(q^{-1})\boldsymbol{M}_*(q), \tag{8}$$

is solved for a minimum phase, invertible $[M \times M]$ polynomial matrix spectral factor $\boldsymbol{\beta}(q^{-1})$ and the Diophantine equation

$$q^{-t_m}\boldsymbol{S}(q^{-1})\boldsymbol{C}(q^{-1})\boldsymbol{C}_*(q)\boldsymbol{B}_*(q) = \boldsymbol{Q}(q^{-1})\boldsymbol{\beta}_*(q) + q\boldsymbol{L}_*(q) \tag{9}$$

is solved for polynomial matrices $\boldsymbol{Q}(q^{-1})$ and $\boldsymbol{L}_*(q)$, both having dimensions $[2 \times M]$. Note that the dimensions of the terms in these equations only depend on M and not on N which, for large N, is an advantage computationally. Combining the two results, the optimal filter is a transfer function matrix in right matrix fraction description (MFD) [34] form:

$$\boldsymbol{F}(q^{-1}) = \boldsymbol{Q}(q^{-1})\boldsymbol{\beta}^{-1}(q^{-1}). \tag{10}$$

The spectral factorization equation (8) is quadratic in the coefficients of the polynomial matrix $\boldsymbol{\beta}(q^{-1})$. It can be solved efficiently iteratively, see e.g. [37], to obtain a polynomial matrix spectral factor $\boldsymbol{\beta}(q^{-1})$ which has a causal inverse $\boldsymbol{\beta}^{-1}(q^{-1})$, and for which all zeros of $\det(\boldsymbol{\beta}(z^{-1}))$ are within the unit circle $|z| = 1$ of the complex plane. As a result, the inverse $\boldsymbol{\beta}^{-1}(q^{-1})$ is a rational (transfer function) matrix that represents a stable and causal discrete-time dynamic system.

The equation (9) is linear in the coefficients of $\boldsymbol{Q}(q^{-1})$ and $\boldsymbol{L}_*(q)$. It is equivalent to a linear system of equations in the coefficients of $\boldsymbol{Q}(q^{-1})$ and $\boldsymbol{L}_*(q)$, and introduces the causality constraint due to the factor q^{-t_m} on the left-hand side. It can be shown that $\boldsymbol{L}_*(q) \to 0$ when $t_m \to \infty$, in which case we approach the noncausal Wiener solution, see [38].

Since $\boldsymbol{\beta}^{-1}(q^{-1})$ is guaranteed to be a stable and causal rational matrix, the filter (10) represents a $[2 \times M]$-matrix of stable and causal discrete-time filters. Since $\boldsymbol{\beta}^{-1}(q^{-1})$ has infinite impulse response, $\boldsymbol{F}(q^{-1})$ represents a matrix of IIR filters. This filter bank could be realized in state-space form, based on the MFD representation $\boldsymbol{Q}(q^{-1})\boldsymbol{\beta}^{-1}(q^{-1})$, see [34], but it is often more convenient to approximate it by a FIR filter matrix by truncating its impulse responses. In the following, $\boldsymbol{F}(q^{-1})$ is assumed to be a FIR filter matrix.

2) Power Spectrum Correction: The optimal filter criterion given by (7) does not guarantee that the power spectrum of the estimated ear signals will match that of the reference ear signals. The maximum value of the error criterion is bounded by its value when the filter gain is zero. Thus, if no filter can substantially reduce the error below this bound at some frequency, the MMSE-optimal filter tends to attenuate that frequency. This is typically undesirable from a perceptual standpoint (c.f. Section IV-B2) and the MMSE-optimal filter is therefore adjusted for correct ear signal power spectrum. The final corrected filter $\boldsymbol{F}^c(q^{-1})$ becomes

$$\boldsymbol{F}^c(q^{-1}) = \boldsymbol{\Gamma}(q^{-1})\boldsymbol{F}(q^{-1}), \tag{11}$$

where $\boldsymbol{\Gamma}(q^{-1})$ is a $[2 \times 2]$ diagonal polynomial matrix that is chosen so that the power spectra of the estimated and reference ear signals become approximately equal. In the frequency domain, this equality can be expressed as

$$\mathscr{F}^\omega\left\{\boldsymbol{\Gamma}\boldsymbol{F}\boldsymbol{B}\boldsymbol{C}\boldsymbol{C}_*\boldsymbol{B}_*\boldsymbol{F}_*\boldsymbol{\Gamma}_*\right\}_{(k,k)} \simeq \mathscr{F}^\omega\left\{\boldsymbol{S}\boldsymbol{C}\boldsymbol{C}_*\boldsymbol{S}_*\right\}_{(k,k)}, \tag{12}$$

where the index $k \in \{1, 2\}$ selects the diagonal element corresponding to the left or right ear power spectrum respectively. To fulfill this relation, the elements of $\boldsymbol{\Gamma}(q^{-1})$ are chosen as minimum phase polynomials with power spectrum defined by

$$\mathscr{F}^\omega\left\{\boldsymbol{\Gamma}\boldsymbol{\Gamma}_*\right\}_{(k,k)} = \frac{\mathscr{F}^\omega\left\{\boldsymbol{S}\boldsymbol{C}\boldsymbol{C}_*\boldsymbol{S}_*\right\}_{(k,k)}}{\mathscr{F}^\omega\left\{\boldsymbol{F}\boldsymbol{B}\boldsymbol{C}\boldsymbol{C}_*\boldsymbol{B}_*\boldsymbol{F}_*\right\}_{(k,k)} + \rho(\omega)}, \tag{13}$$

where $\rho(\omega)$ is a small regularization parameter.

C. Filter Regularization

It is desirable to avoid excessive filter gains, both because this amplifies microphone electrical self-noise and increases sensitivity to model errors in $\boldsymbol{B}(q^{-1})$. The filter gains of the optimal filter $\boldsymbol{F}(q^{-1})$ generally depend on the frequency-dependent conditioning of the filter design problem, which can be adjusted using regularization.

The additive noise model $\boldsymbol{M}(q^{-1})$ controls regularization and is selected to specify the SNR of the measurement signal vector $\boldsymbol{y}(t)$ for the nominal signal level in the model sound field. To find an expression for the SNR of $\boldsymbol{y}(t)$, consider a covariance polynomial matrix $\boldsymbol{P}_y(q, q^{-1})$ containing the covariance sequences between the signals in $\boldsymbol{y}(t)$, defined as

$$\{\boldsymbol{P}_y(q, q^{-1})\}_{deg=n} := E[\boldsymbol{y}(t)^T\boldsymbol{y}(t-n)]. \tag{14}$$

Here, the time-lag n translates to a corresponding coefficient matrix for polynomial degree n in $\boldsymbol{P}_y(q, q^{-1})$. The signal component $\boldsymbol{P}_s(q, q^{-1})$ and noise component $\boldsymbol{P}_n(q, q^{-1})$ of $\boldsymbol{P}_y(q, q^{-1})$ are given by

$$\boldsymbol{P}_y(q, q^{-1}) = \boldsymbol{P}_s(q, q^{-1}) + \boldsymbol{P}_n(q, q^{-1}), \tag{15}$$

$$\boldsymbol{P}_s(q, q^{-1}) = \boldsymbol{B}\boldsymbol{C}\boldsymbol{C}_*\boldsymbol{B}_*, \tag{16}$$

$$\boldsymbol{P}_n(q, q^{-1}) = \boldsymbol{M}\boldsymbol{M}_*. \tag{17}$$

We can note that $\boldsymbol{P}_y(q, q^{-1}) = \boldsymbol{\beta}(q^{-1})\boldsymbol{\beta}_*(q)$. The Fourier transform of the m^{th} diagonal element of $\boldsymbol{P}_s(q, q^{-1})$ and $\boldsymbol{P}_n(q, q^{-1})$ correspond to the respective contribution to the power spectral density of the m^{th} microphone signal, thus the SNR of $\boldsymbol{y}(t)$, at frequency ω and for microphone index m, can be calculated as

$$SNR_m(\omega) = \frac{\mathscr{F}^\omega\left\{\boldsymbol{P}_s(q, q^{-1})\right\}_{(m,m)}}{\mathscr{F}^\omega\left\{\boldsymbol{P}_n(q, q^{-1})\right\}_{(m,m)}}. \tag{18}$$

As seen, the SNR of $\boldsymbol{y}(t)$ depends on $\boldsymbol{C}(q^{-1})$, $\boldsymbol{M}(q^{-1})$, and $\boldsymbol{B}(q^{-1})$. A suggestion is to specify the model sound field $\boldsymbol{C}(q^{-1})$ freely first, and then select the measurement signal noise model $\boldsymbol{M}(q^{-1})$ to obtain the desired SNR level.

In the application example in Section IV, all microphones in the array can be assumed to have equal properties and $\boldsymbol{M}(q^{-1})$ was selected to model a constant frequency independent SNR. This is a reasonable general starting point when it is desired to apply equal penalty to high filter gains at all frequencies. The microphone noise covariance polynomial matrix $\boldsymbol{P}_n(q, q^{-1})$ is then constructed from a scaled average of the diagonal elements of the microphone signal covariance polynomial matrix:

$$\boldsymbol{P}_n(q, q^{-1}) = \boldsymbol{M}\boldsymbol{M}_* = \sigma_v^2 \operatorname{diag}_M\left\{\frac{\operatorname{tr}(\boldsymbol{P}_s(q, q^{-1}))}{M}\right\}, \tag{19}$$

where σ_v^2 is a noise variance parameter, $tr(\cdot)$ denotes the matrix trace operation and $\mathrm{diag}_M\{\cdot\}$ means constructing a $[M \times M]$ diagonal matrix from the scalar element within the brackets. For the case where all microphones have identical useful signal power, the resulting measurement signal SNR then becomes $SNR_m = 1/\sigma_v^2$. Note that it is sufficient to calculate $\boldsymbol{M}(q^{-1})\boldsymbol{M}_*(q)$, as $\boldsymbol{M}(q^{-1})$ is not explicitly needed to solve the spectral factorization in (8).

In theory, $\boldsymbol{M}(q^{-1})$ could be chosen to accurately model the electrical self-noise spectrum of the microphone model used. In general, this choice of $\boldsymbol{M}(q^{-1})$ is not necessarily suitable since the actual SNR of recorded microphone signals is anyhow usually unknown and time-varying. Rather, how much filter regularization to apply is a balance between MSE-performance, sensitivity to model errors, filter noise amplification, and (for general binaural recording) how subjectively disturbing a certain background noise spectrum is. Some manual tuning of $\boldsymbol{M}(q^{-1})$ is thus typically required for each specific application and design goal.

Finally, it can be noted that the relative scaling of $\boldsymbol{C}(q^{-1})$, $\boldsymbol{B}(q^{-1})$ and $\boldsymbol{M}(q^{-1})$ affects filter regularization, but a scaling of $\boldsymbol{S}(q^{-1})$ results in the same optimal filter $\boldsymbol{F}(q^{-1})$, only scaled correspondingly.

D. Performance Metrics

Some key filter design performance metrics used in the application example in Section IV are described in the following.

The normalized MSE (NMSE) equals the error power in the estimated ear signals normalized by the signal power in the reference ear signals, indicating how well the beamformer approximates the HRTF beampattern target at each frequency. The modeling error for all source directions is needed in its calculation and is given by

$$\boldsymbol{E}(q^{-1}) = q^{-t_m}\boldsymbol{S}(q^{-1})\boldsymbol{C}(q^{-1}) - \boldsymbol{F}(q^{-1})\boldsymbol{B}(q^{-1})\boldsymbol{C}(q^{-1}). \quad (20)$$

Further, define the reference ear signal covariance polynomial matrix as

$$\boldsymbol{P}_{ref}(q, q^{-1}) = \boldsymbol{S}\boldsymbol{C}\boldsymbol{C}_*\boldsymbol{S}_*, \quad (21)$$

adhering here to the definition of a covariance polynomial matrix given by (14). The NMSE can then be defined as

$$NMSE^k(\omega) = \frac{\mathscr{F}^{\omega}\left\{\boldsymbol{E}\boldsymbol{E}_*\right\}_{(k,k)}}{\mathscr{F}^{\omega}\left\{\boldsymbol{P}_{ref}\right\}_{(k,k)}}, \quad (22)$$

and it attains a value between 0 and 1 for each frequency. Notice that NMSE is evaluated using $\boldsymbol{F}(q^{-1})$ without power spectrum correction.

Another perceptually important metric is the frequency-dependent coherence between the estimated ear signals $\hat{\boldsymbol{f}}(t)$. Define the estimated ear signal covariance polynomial matrix as:

$$\boldsymbol{P}_{est}(q, q^{-1}) = \boldsymbol{F}^c\boldsymbol{P}_s\boldsymbol{F}^c_*. \quad (23)$$

The interaural coherence of the estimated ear signals then becomes

$$C_{lr}(\omega) = \frac{\left|\mathscr{F}\left\{\boldsymbol{P}_{est(1,2)}\right\}_{\omega}\right|}{\sqrt{\mathscr{F}\left\{\boldsymbol{P}_{est(1,1)}\boldsymbol{P}_{est(2,2)}\right\}_{\omega}}} \quad (24)$$

and should ideally equal the coherence of the reference ear signals, which can be calculated with the same expression by substituting $\boldsymbol{P}_{ref}(q, q^{-1})$ in place of $\boldsymbol{P}_{est}(q, q^{-1})$.

The level of microphone self-noise in the estimated ear signals is important to consider. It has an impact especially when recording quiet sound fields, i.e. in the absence of any sound that can mask the noise. A relevant question to ask is how strong the noise is in the estimated ear signals compared to the noise level of a recording taken with a single microphone from the array (where the recordings have been adjusted for similar useful signal gain). A metric that answers this question, which is common in the context of uni-directional beamformer design, is White Noise Gain (WNG). It measures the power ratio of useful signal gain (i.e. beampattern gain) to filter noise gain and is normally defined for a single direction. For the application of HRTF beampattern synthesis, it is more relevant to consider an average of WNG over all directions [39]. The power average of WNG in all measured directions can be defined as

$$\begin{aligned} WNG_{avg}^k(\omega) = 10\log_{10}\mathscr{F}^{\omega}\{\boldsymbol{P}_{est,(k,k)}\} \\ - 10\log_{10}\mathscr{F}^{\omega}\{\mathrm{tr}(\boldsymbol{P}_s)/M\} \\ - 10\log_{10}\mathscr{F}^{\omega}\{\boldsymbol{F}^c_{(k,:)}\boldsymbol{F}^c_{*(:,k)}\}. \end{aligned} \quad (25)$$

Here, the first term represents the estimated ear signal energy, the second term represents the average energy of the recorded microphone signals, and the third term represents the filter noise gain. The first and second terms together represent the average beampattern gain. The index k selects the left or right ear filter.

If the filter design model sound field can be considered diffuse, then $WNG_{avg}(\omega)$ approximates the relative SNR of the estimated ear signals compared to the SNR of a single microphone capsule when recording diffuse sound fields. It follows that $-WNG_{avg}(\omega)$ indicates the relative noise level of the estimated ear signals compared to that of a single microphone capsule, in the absence of a signal.

IV. Application Example

The purpose of the application example is to illustrate the use of the proposed filter design framework and demonstrate the feasibility of the described BRIR modeling method by investigating a practical example and evaluating the resulting performance. BRIRs are modeled for the torsoless Neumann KU100 artificial head and a direct comparison is made between modeled and measured BRIRs using a KU100 head available in the lab.

The room chosen for BRIR acquisition as well as for microphone array measurements is a large, undamped (RT60 around 1.4 s), mostly empty storage-facility type room. It is 7.5 × 11 m with around 5 m ceiling height.

A commercially available microphone array was selected, Zylia ZM-1. It is a consumer-grade spherical microphone array with 19 microphone channels, about 10 cm diameter, and an ASIO driver interface over USB. While the filter design framework puts no restrictions on the array geometry, a spherical microphone array is suitable for this application because the design of the binaural estimation filter $\boldsymbol{F}_r(q^{-1})$, for the reflected sound part of the BRIR, makes use of its approximately uniform beamforming performance in all directions.

We use a right-handed coordinate system where azimuth is the counter-clockwise angle in the horizontal plane relative to the positive x-axis and elevation is the angle relative to the horizontal plane. The coordinate system is defined in relation to the microphone array if nothing else is indicated. Additionally, we use yaw, pitch & roll Euler-angles to denote the head pose (i.e., look direction) of the "virtual head" that the binaural estimation filter implements using the microphone array.

A. Measurements

All impulse response measurements were carried out using the logarithmic sine-sweep method [40] with 4 s long sweeps. The microphone array impulse responses in $\boldsymbol{B}(q^{-1})$ were measured using a single speaker by rotating the microphone array to different orientations. This was done using a custom-designed measurement robot, controlled via a serial interface, and able to rotate the microphone array to any orientation with $<1^\circ$ precision. A Tukey window $\mathcal{W}_B$ of length 4 ms with cosine fraction 0.5 was applied to each measured impulse response in $\boldsymbol{B}(q^{-1})$ to emulate anechoic measurements.

The microphone array was placed 1.6 m above the floor. A Genelec 8010 A speaker was used as sound source, placed at the same height and at 2.1 m distance. This gave a reflection-free time window of around 5 ms before the first reflection. The speaker was chosen due to its small size and broadband response, to roughly represent a point source. A sample rate of 48 kHz was used for all measurements.

A spatial sampling grid consisting of $N = 300$ points equidistributed over a spherical surface [41] was chosen for the measurement of $\boldsymbol{B}(q^{-1})$. This gave $\boldsymbol{B}(q^{-1})$ dimensions $[19 \times 300]$ and the target HRIR polynomial matrix $\boldsymbol{S}(q^{-1})$ got dimensions $[2 \times 300]$. The choice of the grid density is discussed further in the sections below.

We chose to model BRIRs for the same Genelec speaker in the same room. The microphone array RIR, $\boldsymbol{G}(q^{-1})$, of size $[19 \times 1]$, was measured using the same setup with the speaker in front of the array, i.e. at 0° azimuth, 0° elevation relative to the array (the setup was moved slightly so that $\boldsymbol{G}(q^{-1})$ would not correspond exactly to a measured grid point in $\boldsymbol{B}(q^{-1})$). The RIRs in $\boldsymbol{G}(q^{-1})$ were truncated to 0.83 s (40000 samples). After this time, the decay tails started to be dominated by measurement noise.

1) Compensation for Measurement Speaker Response: The design theory for the binaural estimation filter in Section III assumes that the anechoic array model $\boldsymbol{B}(q^{-1})$ only contains the dynamics of the microphone array. In practice however, it includes the dynamics of the speaker used to measure it. The subsequent effect on the filter design was largely eliminated by convolving also the target HRIRs in $\boldsymbol{S}(q^{-1})$ with the speaker impulse response. To this end, the on-axis speaker impulse response $h_{lsp}(q^{-1})$ was measured using the same setup with an Earthworks M30 microphone, calibrated for a flat response, in place of the Zylia microphone array.

The measurement of $\boldsymbol{B}(q^{-1})$ can be modeled as

$$\boldsymbol{B}(q^{-1}) = \mathcal{W}_B \left\{ \boldsymbol{B}_{arr}(q^{-1}) h_{lsp}(q^{-1}) \right\}, \tag{26}$$

where $\boldsymbol{B}_{arr}(q^{-1})$ represents the response of the microphone array, $h_{lsp}(q^{-1})$ the response of the measurement speaker, and $\mathcal{W}_B$ is the time window applied to the measurements. The modified target response becomes

$$\boldsymbol{S}(q^{-1}) = \boldsymbol{S}_{hrir}(q^{-1}) \mathcal{W}_B \left\{ h_{lsp}(q^{-1}) \right\}. \tag{27}$$

Here, $\boldsymbol{S}_{hrir}(q^{-1})$ represents the target HRIRs.

B. Filter Design

1) Parameter Choices: The smoothing lag was set to $t_m = 480$ samples, corresponding to a latency of 10 ms, which provides a good margin for the filter pre-response and thus gives full performance with respect to MSE. A constant frequency-independent SNR was modeled for the microphone signals $\boldsymbol{y}(t)$, as described in Section III-C – an SNR of $10 \log_{10}(1/\sigma_v^2) = 20$ dB was generally used, except for the plots that compare the effect of different SNR levels.

To define the target HRIRs, $\boldsymbol{S}_{hrir}(q^{-1})$, we used the public HRTF database for Neumann KU100 published in [42]. The database HRIRs are densely sampled with 2° resolution in azimuth/elevation and have a length of 128 samples (2.7 ms). Nearest-neighbor interpolation was used to pick HRIRs out of the database for the 300 point grid used in $\boldsymbol{S}_{hrir}(q^{-1})$. A small spectral adjustment was applied to the database HRTFs to make them more similar to the HRTFs of the KU100 unit available in the lab. This was done by measuring the average HRTF spectrum magnitude of the lab KU100 in the horizontal plane and applying a single minimum phase EQ filter to all HRTFs in the database to get a matching average spectrum in the horizontal plane.

2) Filter Design for Reflected Sound: Equal weight of the estimation error in all directions was specified for the design of $\boldsymbol{F}_r(q^{-1})$ by setting all the diagonal elements of $\boldsymbol{C}(q^{-1})$ equal to the same scalar value (representing a diffuse model sound field). The plots in this section is for a filter design that implements a virtual head looking straight ahead (yaw, pitch & roll equals 0°). Fig. 4 shows the 19 calculated filter responses for the left ear design. The energy of the filter responses is distinctly located in time, with a short pre-response, suggesting that the alternative noncausal frequency-domain solution to the optimal causal filter, discussed in Section III-B1, would also be feasible for this application example.

Fig. 5 shows the NMSE according to (22) for several filter designs with different levels of regularization, obtained by varying the modeled microphone signal SNR (given by $1/\sigma_v^2$). The general trend is that the MSE increases towards high frequencies where microphone array spatial aliasing, increasingly complex HRTF beam patterns, and a limited number of microphone channels prevent accurate synthesis of the target beam pattern. Modeling a higher SNR by decreasing σ_v^2 results in better MSE-performance at low frequencies, but the price is higher

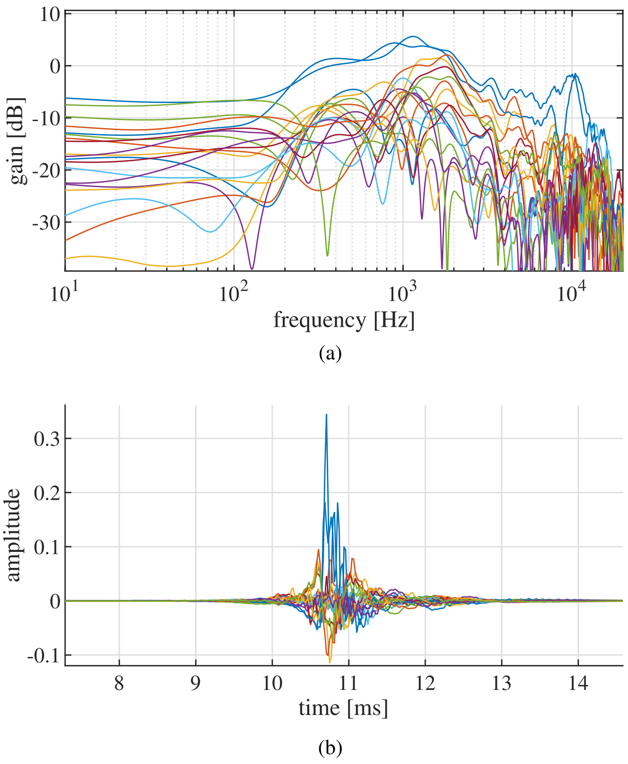


Fig. 4. (a) magnitude and (b) time responses of the reflected sound filter $\boldsymbol{F}_r(q^{-1})$, left ear.

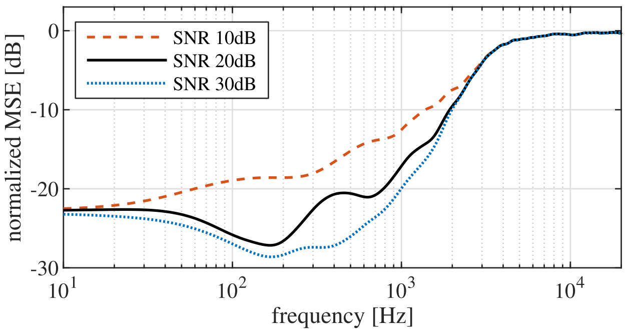


Fig. 5. NMSE for reflected sound filter $\boldsymbol{F}_r(q^{-1})$ design, left ear, for three different $SNR_m = \sigma_v^{-2}$.

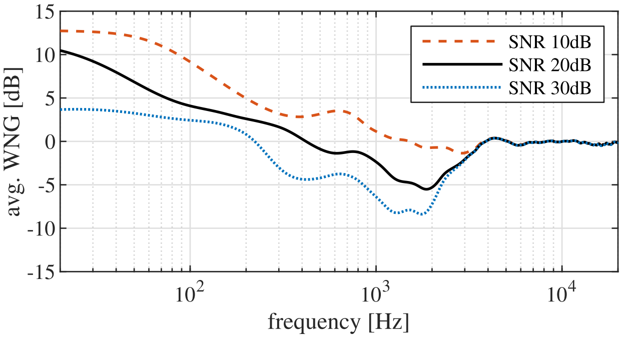


Fig. 6. $WNG_m^1(f)$ for reflected sound filter $\boldsymbol{F}_r(q^{-1})$ design, left ear, for three different $SNR_m = \sigma_v^{-2}$.

filter gains and lower WNG as can be seen in Fig. 6, which shows the average WNG as a function of the SNR parameter.

Considering the effect of the grid density on the reflected sound filter design, the grid should be dense enough to accurately capture the spatial variability of the synthesized beampatterns and of the target HRTFs. The NMSE in Fig. 5 shows that accurate HRTF beampatterns cannot be synthesized with the used array above 3–4 kHz, and we can assume, for this example, that the grid density choice is mostly critical below 3–4 kHz. A SH-decomposition of the KU100 HRTF data set shows that the spatial variability increases with frequency [23] and indicates significant energy up to an SH-order of around ten at 3–4 kHz. Since the 300-point grid can accurately represent SH basis functions of this order, it should be sufficiently dense to largely avoid spatial aliasing in the considered frequency range when sampling the target HRTFs.

The large MSE at high frequencies indicates that there will be wave-form level errors in the reflected part of the modeled BRIRs at high frequencies. To discuss the perceptual impact of this, we first note that early and late reflected parts of a BRIR have different structure and perceptual effects [43], and it is of interest to consider both the modeling of distinct early reflections and diffuse late reverberation. Other studies on auralization indicate that certain wave-form level errors in the reflected part of the BRIR are perceptually benign. For example, the earlier discussed study on “hybrid Ambisonics” [25] found that the perceived quality of auralization (of the two rooms included in their listening test) ceased to improve beyond an SH-order of three, which is also the maximum SH-order that the array we use can support.

Another Ambisonics-based method for auralization presented in [31] provides some clues about perceptually important parameters of the reflected part of the BRIR. They model BRIRs based on a first-order B-format RIR, and design a B-format binaural decoder optimized to correctly reproduce statistical properties of diffuse reverb, namely interaural coherence, power spectrum and decay rate. A listening test indicates that the modeled BRIRs are perceptually very similar to reference measured BRIRs.

We likewise assume that if the sound field in the auralized room is close to diffuse, it is necessary to reproduce interaural coherence, power spectrum, and decay rate of the reflected sound part of the BRIR for a good perceptual result. We also argue that since we have already assumed in our modeling process that no information is available about the directionality of the reflected sound field, the best design choice we can make is to ensure good diffuse field properties.

The power spectrum is effectively equalized to be correct for diffuse sound fields by the power spectrum correction filter defined by (11)–(13). Fig. 7 shows the magnitude response of the power spectrum correction filter. It has a rise towards high frequencies, indicating that the MMSE-optimal filters have low gain in this frequency range (c.f. Section III-B2).

The interaural coherence for the model sound field is shown in Fig. 8 and was calculated according to (24). The resulting ear

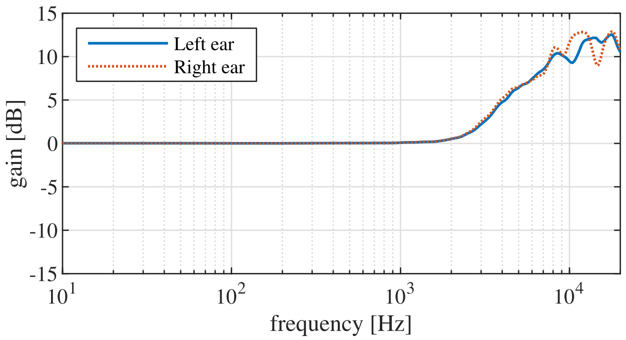


Fig. 7. Magnitude response of power spectrum correction filter $\boldsymbol{\Gamma}(q^{-1})$ for reflected sound filter $\boldsymbol{F}_r(q^{-1})$ design.

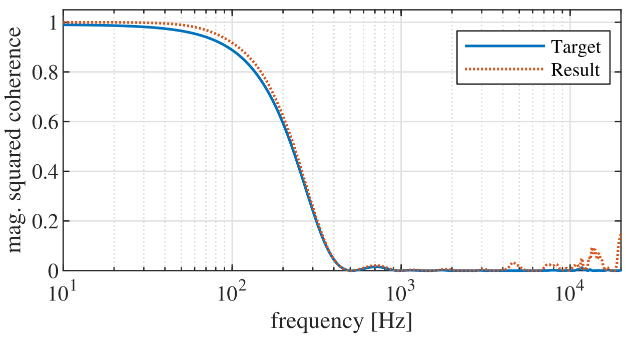


Fig. 8. Interaural coherence $C_{lr}(\omega)$, magnitude squared, of estimated and reference ear signals for the reflected sound filter $\boldsymbol{F}_r(q^{-1})$ design.

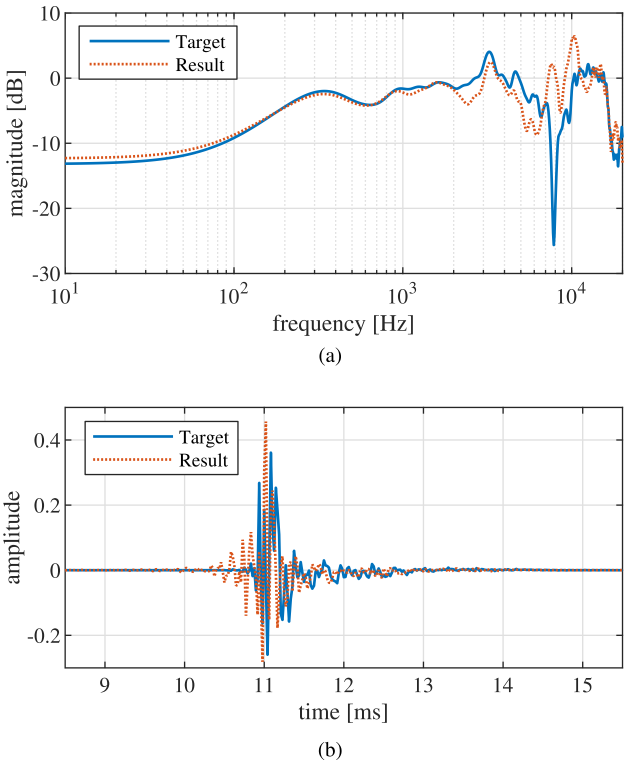


Fig. 9. (a) magnitude and (b) impulse array response in loudspeaker direction (0° azimuth, 0° elevation) for reflected sound filter design, HRTF target $\boldsymbol{S}(q^{-1})$ vs. result $\boldsymbol{F}_r(q^{-1})\boldsymbol{B}(q^{-1})$ (left ear).

signal coherence is similar to the target ear signal coherence over the full frequency range, with a slight deviation primarily at low frequencies. Interestingly, the coherence gets modeled correctly (i.e., has a low value) also above the spatial aliasing frequency of the array where the MSE is high. The synthesized beampatterns in this frequency range, which upon inspection look chaotic and random in character, and different for the two ears, apparently lead to uncorrelated ear signals in a diffuse sound field.

Since the MSE of the reflected sound filter design is large above 2–3 kHz, individual reflections in the modeled BRIRs have a spectral coloration at high frequencies. Fig. 9 illustrates this and shows the simulated array response (given by the expression $\boldsymbol{F}_r(q^{-1})\boldsymbol{B}(q^{-1})$) compared to the target response for frontal sound incidence. The error is relatively low up to around 2 kHz, after which the error is significant. The trend is similar for other directions. Fig. 9(b) also shows a pre-ringing of about 1 ms, representing a slight time-smearing of energy at high frequencies in the reflected sound part of the modeled BRIRs.

The coloration of individual reflections at high frequencies implies larger perceptual errors when auralizing rooms with more directional reflected sound fields. However, comparable coloration of individual reflections would also occur in the studies discussed above [25], [31], which still demonstrated good perceptual results. For many "normal" rooms, the perceived coloration may thus be small. The spectral effect of reflections is averaged over many directions, and if the spatial energy distribution of reflected sound is relatively uniform, it may be more perceptually relevant to consider the average power response of the array for all directions, rather than the spectral distortion in single directions. See e.g. [25], [44] for related discussion.

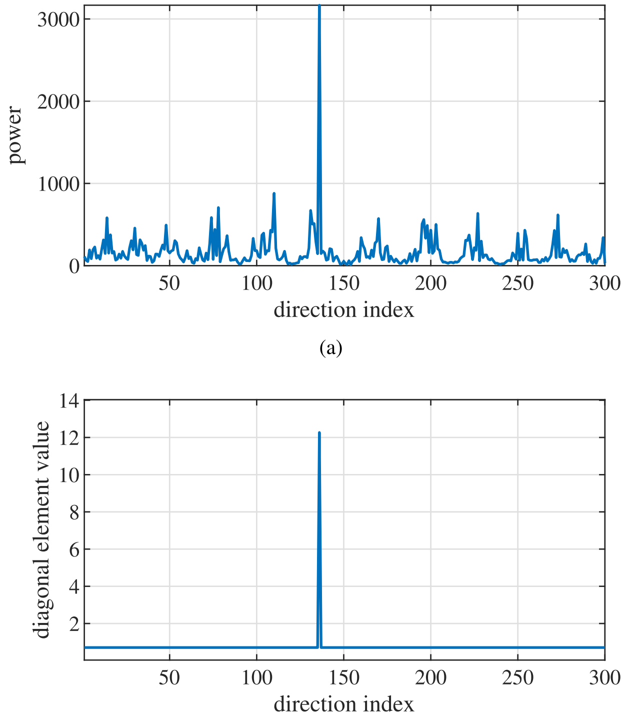
3) Filter Design for Direct Sound: For the design of $\boldsymbol{F}_d(q^{-1})$, the spatial sound power distribution $\boldsymbol{C}(q^{-1})$ of the model sound field was specified to mimic the directionality of the direct sound RIR component, $\boldsymbol{G}_d(q^{-1})$. To calculate $\boldsymbol{G}_d(q^{-1})$, a Tukey window of length 5 ms with cosine fraction 0.5 was applied to $\boldsymbol{G}(q^{-1})$.

The application example represents a simple case where the auralized speaker is full-range, close to a point-source, and the RIR $\boldsymbol{G}(q^{-1})$ was measured with the speaker in a direction close to one of the grid points used for measuring the array model $\boldsymbol{B}(q^{-1})$. For this case it is sufficient to find a single DoA of the direct sound (c.f. discussion in Section V).

A straightforward method to find the direct sound DoA, is to correlate the single column of $\boldsymbol{G}_d(q^{-1})$ with each column of the full sphere measurements in $\boldsymbol{B}(q^{-1})$ to find the best matching direction. The sound power coming from direction i is then proportional to

$$pwr_i = \left\{\boldsymbol{B}_{*(i,:)}\boldsymbol{G}_d\boldsymbol{G}_{d*}\boldsymbol{B}_{(:,i)}\right\}_{deg=0}. \tag{28}$$

This expression corresponds to a full-band version of a narrow-band formula for a conventional beamformer [45], integrating power over frequency. Evaluating (28) for all directions gave the result shown in Fig. 10(a). The DoA of the direct sound can be identified by the peak at direction index 136, corresponding to 0° azimuth, 0° elevation, as expected since $\boldsymbol{G}(q^{-1})$ was measured with the loudspeaker close to this direction.



(a)

(b)

Fig. 10. (a) Steered response power indicating DoA of direct sound. (b) Sound field model for the direct sound filter design: scalar diagonal elements of $\boldsymbol{C}(q^{-1})$.

The diagonal elements of $\boldsymbol{C}(q^{-1})$ were correspondingly selected as scalars, shown in Fig. 10(b), specifying dominant power for the identified DoA. The power for other directions was not set to exactly zero, in the hope to provide some additional design robustness without compromising the performance in the direct sound direction.

Having calculated $\boldsymbol{F}_d(q^{-1})$, the resulting simulated array magnitude and impulse response in the identified DoA (0° azimuth, 0° elevation) have insignificant error compared to target, as expected since this direction was weighted high by the selection of $\boldsymbol{C}(q^{-1})$. Minor deviations due to filter regularization are expected, and the resulting magnitude response is within +-0.5 dB of the target response over the full frequency range.

C. BRIR Evaluation

Two sets of BRIRs are evaluated in the following, using the filters designed above. The first set is referred to as $BRIR_{split}$ and was modeled according to the suggested procedure where the direct and reflected time segments of the BRIRs are modeled separately according to (1). The second set is referred to as $BRIR_{nosplit}$ and does not use separate modeling of the direct and reflected parts, which was achieved in practice by setting the filter $\boldsymbol{F}_d(q^{-1})$ equal to $\boldsymbol{F}_r(q^{-1})$ before evaluating (1).

The BRIRs were modeled for a loudspeaker direction of 0° azimuth, 0° elevation, as mentioned previously. It is of interest, however, to evaluate the BRIR modeling accuracy for a few different head poses, which puts the loudspeaker in other angles relative to the (virtual) head. Changing the head pose corresponds to changing the HRIRs that make up $\boldsymbol{S}(q^{-1})$. Rotating the coordinate system that the HRIRs are defined for in an opposite direction to the simulated head movement provides the intended result. The filter design and BRIR modeling process were thus repeated for a set of head poses in the horizontal plane: 0°, 44° and 90° yaw angle. Reference BRIRs were measured by mounting a Neumann KU100 artificial head on the measurement robot in the same position as the microphone array, turning the head to face the corresponding directions, and measuring BRIRs from the Genelec speaker.

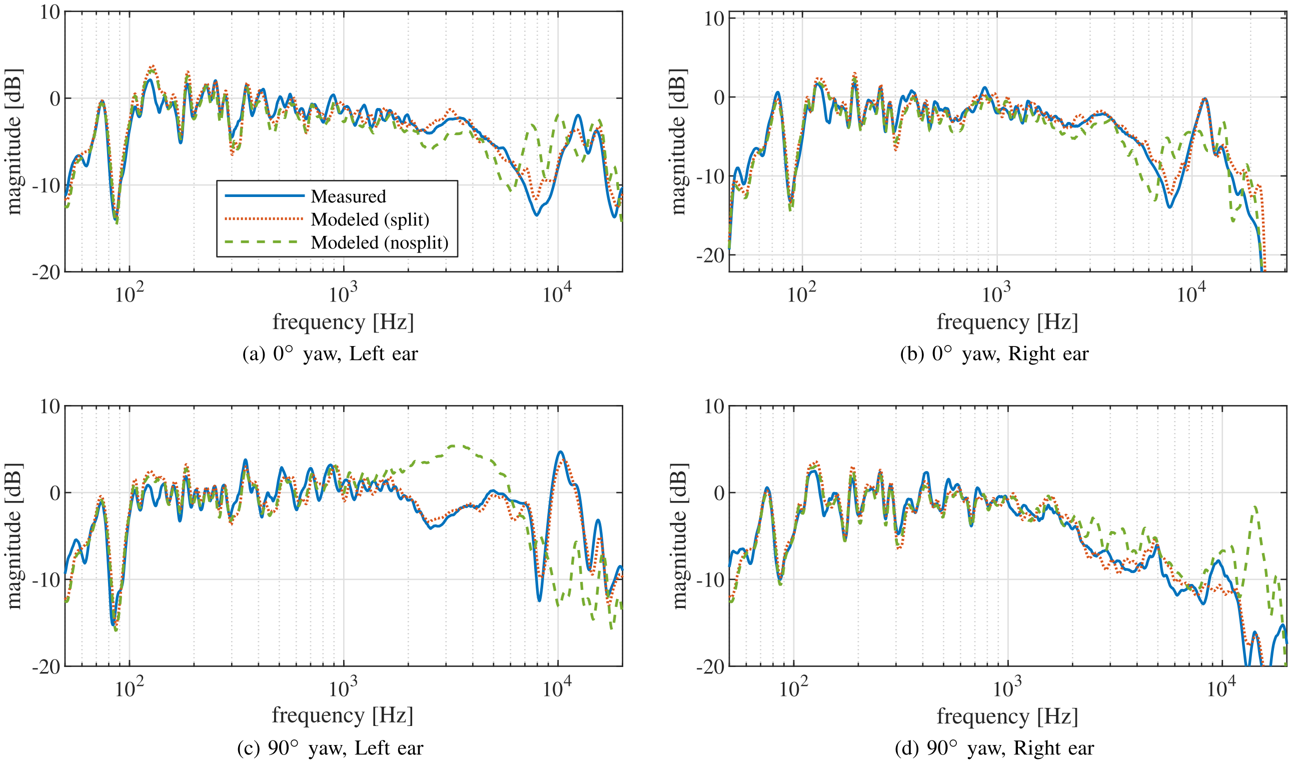
Fig. 11 shows the power spectrum of the measured and modeled BRIRs for two head yaw angles: 0° (loudspeaker straight ahead) and $-90°$ (loudspeaker to the left of the head). The measured and modeled responses correspond well for $BRIR_{split}$. Some deviations are expected due to the differences in HRTF responses between the HRTF database used and the KU100 unit used for reference measurements. Below the lowest shown frequency of 50 Hz the deviation increases, which is attributed to low measurement SNR in this frequency region due to the limited low-frequency extension of the small measurement speaker used.

For $BRIR_{nosplit}$, the modeling becomes worse above about 1.5 kHz with relatively large deviations from the measured responses. This behavior can be explained by poor high-frequency modeling of direct sound in $BRIR_{nosplit}$ and the fact that the direct sound power makes up a growing fraction of the total sound power as the frequency increases.

The detailed waveform of the first 40 ms of $BRIR_{split}$, for the 0° yaw angle and the left ear, is shown in Fig. 12(a). The BRIR envelope is reproduced approximately, with deviations due to the approximate modeling of the high-frequency BRIR reflected sound part. Fig. 12(b) shows a low-pass filtered version of the same BRIR (Butterworth, 5th order, $f_c = 1500$ Hz) which shows that the detailed BRIR waveform is reproduced accurately in the frequency range where the MSE for the reflected sound filter design is low.

To evaluate the energy decay envelope of the modeled BRIRs in $BRIR_{split}$ versus the measured BRIRs, backward integration of the mathematical square of the BRIRs was performed [46] with the result shown for the left ear, 0° yaw angle in Fig. 13. The modeled energy decay is very similar to the measured reference.

For accurate modeling of the BRIR decay envelope it is necessary to measure the RIRs in $\boldsymbol{G}(q^{-1})$ with sufficient SNR, so that the interesting part of the decay tails is not covered in measurement noise. The SNR gained or lost in the HRTF beamforming process is determined by the WNG in the design of $\boldsymbol{F}_r(q^{-1})$. The WNG for the evaluated filters is shown as the solid line in Fig. 6 and indicates a rather benign loss of SNR by at most -5 dB in a band around 1–3 kHz, and a slight increase in SNR below 400 Hz. It is concluded that it is possible to model the BRIR decay tail envelope with high precision, as long as $\boldsymbol{G}(q^{-1})$ can be measured with high SNR.



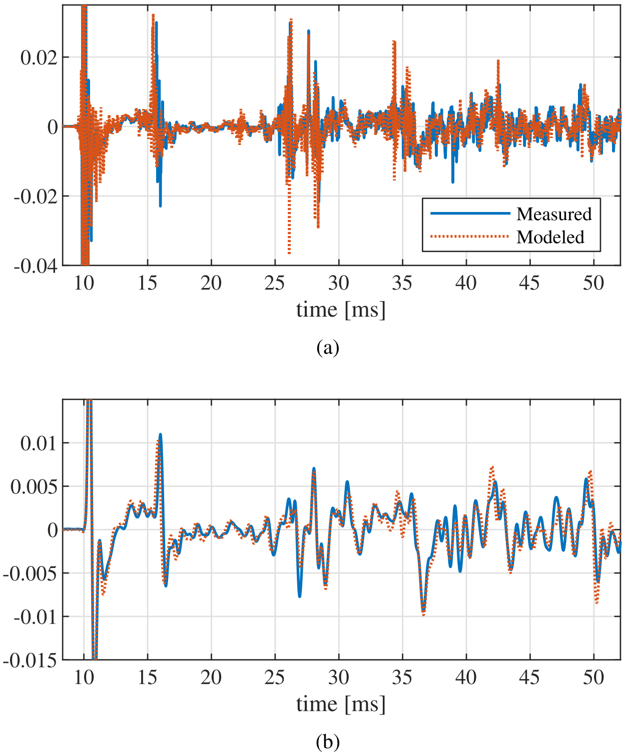
(a) 0° yaw, Left ear

(b) 0° yaw, Right ear

(c) 90° yaw, Left ear

(d) 90° yaw, Right ear

Fig. 11. Power spectrum (with 1/6 oct. smoothing) of the modeled BRIRs compared to the measured BRIRs. Two angles of head yaw are shown, plots (a) and (b) are for 0° yaw, and (c) and (d) are for 90° yaw.



(a)

(b)

Fig. 12. First 40 ms of measured vs. modeled BRIR for 0° yaw. (a) Original response, (b) low-pass filtered, $f_c = 1500$ Hz, 5th order Butterworth.

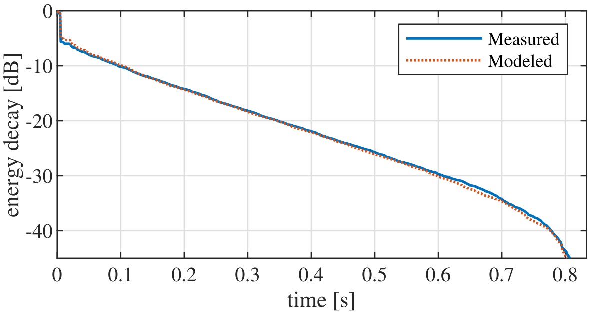


Fig. 13. Energy decay curves of measured versus modeled BRIR (for 0° yaw, left ear).

D. Listening Test

A listening test was conducted to evaluate the subjective similarity of the measured and modeled BRIRs in the application example. The listening test approximately followed ITU-R BS.1534-1 (MUSHRA). Four BRIR alternatives were evaluated in the test, including a hidden reference and anchor. The measured KU100 BRIRs were designated as reference and the anchor was a processed version of the reference with a time window applied to keep only the first 5 ms of the direct sound, setting the reflected sound part to zero. The BRIRs were labeled as follows:

- $BRIR_{split}$: modeled using the suggested procedure
- $BRIR_{nosplit}$: same binaural estimation filter for direct and reflected sound
- *Reference:* measured KU100 BRIRs

- *Anchor:* measured KU100 BRIRs, windowed

The test was run on a laptop using a GUI, operated by the test subjects themselves with no one else present, and used a Roland QuadCapture UA-55 sound card and a pair of Sennheiser HD650 headphones. The headphones were equalized to have an approximately flat frequency response as measured with the Neumann KU100 artificial head. A minimum phase FIR filter was used for the purpose, with magnitude response specified as the inverse of an average headphone magnitude response, measured by taking off and putting on the headphones on the KU100 eight times.

The subjects were asked to rate the overall similarity (including both timbral and spatial dimensions) of the different alternatives to the reference. The answers were collected using sliders in the GUI, which had a numerical range from 0 (labeled “Severe difference”) to 100 (labeled “No difference”).

To account for variation in BRIR modeling performance for different loudspeaker directions, BRIRs for the three modeled yaw angles of 0°, 44° and 90° were included in the test. Head poses in the horizontal plane were chosen since measurements of KU100 reference BRIRs could only be done with high angular precision in the horizontal plane using the available equipment, and the objective modeling performance for elevated angles is assumed to be similar.

The audio material selected for the test was music, with $\sim$ 10 s excerpts from two songs, down-mixed to mono by adding the left/right channels. The first one was *”Before You Accuse Me”* by Eric Clapton (from the 1989 Journeyman studio album), containing a broadband snare drum and distorted guitar, especially revealing of spectral and reverb decay differences. The second excerpt was from the song *”Love Over Gold”* by Dire Straits, containing piano and vocals. All audio samples were high-pass filtered at 50 Hz with a 5th order Butterworth filter to avoid influence from the previously noted low measurement SNR below this frequency and resulting differences between the measured and modeled BRIRs.

The subjects received detailed instructions before the test. They first had a training phase where they listened to the processed samples (which weren’t labeled) and compared them to the reference to learn the magnitude of the differences they were about to judge. They could also adjust the volume to a comfortable level before starting the test.

The three yaw angles and two audio samples gave each subject six trials to complete, which took on average around 15 minutes. The BRIR alternatives were labeled A–D in the GUI with randomized order for each trial and each subject. The subjects could switch instantly between listening to one of the BRIR alternatives and the reference as many times as they wanted before making a decision.

Twenty-one subjects participated in the test. In each trial, it was required that at least one BRIR alternative be given a rating of 100 due to the presence of the hidden reference. Seven of the subjects were excluded from the final results since they rated the hidden reference below 90 in one or more trials. The fourteen remaining subjects included twelve men and two women, aged 24 to 49 and with self-reported normal hearing. About half had substantial prior experience with critical listening.

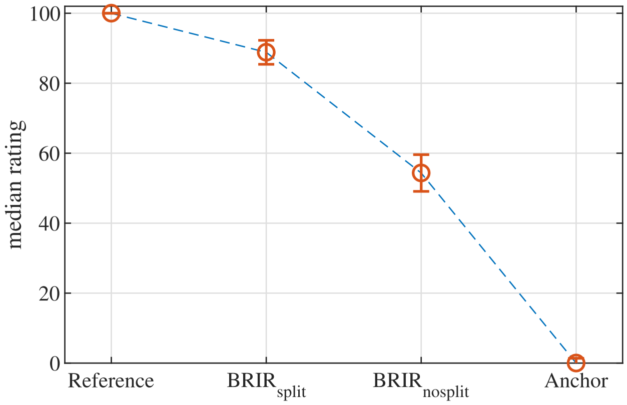


Fig. 14. Median ratings of BRIR alternatives, with regard to overall perceived difference to reference (non-normalized data). IQR-based 95% confidence intervals.

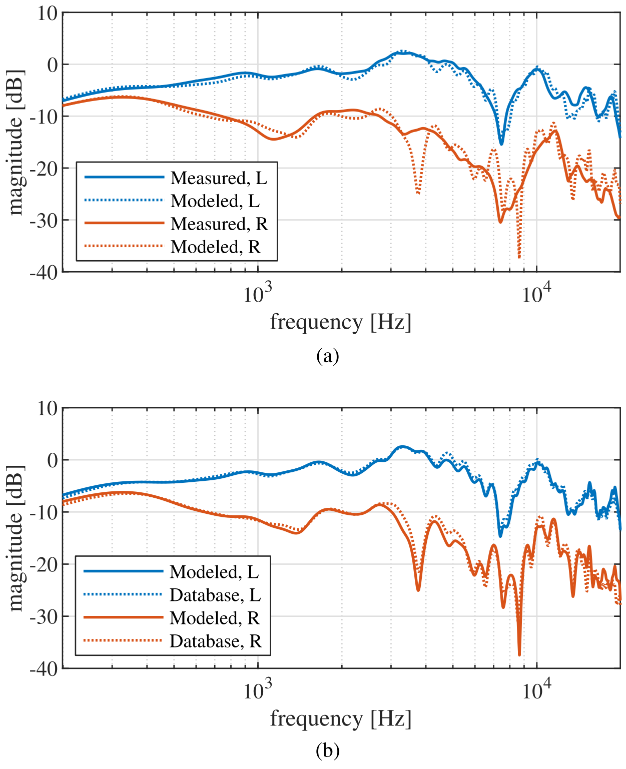
The test result in the form of median ratings for the four BRIR versions is shown in Fig. 14, together with IQR-based 95% confidence intervals [47]. The BRIRs modeled using the suggested procedure ($BRIR_{split}$) got a rating of 88.8, which indicates a relatively small difference to the reference, in comparison with BRIRs that were modeled with the same binaural estimation filter for the direct and reflected sound parts of the RIR ($BRIR_{nosplit}$) which got a rating of 54.3. Several subjects commented informally that two of the test alternatives were very similar to each other and the reference, presumably referring to $BRIR_{split}$ and the hidden reference.

V. Discussion

The listening test results can be compared to other studies on auralization based on Ambisonics. We expect that the waveform-level error in the reflected sound part of the BRIR could be made almost equally low if we had used Ambisonics-based auralization in our application example. We confirmed this informally. Without going into detail, we can note that our proposed filter design framework can be used to calculate an Ambisonics encoder filter by letting the target polynomial matrix $\boldsymbol{S}(q^{-1})$ equal spatially sampled SH basis functions and leaving the other design parameters as in the application example. And an Ambisonics binaural decoder filter can be calculated by letting the measurements $\boldsymbol{B}(q^{-1})$ equal SH basis functions and letting $\boldsymbol{S}(q^{-1})$ equal anechoic HRTFs as in the application example.

The result that $BRIR_{nosplit}$ got a low rating, representing auralization without separate rendering of the direct sound, is consistent with similar results found using Ambisonics, in e.g. [21] (noting that the array we used is limited to third-order Ambisonics). The low perceptual rating is expected due to the inferior high-frequency modeling of direct sound (c.f. Fig. 9), which leads both to severe spectral distortion of the direct sound and a failure to synthesize correct interaural level and time differences.

The higher rating of $BRIR_{split}$ is fully explained by the improved modeling of the direct sound. That the perceptual difference to reference is small is consistent with the results of



(a)

(b)

Fig. 15. Magnitude response comparisons between: (a) the direct sound parts of the measured BRIR (*Measured*) and the modeled BRIR (*Modeled*), for 44° yaw. (b) the direct sound part of the modeled BRIR and the corresponding HRTF target response from the HRTF database used (*Database*).

e.g. [25]. There it was demonstrated that a third-order Ambisonics rendering of reverb can be enough for a good perceptual result in auralization, when the direct sound is rendered separately.

To explain the remaining perceptual difference between $BRIR_{split}$ and the reference, we analyze the direct and reflected sound parts of the modeled BRIRs separately for the 44° yaw angle.

Fig. 15(a) shows the magnitude response of the direct sound part (using a 4 ms window) and compares the measured KU100 reference to the modeled BRIR. A significant difference can be seen, which will have affected the listening test. The difference stems mainly from differences between the database KU100 HRTFs used in the binaural estimation filter design and our reference KU100 artificial head. The database HRIRs contain some very early reflections that are visible as wiggles in the HRTF magnitude response, and the modeled BRIRs consequently also contain these reflections. Fig. 15(b) compares the modeled BRIR direct sound part to the target HRTF response for the identified loudspeaker direction, and the modeled BRIR response is very close to the target.

The reflected sound spectrum of the modeled BRIRs has slightly more energy at high frequencies compared to the measured BRIRs, as illustrated in Fig. 16 for the 44° yaw angle. The difference of around 2 dB above 4 kHz was similar for the other yaw angles. The reason for the difference remains unclear, but we note that the power spectrum correction filter $\boldsymbol{\Gamma}(q^{-1})$ for the reflected sound filter $\boldsymbol{F}_r(q^{-1})$ is designed assuming an isotropic sound field, and the reflected sound field in our auralized room likely has some directivity, c.f. [48].

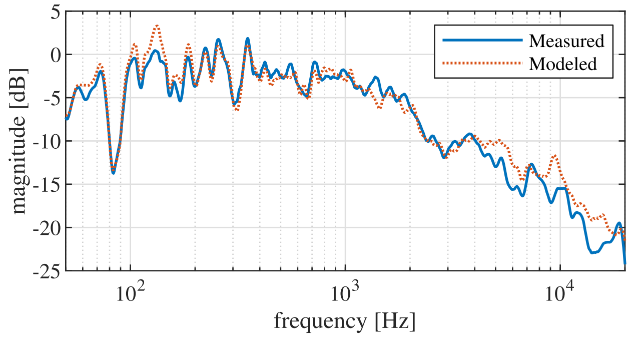


Fig. 16. Magnitude response comparison between the reflected sound parts of the measured BRIR (*Measured*) and the modeled BRIR (*Modeled*), for 44° yaw (1/6 oct. smoothing).

The above discussion attempts to explain why $BRIR_{split}$ was audibly different from the reference. Of course, the listening test only evaluated the overall difference to the reference. It does not provide insight into specific perceptual aspects like timbre, spaciousness, or perceived direction, but it does support the general validity of the BRIR modeling method.

In the application example, the auralized loudspeaker direction almost coincides with one of the grid points used in the binaural estimation filter design. Our choice to weight a single direction highly in the direct sound binaural estimation filter design thus works well. In Fig. 15(b), a small difference can still be seen between the direct sound part of the modeled BRIR and the target HRTF for the corresponding direction, presumably because the auralized speaker direction deviated slightly from the estimated DoA.

In more general cases, the auralized loudspeaker may be in any position between grid points. DoA-estimation by evaluating (28) then gives multiple peaks corresponding to grid points close to the loudspeaker direction. The loudspeaker also may not resemble a point source. More research is needed to find a robust way to specify $\boldsymbol{C}(q^{-1})$ in these cases and select an appropriate grid resolution. A hypothesis is that the synthesized beampattern can be made to approach the target also in between grid points by specifying dominant sound power in $\boldsymbol{C}(q^{-1})$ for grid points covering an angular region that includes the loudspeaker direction. This would, however, require a dense enough grid to avoid over-fitting to the grid points.

Lastly, the suggested separate modeling of direct and reflected sound is not feasible if a direct sound RIR time segment can not be identified. That may happen for e.g. sound systems in automotive cabins with multiple reflecting surfaces close to the speakers and the listening position.

VI. Conclusion

A new filter design framework for estimating binaural signals from microphone array signals was presented, using a causal Wiener filtering formulation. A general problem formulation was used where the waveform-level error of the binaural signal estimate is minimized under the conditions of a specified sound field spatial energy distribution and a microphone noise model. Compared to Ambisonics-based methods, there is no restriction on the geometry the problem is defined in, and directional weighting of the estimation error and the use of general microphone array geometries is straightforward.

A BRIR modeling method was also presented that demonstrates the use of the proposed filter design framework and describes a complete process for acquiring BRIRs from microphone array RIR measurements, minimizing perceptual and waveform-level differences to measured BRIRs. Direct and reflected sound is modeled separately. The method facilitates fast BRIR data set acquisition for dynamic binaural synthesis and is a viable alternative to Ambisonics-based binaural room auralization.

An application example provided experimental validation of the BRIR modeling procedure and used a 19-channel SMA. A listening test indicated a small perceptual difference between measured versus modeled BRIRs. Our results complement recent research on Ambisonics-based auralization, which likewise demonstrated good perceptual results using microphone arrays of similar complexity as the one we used when rendering the direct sound separately with high accuracy. We showed that the waveform-level error of the reflected sound part of the BRIR can be made low up to at least 1.5 kHz with the chosen array. At higher frequencies, we aimed for correct statistical properties of the diffuse part of the reflected sound field. The listening test results indicated that this was sufficient for a low perceptual error for the room we used. More research is needed to find to which degree this generalizes to rooms with more directional reflected sound fields, e.g. small rooms with salient early reflections.

Future research could also investigate improved robust modeling of the direct sound for more demanding room acoustical scenarios, as discussed in the previous section, and benchmark perceptual performance with parametric methods like SDM. Other applications for the presented filter design framework could also be investigated, like binaural rendering of general microphone array recordings.

Acknowledgment

The authors wish to thank Dirac Research for supporting this work, with special thanks to Tomas Olofsson for comments and feedback. We are grateful also for the helpful comments of the anonymous reviewers.

References

[1] W. Zhang, P. Samarasinghe, H. Chen, and T. Abhayapala, "Surround by sound: A review of spatial audio recording and reproduction," Appl. Sci., vol. 7, no. 5, May 2017, Art. no. 532.

[2] H. Møller, "Fundamentals of binaural technology," *Appl. Sci.*, vol. 36, pp. 171–218, Dec. 1992.

[3] A. Lindau, "Binaural resynthesis of acoustic environments. technology and perceptual evaluation," Ph.D. dissertation, Beuth University of Applied Sciences Berlin, Jun. 2014.

[4] M. Kleiner, B. Dalenbäck, and P. Svensson, "Auralization–an overview," *J. Audio Eng. Soc.*, vol. 41, no. 11, pp. 861–875, Nov. 1993.

[5] P. Mackensen, U. Felderhof, G. Theile, U. Horbach, and R. Pellegrini, "Binaural room scanning – a new tool for acoustic and psychoacoustic research," *J. Acoustical Soc. Amer.*, vol. 105, no. 2, pp. 1343–1344, Jan 1999.

[6] A. Lindau and S. Weinzierl, "FABIAN - An instrument for software-based measurement of binaural room impulse responses in multiple degrees of freedom," in *Proc. 24th Tonmeistertagung*, Jan 2006, pp. 621–625.

[7] H. Møller, M. Sørensen, C. Jensen, and D. Hammershøi, "Binaural technique: Do we need individual recordings?," *J. Audio Eng. Soc.*, vol. 44, pp. 451–464, Jun. 1996.

[8] E. Rasumow, "Synthetic reproduction of head-related transfer functions by using microphone arrays," Ph.D. dissertation, School of Medicine and Health Sciences, University of Oldenburg, 2015.

[9] B. Bernshütz, "Microphone arrays and sound field decomposition for dynamic binaural recording," Ph.D. dissertation, University of Technology, Berlin, 2016.

[10] C. D. Salvador, S. Sakamoto, J. Treviño, and Y. Suzuki, "Design theory for binaural synthesis: Combining microphone array recordings and head-related transfer function datasets," *Acoustical Sci. Technol.*, vol. 38, no. 2, pp. 51–62, Mar. 2017.

[11] L. Madmoni, J. Donley, V. Tourbabin, and B. Rafaely, "Beamforming-based binaural reproduction by matching of binaural signals," in *Proc. AES Int. Conf. Audio Virtual Augmented Reality*, Aug. 2020. [Online]. Available: https://www.aes.org/e-lib/browse.cfm?elib=20878

[12] J. Chen, B. Van Veen, and K. Hecox, "External ear transfer function modeling: A beamforming approach," *J. Acoustical Soc. Amer.*, vol. 92, pp. 1933–44, Nov. 1992.

[13] Z. Li and R. Duraiswami, "Headphone-based reproduction of 3D auditory scenes captured by spherical/hemispherical microphone arrays," *Proc. IEEE Int. Conf. Acoustical, Speech, Signal Process.*, vol. 5, pp. 337–340, 2006.

[14] E. G. Williams, *Fourier Acoustics*. London, U.K.: Academic Press, 1999.

[15] F. Zotter and M. Frank, *Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*. Berlin, Germany: Springer, 2019.

[16] L. Brännmark, A. Bahne, and A. Ahlen, "Compensation of loudspeaker-room responses in a robust MIMO control framework," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, pp. 1201–1216, Jun. 2013.

[17] S. Widmark, "Causal IIR audio precompensator filters subject to quadratic constraints," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1897–1912, Oct. 2018.

[18] B. Bernhardsson and M. Sternad, "Feedforward control is dual to deconvolution," *Int. J. Control*, vol. 57, no. 2, pp. 393–405, 1993.

[19] O. Kirkeby, P. A. Nelson, H. Hamada, and F. Orduna-Bustamante, "Fast deconvolution of multichannel systems using regularization," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 2, pp. 189–195, Mar. 1998.

[20] B. Rafaely, "Analysis and design of spherical microphone arrays," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 135–143, Jan. 2005.

[21] J. Ahrens and C. Andersson, "Perceptual evaluation of headphone auralization of rooms captured with spherical microphone arrays with respect to spaciousness and timbre," *J. Acoustical Soc. Amer.*, vol. 145, pp. 2783–2794, Apr. 2019.

[22] T. Lübeck, H. Helmholz, J. Arend, C. Pörschmann, and J. Ahrens, "Perceptual evaluation of mitigation approaches of impairments due to spatial undersampling in binaural rendering of spherical microphone array data," *J. Audio Eng. Soc.*, vol. 68, pp. 428–440, Jul. 2020.

[23] M. Zaunschirm, C. Schörkhuber, and R. Höldrich, "Binaural rendering of ambisonic signals by head-related impulse response alignment and a diffuseness contraint," *J. Acoustical Soc. Amer.*, vol. 143, no. 6, pp. 3616–3627, June 2018.

[24] C. Schörkhuber, M. Zaunschirm, and R. Höldrich, "Binaural rendering of ambisonic signals via magnitude least squares," in *Proc. DAGA*, Mar. 2018, pp. 339–342.

[25] I. Engel, C. Henry, S. V. Amengual Garí, P. W. Robinson, and L. Picinali, "Perceptual implications of different ambisonics-based methods for binaural reverberation," *J. Acoust. Soc. Amer.*, vol. 149, no. 2, pp. 895–910, Feb. 2021.

[26] S. Tervo, J. Pätynen, A. Kuusinen, and T. Lokki, "Spatial decomposition method for room impulse responses," *J. Audio Eng. Soc.*, vol. 61, pp. 16–27, Jan. 2013.

[27] M. Zaunschirm, M. Frank, and F. Zotter, "Binaural rendering with measured room responses: First-order ambisonic microphone vs. dummy head," *Appl. Sci.*, vol. 10, no. 5, p. 1631, Feb. 2020.

[28] J. Merimaa and V. Pulkki, "Spatial impulse response rendering I: Analysis and synthesis," *J. Audio Eng. Soc.*, vol. 53, no. 12, pp. 1115–1127, Dec. 2005.

[29] L. McCormack, V. Pulkki, A. Politis, O. Scheuregger, and M. Marschall, "Higher-order spatial impulse response rendering: Investigating the perceived effects of spherical order, dedicated diffuse rendering, and frequency resolution," *J. Audio Eng. Soc.*, vol. 68, no. 5, pp. 338–354, May 2020.

[30] P. Stade, J. Arend, and C. Pörschmann, “A parametric model for the synthesis of binaural room impulse responses,” *Proc. Meetings Acoust. 173EAA*, vol. 30, 2017, Art. no. 015006.

[31] F. Menzer, C. Faller, and H. Lissek, “Obtaining binaural room impulse responses from B-format impulse responses using frequency-dependent coherence matching,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, pp. 396–405, Feb. 2011.

[32] S. V. Amengual Garí, O. Brimijoin, H. Hassager, and P. Robinson, “Flexible binaural resynthesis of room impulse responses for augmented reality research,” in *Proc. EAA Spatial Audio Signal Process. Symp.*, Sep. 2019, pp. 161–166.

[33] V. R. Algazi, R. O. Duda, and D. M. Thomson, “Motion-tracked binaural sound,” *J. Audio Eng. Soc.*, vol. 52, no. 11, pp. 1142–1156, Nov. 2004.

[34] T. Kailath, *Linear Systems*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1980.

[35] V. Kučera, *Analysis and design of discrete linear control systems*. Hemel Hempstead: Prentice-Hall, 1991.

[36] A. Ahlén and M. Sternad, “Wiener filter design using polynomial equations,” *IEEE Trans. Signal Process.*, vol. 39, no. 11, pp. 2387–2399, Nov. 1991.

[37] J. Ježek and V. Kučera, “Efficient algorithm for matrix spectral factorization,” *Automatica*, vol. 21, no. 6, pp. 663–669, 1985.

[38] A. Barkefors, M. Sternad, and L.-J. Brännmark, “Design and analysis of linear quadratic gaussian feedforward controllers for active noise control,” *IEEE Trans. Speech Audio Process.*, vol. 22, no. 12, pp. 1777–1791, Dec. 2014.

[39] E. Rasumow, M. Blau, M. Hansen, S. Doclo, S. Par, V. Mellert, and D. Püschel, “The impact of the white noise gain (WNG) of a virtual artificial head on the appraisal of binaural sound reproduction,” in *Proc. EAA Joint Symp. Auralization Ambisonics*, Berlin, Germany, Apr. 2014, pp. 174–180.

[40] S. Müller and P. Massarani, “Transfer-function measurement with sweeps,” *J. Audio Eng. Soc.*, vol. 49, no. 6, pp. 443–471, Jun. 2001.

[41] M. Deserno, “How to generate equidistributed points on the surface of a sphere,” May 2004, Accessed: Jun. 2020, [Online]. Available: https://www.cmu.edu/biolphys/deserno/pdf/sphere_equi.pdf.

[42] B. Bernschütz, “A spherical far field HRIR/HRTF compilation of the Neumann KU 100,” in *Proc. AIA-DAGA Conf. Acoust.*, 2013, pp. 592–595.

[43] F. Toole, “Loudspeakers and rooms for sound reproduction–a scientific review,” *J. Audio Eng. Soc.*, vol. 54, pp. 451–476, Jan. 2012.

[44] T. Lübeck, C. Pörschmann, and J. M. Arend, “Perception of direct sound, early reflections, and reverberation in auralizations of sparsely measured binaural room impulse responses,” in *Proc. AES Int. Conf. Audio Virtual Augmented Reality*, Aug. 2020. [Online]. Available: https://www.aes.org/e-lib/browse.cfm?elib=20878

[45] P. Stoica and R. Moses, *Spectral Analysis of Signals*. Upper Saddle River, NJ, USA: Prentice-Hall, 2005.

[46] M. R. Schroeder, “New method of measuring reverberation time,” *J. Acoust. Soc. Amer.*, vol. 37, no. 3, pp. 409–412, 1965.

[47] J. Chambers, W. Cleveland, B. Kleiner, and P. Tukey, “Notched box plots,” in *Graphical Methods Data Anal.*, MIT Press, 1983, ch. 3-4, pp. 60–63.

[48] B. N. Gover, J. G. Ryan, and M. R. Stinson, “Measurements of directional properties of reverberant sound fields in rooms using a spherical microphone array,” *J. Acoust. Soc. Amer.*, vol. 116, no. 4, pp. 2138–2148, 2004.

Viktor Gunnarsson (Student Member, IEEE) received the M.Sc. degree in sound and vibration and the B.Sc. degree in electrical engineering from the Chalmers University of Technology, Gothenburg, Sweden, in 2010. Since then, he has been with Dirac Research AB, Uppsala, Sweden, as an Engineer, Inventor, and Researcher in the field of digital sound optimization. In parallel, he is currently working toward the Ph.D. degree in signal processing with Uppsala University, Uppsala, Sweden.

His research interests include spatial audio perception and reproduction, sound system simulation, acoustics and psychoacoustics of loudspeakers and rooms, and in general the philosophy of sound reproduction and the pursuit of the perfect sound.

Mikael Sternad (Senior Member, IEEE) received the Ph.D. degree in automatic control from Uppsala University, Uppsala, Sweden, in 1987. He is currently a Professor of automatic control with the Department of Electrical Engineering, Uppsala University.

His research focuses on signal processing applied to mobile broadband communication problems, such as channel prediction schemes for fast link adaptation, scheduling and coordinated multipoint transmission. He has acted as a Project Leader of the national 4G research project Wireless IP and the Swedish Research Council Framework project Dynamic Multipoint Transmission, and also several VINNOVA projects. He was engaged in the EU WINNER which formed the basis for the 4G wireless standardization effort, the Artist4G project, and the EU FP7 project METIS.

He is also working on sound field control, acoustic zones and personal audio, 3D sound and recording and rendering for virtual reality applications. He is Co-Founder and was the Chairman (during 2001–2005) of Dirac Research AB which is active in these fields. A research interest is robust and adaptive MIMO feedforward control algorithms. These can be applied to wireless transmission (network MIMO, or CoMP), and also to sound field control using multiple loudspeakers.