Identification and matching of room acoustics with moving head-worn microphone arrays

Deppisch *et al.*
Journal on Audio, Speech, and Music Processing (2026) 2026:30
https://doi.org/10.1186/s13636-026-00466-1

Journal on Audio,
Speech, and Music Processing

EMPIRICAL RESEARCH

Open Access

Identification and matching of room acoustics with moving head-worn microphone arrays

Thomas Deppisch[1*], Sebastià V. Amengual Garí[2], Paul Calamia[2] and Jens Ahrens[1]

Abstract

Head-worn devices such as smartglasses and headsets are the predominant form factor for augmented reality and telepresence applications, where real-world environments are augmented with virtual sound sources. For these sources to appear perceptually convincing, the acoustics of their virtual environment must closely match the room acoustics of the physical space. Estimating room impulse responses (RIRs) in this setting is challenging because practical scenarios require small arrays that continuously move with the user's head. This study presents a method for the blind identification of RIRs from speech signals captured with a moving head-worn microphone array and the subsequent rendering of virtual sound sources based on perceptually relevant acoustic parameter estimates. A motion-aware signal model that estimates spatial RIRs as sound field coefficients and incorporates position tracking data is compared against an omnidirectional model and a baseline. Numerical results show that the motion-aware model provides the most accurate acoustic parameter estimates when used in an informed setting with the true reference signal. In the blind setting, however, its advantage largely diminishes. In a listening experiment, renderings based on the omnidirectional model are rated as most similar to the reference condition and are most often associated with the correct room significantly above chance. The findings highlight the practical relevance of the proposed framework, with the omnidirectional model offering robust and perceptually convincing performance, while the motion-aware model remains promising for parameter estimation.

Keywords Acoustics, Augmented reality, Binaural, Microphone array, Room impulse response, Smartglasses

1 Introduction

The growing adoption of smartglasses and virtual/augmented reality (VR/AR) headsets has accelerated the integration of head-worn microphone arrays into consumer devices. These devices enable audio capture from the user's perspective and playback through near-ear loudspeakers, allowing for augmentation of real environments with virtual sounds. For convincing augmentation, the acoustics of the real and virtual environments must match perceptually, so that virtual sound sources blend seamlessly with the physical scene [1]. Without accurate acoustic matching, externalization, distance perception, and the overall plausibility of virtual sources are impaired [2–4]. Among the most critical parameters for perceptual matching are the reverberation time (RT) and measures of early-to-late energy such as the direct-to-reverberant energy ratio (DRR) [1, 5, 6].

In practice, the acoustic properties of the user's environment are unknown, and dedicated measurements are infeasible. Instead, naturally occurring sounds, such as speech, can be used to estimate acoustic parameters [7–11] or room impulse responses (RIRs) [12–19]. While RIRs have the advantage of capturing all linear, time-invariant acoustic properties at once, their estimation

*Correspondence:
Thomas Deppisch
thomas.deppisch@chalmers.se
[1] Chalmers University of Technology, Gothenburg 412 96, Sweden
[2] Reality Labs Research, Meta, Redmond 98052, WA, USA

is challenging due to long acoustically relevant RIR lengths of several hundred milliseconds, which impedes the convergence of conventional blind estimation techniques [20].

Recent work has addressed these challenges with a robust two-stage framework for blind multichannel RIR estimation [20–22]. The method first constructs a pseudo-reference signal via dereverberation and beamforming, and then estimates the RIR with a multichannel Wiener or recursive least-squares (RLS) filter. The approach converges reliably in static conditions and outperforms conventional direct parameter estimation. A study on smartglasses recordings demonstrated its feasibility and perceptually effective matching performance in stationary scenarios [23].

Instead of dividing the room acoustic matching problem into estimation and rendering, some recent machine-learning approaches jointly tackle both steps, either analyzing microphone signals [24–26] or by exploiting visual information [27, 28]. Head-worn microphone arrays pose a unique challenge for acoustic parameter estimation and matching algorithms due to their movement with the user's head. The aforementioned works, however, only consider scenarios with stationary microphones, and room acoustic matching with signals from a moving array has not been attempted thus far.

Existing work involving RIR and sound field estimation from moving microphones primarily considers controlled setups to measure and reconstruct RIRs over a spatial grid from one or several moving microphones while using a known, sometimes specifically designed excitation signal [29–35]. These approaches do not consider head-worn arrays, natural movement, or blind estimation. A few studies have modeled signals in the spherical harmonic domain, updating coefficients across microphone positions [33, 35].

In related works, the authors recently proposed an RLS filter that supports the estimation of multichannel RIRs as sound field coefficients in dynamic conditions by incorporating tracking data for array orientation and position [36, 37]. In this approach, RIRs are represented as circular harmonic (CH) sound field coefficients that remain valid in a spatial region around the array, accommodating both rotations and translations. A feasibility study using a controlled setup with an equatorial microphone array moving along a line demonstrated that accurate RIR estimation is achievable at low and mid frequencies with the performance depending on array displacement and the maximum order of the estimated sound field coefficients. These studies however only considered the informed estimation problem, assuming a known reference signal, and did not attempt perceptually valid matching.

The present work extends and combines the framework from [20] and the method from [37] to support realistic use cases with head-worn arrays and natural head movements without requiring analytic array models, and enables the rendering of binaural virtual sources based on estimated acoustic parameters. Two signal models are considered: a motion-aware model that leverages position tracking information and estimates sound field coefficients, and an omnidirectional model that estimates an omnidirectional RIR without requiring tracking. The results are evaluated numerically in terms of the estimation performance of typical acoustic parameters and perceptually through a listening experiment that assesses the perceptual quality of virtual sound sources rendered using the estimated parameters. To our knowledge, this is the first demonstration of blind room acoustic matching from moving head-worn microphone arrays.

2 Array signal models

This section introduces a signal model for estimating spatial room transfer function coefficients during array motion. The model is designed to satisfy two requirements. First, it represents the sound pressure measured within a region around the coordinate origin in terms of a set of spatial room transfer function coefficients defined at the origin. Second, it separates the contributions of the spatial array transfer function, the spatial room transfer function, and the source signal, thereby enabling an array-independent estimate of the spatial room impulse response. As a simpler alternative, an omnidirectional signal model is also proposed that models the omnidirectional sound pressure while ignoring the array motion.

2.1 Motion-aware signal model

To derive such a representation, it is useful to formulate the problem from a sound field perspective. Any source- and sink-free interior sound field can be described by a superposition of infinitely many plane waves with a corresponding continuous plane-wave amplitude density. This plane-wave density is commonly decomposed into a radial and an angular component, where the angular component is often expanded in orthogonal basis functions, such as spherical harmonics (SHs) on the sphere or circular harmonics (CHs) on the circle [38].

Head-worn microphone arrays are typically embedded in devices such as smartglasses or head-mounted displays, which constrain microphone placement to a limited aperture, usually on a single circumferential contour around the front and sides of the head. When microphones are approximately positioned on a circumferential contour around the head, CHs

$$C_m(\phi) = e^{\mathrm{i}m\phi}\,, \tag{1}$$

provide a suitable basis [39]. Employing the plane-wave decomposition in the CH domain, the sound pressure at the angular frequency ω and a position in the 2D plane defined by radius r and angle ϕ due to a source s is given by [38]

$$p(\omega, r, \phi) = s(\omega) \sum_{m=-\infty}^{\infty} C_m(\phi)\, b_m(\omega r)\, a_m(\omega)\,. \tag{2}$$

Here, a_m are the CH coefficients representing the directional plane wave density and b_m are radial terms. After limiting the CH coefficients to order $N \in \mathbb{N}_0$ such that their degree $m \in [-N, N]$, the CH-domain sound field coefficients at a translated and rotated origin described by position $(\phi_\mathrm{t}, r_\mathrm{t})$ and azimuthal rotation α are [37]

$$\boldsymbol{a}_\mathrm{t}(\omega, \alpha, \phi_\mathrm{t}, r_\mathrm{t}) = \boldsymbol{R}(\alpha)\, \boldsymbol{T}(\omega, \phi_\mathrm{t}, r_\mathrm{t})\, \boldsymbol{a}(\omega)\, s(\omega)\,, \tag{3}$$

where the CH coefficients for all degrees up to order N are stacked in the vectors $\boldsymbol{a} \in \mathbb{C}^{2N+1}$ and $\boldsymbol{a}_\mathrm{t} \in \mathbb{C}^{2N+1}$ for the original and the translated coefficients, respectively. The array translation is expressed using the translation matrix $\boldsymbol{T} \in \mathbb{C}^{2N+1\times 2N+1}$ containing the elements $T_{m,\mu} = C_{\mu-m}(\phi_\mathrm{t})\, \mathrm{i}^{\mu-m}\, \mathrm{J}_{\mu-m}(\mathrm{k}r_\mathrm{t})$ at row index $m \in [-N, N]$ and column index $\mu \in [-N, N]$, where J_m are the Bessel functions. The diagonal rotation matrix $\boldsymbol{R} \in \mathbb{C}^{2N+1\times 2N+1}$ contains $e^{\mathrm{i}m\alpha}$ terms on its main diagonal.

Given the CH-domain coefficients of the rotated, translated sound field in (3), the sound pressure captured by a microphone array is obtained by applying the spatial array response matrix $\boldsymbol{D} \in \mathbb{C}^{M\times 2N+1}$,

$$\boxed{\boldsymbol{p}(\omega, \alpha, \phi_\mathrm{t}, r_\mathrm{t}) = \boldsymbol{D}(\omega)\, \boldsymbol{a}_\mathrm{t}(\omega, \alpha, \phi_\mathrm{t}, r_\mathrm{t})\,.} \tag{4}$$

The spatial array response matrix $\boldsymbol{D}$ maps the translated CH coefficients to the microphone domain while incorporating the spatial array transfer function, accounting for array geometry, sensor characteristics, and scattering effects of the (head-worn) array. This signal model describes the sound pressure for a rotated, translated array and is valid in a source- and sink-free region around the coordinate origin. It separates acoustic contributions of the microphone array via the spatial array transfer function captured in $\boldsymbol{D}$, the room acoustics via the spatial room transfer function $\boldsymbol{a}$, and the source signal s. The entries of $\boldsymbol{D}$ have also been described as wearable-device-related transfer function coefficients [40]. This signal model is referred to as the motion-aware signal model in the following.

Although this signal model only supports 2D sound fields, it has been shown to be applicable for movement in a 2D plane within a 3D sound field [37]. CH coefficients obtained from the array then represent a horizontal projection of the sound field [39]. While the signal model can similarly be formulated using SHs and corresponding rotation and translation theorems to describe movement in a 3D sound field, the increased number of $(N+1)^2$ SH instead of $2N+1$ CH coefficients requires a significantly larger amount of microphones distributed around the full head, which prevents a practical application of the method with current head-worn microphone arrays.

2.2 Spatial array response matrix for arbitrary arrays

In [37], the spatial array response matrix was analytically described for an equatorial microphone array,

$$\boldsymbol{D}_\mathrm{EMA}(\omega) = \boldsymbol{C}_\mathrm{M}\, \mathrm{diag}\{\boldsymbol{b}(\omega)\}\,, \tag{5}$$

where $\boldsymbol{C}_\mathrm{M} \in \mathbb{C}^{M\times 2N+1}$ contains the CHs evaluated at the M microphone directions, and $\boldsymbol{b} \in \mathbb{C}^{2N+1}$ contains radial terms for a rigid sphere. In the present work, this concept is extended to accommodate arbitrary arrays for which array transfer functions (ATFs) are available. For head-worn microphone arrays, the array response matrix additionally accounts for the influence of the wearer's head on the recorded sound field.

To estimate $\boldsymbol{D}$ from anechoic ATFs, we describe the pressure $\boldsymbol{p} \in \mathbb{C}^M$ picked up by the array in response to a plane wave from direction ϕ_pw in the CH domain [41],

$$\boldsymbol{p}(\omega, \phi_\mathrm{pw}) = \boldsymbol{D}(\omega)\, \boldsymbol{c}^*(\phi_\mathrm{pw})\,, \tag{6}$$

where $\boldsymbol{c} \in \mathbb{C}^{2N+1}$ contains the CHs, and $(\cdot)^*$ denotes the complex conjugate. For D ATF measurement directions, the ATFs $\boldsymbol{P}_\mathrm{A} \in \mathbb{C}^{M\times D}$ are then described by

$$\boldsymbol{P}_\mathrm{A}(\omega) = \boldsymbol{D}(\omega)\boldsymbol{C}_\mathrm{D}^\mathrm{H}\,, \tag{7}$$

and the spatial array response matrix can be estimated using the least-squares-optimal solution from [41]

$$\hat{\boldsymbol{D}}(\omega) = \boldsymbol{P}_\mathrm{A}(\omega)\, \boldsymbol{W}\boldsymbol{C}_\mathrm{D} \left(\boldsymbol{C}_\mathrm{D}^\mathrm{H} \boldsymbol{W}\boldsymbol{C}_\mathrm{D}\right)^{-1}. \tag{8}$$

Here, $\boldsymbol{C}_\mathrm{D} \in \mathbb{C}^{D\times 2N+1}$ contains the CHs evaluated at the D measurement directions, $(\cdot)^\mathrm{H}$ denotes the Hermitian conjugate, and $\boldsymbol{W} \in \mathbb{R}^{D\times D}$ is a diagonal weight matrix to balance non-uniform measurement grids. For uniform measurement grids, no sampling weights are required, $\boldsymbol{W} = \boldsymbol{I}$. The inverse in (8) is of full rank for $D \geq 2N+1$ measurement positions that are uniformly distributed on a circle.

2.3 Omnidirectional signal model

Explicitly modeling the array translation with a limited number of CH coefficients is prone to errors at high

frequencies [37], and the perceptual impact of these errors has not been investigated thus far. Perceptually important room acoustic properties like RT and DRR do not change rapidly with position. Thus, an alternative simplified signal model may neglect all positional information and focus on describing an omnidirectional room transfer function. This transfer function can be obtained by solely considering the zeroth-order CH coefficient a_0 carrying the omnidirectional information,

$$\boxed{\boldsymbol{p}_0(\omega) = \boldsymbol{d}(\omega)\, a_0(\omega)\, s(\omega)\,,} \tag{9}$$

where $\boldsymbol{d}$ is the column of $\boldsymbol{D}$ corresponding to a_0.

In contrast to (4), where the coefficients $\boldsymbol{a}$ represent the spatial room transfer function between the source and the (fixed) coordinate origin and are valid within a region surrounding the origin, the coefficient a_0 in (9) represents the omnidirectional transfer function between the source and the time-varying array position using a time-varying expansion center. Both signal models separate the influences of the room acoustics, the array, and the source signal.

3 Blind room impulse response estimation

The estimation of spatial room transfer functions (and their time-domain counterpart, spatial RIRs) expressed as CH coefficients $\boldsymbol{a}$ from captured sound pressure $\boldsymbol{p}$ via (4) or (9) is possible with established methods like a multichannel Wiener or recursive least squares (RLS) filter if all other variables in the model are known. It is assumed in the following that the positional information needed for the calculation of $\boldsymbol{R}$ and $\boldsymbol{T}$ are known from dedicated sensors, for example from an inertial measurement unit (IMU) which is often available in head-worn AR/VR devices and smartglasses. The reference signal s is generally not available but can be estimated with the framework from [20, 21, 23] that estimates a pseudo-reference signal $\hat{s}$ from the array signals $\boldsymbol{p}$ using dereverberation and beamforming as illustrated in Fig. 1. The approach is motivated by a common description of RIRs being composed of direct sound, early reflections, and late reverberation. The dereverberation aims to cancel out the late reverberation in the array signals, while the beamformer aims to reduce the impact of early reflections while preserving the direct sound without distortion. The dereverberation and beamforming stages are identical for all methods and operate directly on the microphone-array signals. The distinction between the motion-aware and omnidirectional methods only enters in the definition of the RLS filter in the subsequent RIR estimation step.

While suitable algorithms for the individual blocks have been evaluated for stationary arrays, the present problem of moving arrays necessitates modifications. The weighted prediction error method (WPE) [42] was selected for blind dereverberation in the static scenarios in [20, 23] as it provides a high perceptual signal quality with low distortion, supports multichannel signals, does not require knowledge of the number of sound sources, and preserves channel time differences. Several modifications of the WPE for time-varying scenarios [43–45] and jointly-optimal combinations of dereverberation and beamforming [46–49] have been proposed. In our preliminary tests, a good trade-off between dereverberation performance and computational efficiency was achieved with the frame-adaptive version of the WPE described in [45] and a minimum power distortionless response (MPDR) beamformer [50, ch. 6.2.4]. In practice, the MPDR beamformer is provided with a set of measured ATFs and is supported by direction-of-arrival (DOA) estimation using the multiple signal classification (MUSIC) [51] algorithm, which determines one broadband DOA per frame. Although a combination of the WPE method with a weighted MPDR beamformer was shown to be jointly optimal [46], we chose to use the conventional MPDR as it was more robust in our preliminary tests, i.e., it provided good performance without the need for detailed parameter tuning for different sound fields.

For notational convenience, the dependencies on frequency and position are omitted in the following. With the estimated single-channel pseudo-reference signal $\hat{s}$, we aim to minimize the squared error between the predicted sound pressure from the signal model (4) and the measured sound pressure $\hat{\boldsymbol{p}}$ over B signal blocks, similar to [37]:

$$\min_{\boldsymbol{a}} \frac{1}{B} \sum_{b=0}^{B-1} \|\boldsymbol{D}\boldsymbol{R}_b\boldsymbol{T}_b\boldsymbol{a}\hat{s}_b - \hat{\boldsymbol{p}}_b\|^2 + \delta\|\boldsymbol{a}\|^2\,. \tag{10}$$

Here, δ is a small regularization parameter that penalizes large coefficient norms and stabilizes the estimation.

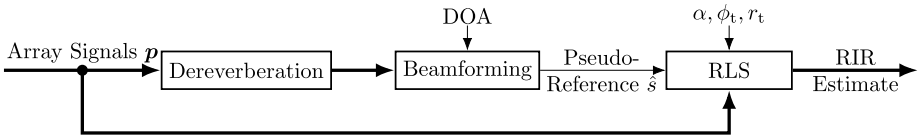

Fig. 1 The SRIR is estimated using an RLS filter with the array signals $\boldsymbol{p}$, the pseudo-reference signal $\hat{s}$, and positional tracking data α, ϕ_t, r_t

The least-squares-optimal solution for the motion-aware signal model from (4) is given by

$$\hat{\boldsymbol{a}} = \left(\frac{1}{B} \sum_{b=0}^{B-1} \hat{s}_b^* \hat{s}_b \boldsymbol{T}_b^{\mathrm{H}} \boldsymbol{R}_b^{\mathrm{H}} \boldsymbol{D}^{\mathrm{H}} \boldsymbol{D} \boldsymbol{R}_b \boldsymbol{T}_b + \delta \boldsymbol{I} \right)^{-1} \frac{1}{B} \sum_{b=0}^{B-1} \hat{s}_b^* \boldsymbol{T}_b^{\mathrm{H}} \boldsymbol{R}_b^{\mathrm{H}} \boldsymbol{D}^{\mathrm{H}} \hat{\boldsymbol{p}}_b \,, \tag{11}$$

and for the omnidirectional signal model from (9) by

$$\hat{a}_{\mathrm{o}} = \left(\frac{1}{B} \sum_{b=0}^{B-1} \hat{s}_b^* \hat{s}_b \, \boldsymbol{d}^{\mathrm{H}} \boldsymbol{d} + \delta \right)^{-1} \frac{1}{B} \sum_{b=0}^{B-1} \hat{s}_b^* \, \boldsymbol{d}^{\mathrm{H}} \hat{\boldsymbol{p}}_b \,. \tag{12}$$

Both sums in (11) and (12) can be updated recursively, resulting in an RLS filter. Depending on whether the room transfer functions are assumed time-invariant or time-varying, the updates can be performed using cumulative averaging or (exponential) moving averaging [37]. In the following, we assume movement in a sound field of stationary sources that can be described by time-invariant transfer functions at each receiver position. Since the array movement is explicitly modeled in the motion-aware signal model, cumulative averaging is applied in this case. In contrast, the omnidirectional model employs exponential averaging as it treats the movement as a time variation of the sound field. Finally, (multichannel CH-domain) RIRs are obtained from the estimated transfer functions via the inverse discrete Fourier transform.

4 Synthesis of binaural room impulse responses

While virtual sound sources could be rendered by directly convolving the estimated RIRs with source signals and binauralizing the result using head-related transfer functions (HRTFs), previous work revealed that such estimates may exhibit narrow-band ringing artifacts due to limitations in the pseudo-reference signal estimation, which can be effectively suppressed using noise-based resynthesis of the late reverberation tail in octave bands [23]. Additionally, preliminary analysis indicated that estimates obtained under array motion exhibit inaccuracies in the spectral magnitude and direction of the reproduced direct sound, as well as in the directional energy distribution, with the magnitude errors being most pronounced at high frequencies. To address these limitations, we propose a resynthesis procedure that models the direct sound as an ideal order-N plane wave with flat magnitude response and creates an isotropic reverberation tail from the omnidirectional RIR estimate through noise-based resynthesis in octave bands. In this process, all directional information and early reflections contained in the estimates are discarded. Accurate modeling of the early reflections may, however, not be critical in practice, since the virtual source is not synthesized at the same position as the source used for estimation. Note that omnidirectional information is available not only from the omnidirectional estimate in (12), but also from the zeroth-order coefficient of the motion-aware estimate in (11). The latter may be more accurate because the motion-aware model explicitly accounts for movement through its higher-order coefficients.

Figure 2 illustrates the proposed synthesis method to produce a binaural room impulse response (BRIR) from room acoustic parameter estimates. The direct sound component is modeled as an ideal plane wave, with its magnitude scaled so that the DRR of the resynthesized response matches the estimated broadband DRR. The reverberation tail is synthesized from uncorrelated white noise for each CH channel. As in [23], the decay envelope of each channel and octave band is shaped to match the RT estimate, while the RMS energy per octave band is scaled to match the RMS energy of the estimate. In contrast to the previous work, the resynthesis is directly applied in the CH domain and the parameter estimates are obtained from the omnidirectional RIR coefficient only. Accordingly, the synthesized reverberation tails of all channels have the same energy decay and RMS energy per octave band after the processing. The use of uncorrelated noise with equal variance across channels ensures that the resulting late reverberation is both diffuse and isotropic [52]. As the synthesis combines an ideal direct sound impulse with an isotropic reverberation tail, it facilitates the generation of RIRs of arbitrary CH order.

Once the CH-domain RIRs are synthesized, they can be efficiently rotated using a CH-domain rotation matrix $\boldsymbol{R}$ to accommodate head tracking during rendering. The final conversion to binaural ear signals is performed using the magnitude least squares method [53].

5 Experiment setup

5.1 Dataset

Experimental data was collected using the custom head-worn microphone array prototype integrated into a pair of sunglasses from Fig. 3. The array consists of seven omnidirectional microphones: three on each side and one centered above the nose bridge. Six optical tracking markers were attached to enable precise position and orientation tracking.

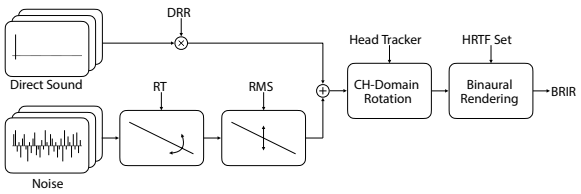

Fig. 2 From the estimated RT, DRR, and RMS energy, a RIR is synthesized by concatenating a direct sound peak and shaped noise for each octave band and channel

Fig. 3 Microphone array prototype with seven microphones and six tracking markers on sunglasses

Prior to the main experiments, anechoic ATFs were measured for the array mounted on a GRAS KEMAR 45BB head and torso simulator. Measurements were taken at 36 equiangular positions with 10° spacing in the horizontal plane to facilitate the computation of the array matrix in (8).

The main experimental data was collected in four acoustically distinct spaces:

- A small office (4.0 m × 3.2 m × 2.7 m, reference distance 2.1 m, DRR −4.3dB)
- A kitchen (6.4 m × 4.0 m × 3.1 m, reference distance 2.8 m, DRR −3.9dB)
- A storage room (12.0 m × 3.0 m × 3.3 m, reference distance 2.4 m, DRR −6.3dB)
- A hallway (40.0 m × 1.4 m × 3.3 m, reference distance 2.5 m, DRR −5.2dB)

The octave-band reverberation times of these spaces in Fig. 4 were obtained from reference measurements and computed with the ITA toolbox [54].

In each room, recordings were made while a participant wearing the array performed two types of movements:

1. Head-only movement: The participant maintained a fixed standing position at the reference distance while performing natural head movements, resulting in array translations up to 30 cm from the origin.

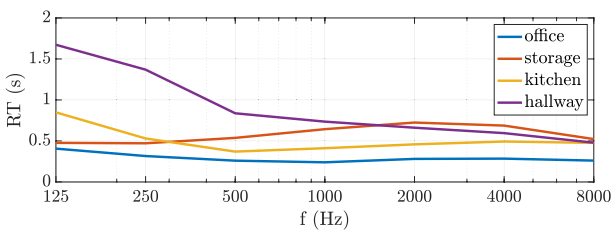

Fig. 4 Octave-band RTs of the four rooms in the dataset

2. Combined movement: The participant moved and rotated their head freely within a 70-cm radius around the reference position.

For both movement types, the subject was instructed to maintain natural head motion primarily in yaw, avoiding deliberate pitch and roll rotations or vertical movements. During the movements, 60 s of either male or female speech was played through a Genelec 8030 A loudspeaker and the recorded signals from the microphone array were then used for the estimation. Movement trajectories were captured using an OptiTrack system comprising four PrimeX13W cameras. Example trajectories for both movement types are shown in Figs. 5 and 6.

Additionally, in each room, a set of RIRs serving as reference was measured using an exponential sine sweep

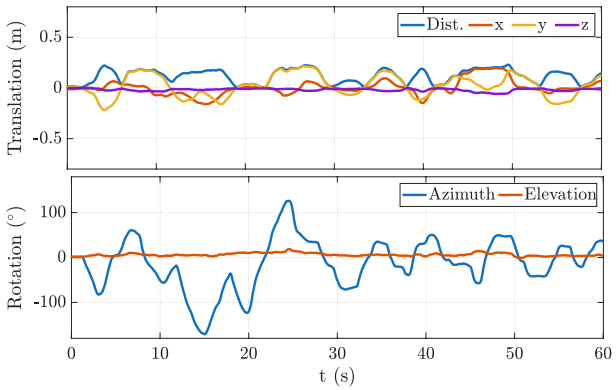

Fig. 5 Example trajectories of array translation distance and Cartesian coordinates (top), and rotation in azimuth and elevation (bottom) of head-only array movement in the small office, where the wearer freely moved their head but kept their feet static

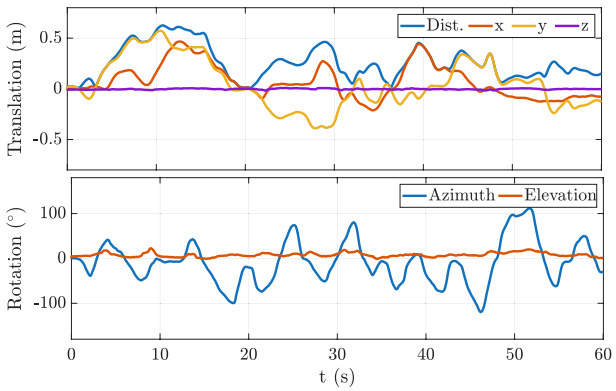

Fig. 6 Example trajectories of array translation distance and Cartesian coordinates (top), and rotation in azimuth and elevation (bottom) of the combined array movement in the kitchen, where the wearer freely moved within a circle of 70 cm radius

with a 22-microphone equatorial microphone array supporting $N = 10$, as in [37].

5.2 Algorithm setup

All processing was performed at a sampling frequency of 48 kHz. The processing chain consisted of three main components shown in Fig. 1: dereverberation, beamforming, and RIR estimation, each with specific parameter configurations.

The WPE dereverberation employed short-time Fourier transform (STFT) processing with a block length of 2048 samples and a hop size of 256 samples, using a square-root Hann window. The algorithm operated with a prediction delay of 20 ms, and second-order signal statistics were averaged over 1.5 s. To optimize performance across the frequency range, the prediction order was frequency-dependent: 32 taps below 500 Hz, 28 taps between 500 Hz and 1 kHz, 24 taps between 1 kHz and 8 kHz, and 12 taps above 8 kHz. The regularization parameter was set to 10^{-4}.

The MPDR beamformer used STFT processing with a block length of 2048 samples, a hop size of 1024 samples, and a square-root Hann window. DOA estimation was performed using the MUSIC algorithm [51], analyzing frequency components between 100 Hz and 1 kHz to determine a single DOA per frame. The DOA estimation employed second-order statistics averaged over 400 ms, and the beamformer's regularization parameter was set to 10^{-2}.

The RLS filter used STFT processing with a block length of 0.5 s (office), 1.0 s (kitchen), 1.5 s (storage), and 2.0 s (hallway). In all cases, the hop size was 25% of the block length, and a square-root Hann window was applied. The motion-aware signal model used a maximum CH order of $N = 2$. The omnidirectional model employed exponential averaging with a forgetting factor of 0.99. For both signal models, the filter's regularization parameter was set to 10^{-2} times the average frequency-dependent microphone self-noise power. Adaptation was terminated either after 60 s or when the RMS energy of the exponentially-smoothed RIR update (smoothing factor 0.99) fell below -20 dB. The array signals were delayed by 5 ms relative to the pseudo-reference during the estimation to ensure a causal result.

A baseline was established using a single-channel RIR estimate from the microphone on the nose bridge, utilizing only the corresponding channel's signal and a pseudo-reference signal estimated by applying WPE dereverberation to the same channel. Otherwise, this baseline used identical parameter settings as the omnidirectional model.

The BRIR synthesis used a CH order of $N = 10$, the maximum order supported by the reference measurements with the equatorial array, and a frequency range from 100 to 16 kHz, where all speech signals had sufficient energy. The RT was estimated as T_{20} in octave bands using the ITA toolbox [54] and the DRR was estimated as the energy ratio of the direct sound (extracted using a window of 0.5 ms before to 1 ms after the direct sound peak) and the rest of the RIR until the noise floor, which was obtained from Lundeby's algorithm [55] using the ITA toolbox.

Binaural rendering was performed using the magnitude least squares method [53] with CH basis functions and 60 horizontal HRTFs from the Neumann KU100 dummy head [56].

6 Numerical evaluation

Since the BRIR synthesis procedure from Sec. 4 relies on accurate estimates of octave-band RT, broadband DRR, and octave-band RMS energy, we evaluate the estimation accuracy of these acoustic parameters. All metrics are computed from the zeroth-order CH coefficient only. Note that the motion-aware signal model from (4) still estimates higher-order coefficients, as these may contribute to a more accurate estimation of the zeroth-order component through the translation model. We compare blind RIR estimation using the motion-aware signal model from (4) and the omnidirectional one from (9) to informed estimation (given the ideal reference signal) with the motion-aware model to identify the impact of inaccuracies in the estimated reference signal, and to the blind baseline approach. The compared methods are summarized in Table 1. Results in this section are presented for each algorithm and movement type, after averaging over results from the different rooms and repetitions with male/female speech. The error metrics are calculated by comparing the estimates with an

Table 1 Overview of the compared methods

Method	Signal model	Estimation	Reference signal
Informed Motion-Aware (MA)	Motion-Aware (4)	Informed	Oracle
Blind Motion-Aware (MA)	Motion-Aware (4)	Blind	Pseudo-Reference
Blind Omnidirectional (Omni)	Omnidirectional (9)	Blind	Pseudo-Reference
Baseline	Single-Channel RIR	Blind	Derev. Single-Channel

omnidirectional RIR derived from the reference RIRs measured with the equatorial microphone array.

Figure 7 shows the average RT estimation error relative to ground truth measurements for both movement types. The algorithms exhibit similar error patterns across movement conditions, with best performance in the mid-frequency range (250 to 2000 Hz) and increased errors toward lower and higher frequencies. Overall, the informed motion-aware (MA) method achieves the lowest average errors of approximately 10% in the mid-frequency range for head-only movements and 10% to 15% for combined movement. Both blind methods (motion-aware and omnidirectional) perform similarly for head-only movements and match the performance of the informed method at low frequencies, but show notably higher errors at mid and high frequencies. For the combined movements, the blind methods overall perform worse than the informed one with the blind motion-aware method having a small advantage over the omnidirectional one at low frequencies, and the blind omnidirectional method outperforming the motion-aware one at high frequencies. The baseline single-microphone approach performs worst, with errors around 15% to 20% at mid frequencies increasing to around 30% (head-only movement) and 40% (combined movement) at frequency extremes. Interestingly, the baseline performs similar to the blind methods at mid frequencies, and even similar to the informed method in the 1 kHz and 2 kHz octave bands for the combined movement.

The average absolute DRR estimation errors in Table 2 show a clear performance hierarchy: The informed (motion-aware) method achieves lowest errors (1.0 dB/1.6 dB for head-only/combined movements), followed by the blind motion-aware method (1.6 dB/3.3 dB),

Table 2 Mean and standard deviation of absolute DRR errors (in dB) of the four estimation methods for head-only and combined movements

	Informed (MA)	Blind (MA)	Blind (Omni)	Baseline
Head-only	1.0 ± 0.9	1.6 ± 1.3	4.9 ± 2.1	24.9 ± 2.3
Combined	1.6 ± 1.3	3.3 ± 0.8	6.0 ± 2.5	24.8 ± 1.4

the blind omnidirectional approach (4.9 dB/6.0 dB), and the baseline (24.9 dB/24.8 dB). The increased errors for larger movements are consistent with the growing model mismatch, except for the baseline where DRR estimation failed entirely.

Figure 8 presents the mean absolute octave-band RMS errors, defined as the average absolute difference between the RMS energy of an estimated RIR and the corresponding ground truth. The informed motion-aware method yields the lowest average errors overall, with the largest mean deviations of approximately 4 dB in the 125 Hz band, moderate mean errors in the 1 kHz to 2 kHz range, and the smallest mean errors of about 1 dB in the remaining bands. Both blind methods perform comparably at low and high frequencies, but their behavior diverges in the mid-frequency range, where the motion-aware model exhibits substantially larger average errors, up to 5 dB for head-only movement and 7 dB for combined movement. The baseline method, in contrast, shows markedly higher errors at low frequencies (up to 7 dB), while its mid-frequency performance ranks among the best, comparable to that of the informed and blind omnidirectional approaches. For the combined-movement scenario, both the baseline and blind omnidirectional methods even slightly outperform the informed method in the 1 kHz to 2kHz bands.

Figure 9 displays the omnidirectional component of the time-domain room impulse responses for the reference

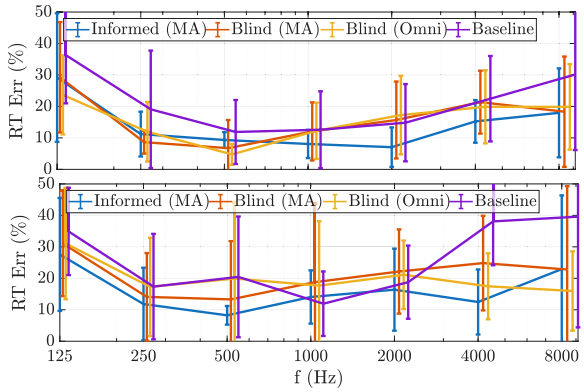

Fig. 7 Mean and standard deviation of relative octave-band RT errors of the informed motion-aware (MA), the blind motion-aware (MA), the blind omnidirectional (Omni), and the baseline method for head-only movement (top) and the combined movement (bottom)

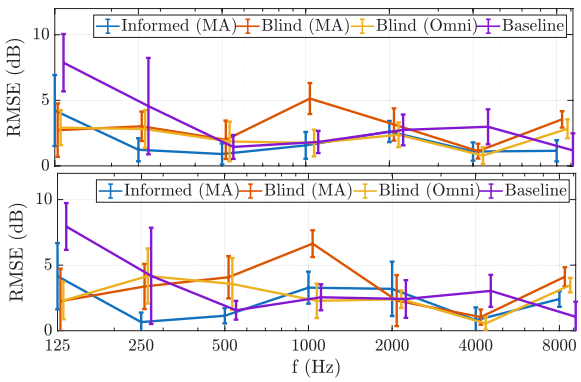

Fig. 8 Mean and standard deviation of absolute octave-band RMS errors of the informed motion-aware (MA), the blind motion-aware (MA), the blind omnidirectional (Omni), and the baseline method for head-only movement (top) and the combined movement (bottom)

measurement, the four estimation methods, and their resynthesized counterparts, for the head-only movement scenario in the kitchen. Because the pseudo-reference signal does not preserve the absolute propagation delay, the estimated RIRs do not recover the direct-sound arrival time. This is not important for the intended virtual source rendering, where reproducing the original delay is unnecessary. Therefore, Fig. 9 shows delay-aligned RIR estimates and resynthesized responses without delay for clarity.

The baseline method's failure to estimate a valid DRR is evident from the exaggerated direct-sound peak relative to the reference. Among the approaches, only the ones using the motion-aware signal model are capable of recovering distinct early reflection peaks.

7 Perceptual evaluation

7.1 Experiment setup

To evaluate the perceptual quality of virtual sound sources rendered using resynthesized BRIRs from the estimated room acoustic parameters, we conducted a listening experiment focusing on assessing their similarity to binaural renderings based on sets of measured RIRs obtained with the equatorial microphone array. To limit complexity, only estimates from the combined movement (cf. Sect. 5.1) were evaluated. Given inherent estimation inaccuracies, we did not aim to achieve perceptual authenticity [57] or transfer-plausibility [58], but rather assessed similarity to renderings of measured reference RIRs through two research questions:

(i) Which estimation method produces (resynthesized) renderings most similar to those based on measured RIRs?

(ii) Are renderings based on estimated, resynthesized BRIRs perceived as more similar to the reference than renderings based on measurements from other rooms?

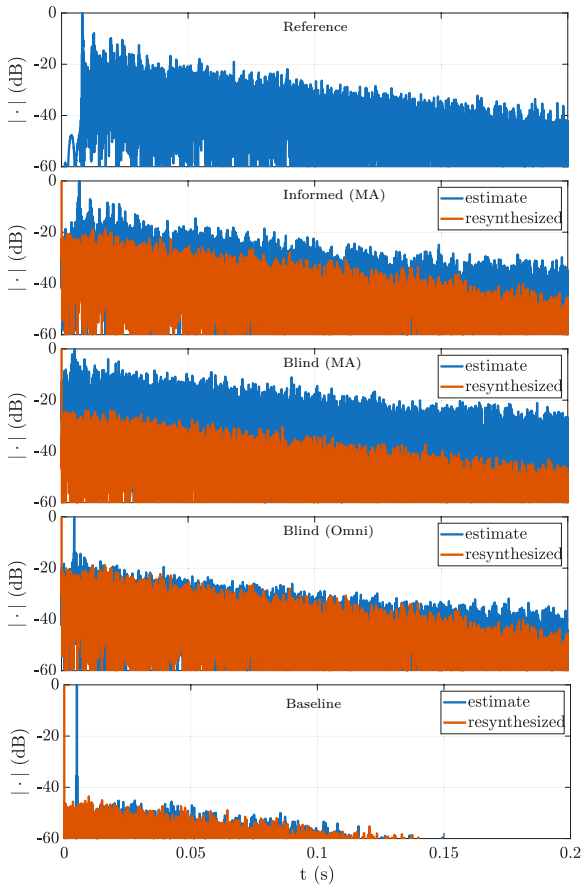

Fig. 9 Omnidirectional time-domain responses (in dB) of reference, estimates, and resynthesized estimates in the kitchen for head-only movement

To answer these questions, we designed a two-part experiment. The first part investigates within-room similarity, and the second part across-room comparison. For both parts of the experiment, BRIRs were synthesized using the procedure from Sect. 4, i.e., from RT, DRR, and octave-band RMS estimates derived from the omnidirectional RIR component only. For the motion-aware model, this corresponds to its zeroth-order CH coefficient. The test stimuli were generated from resynthesized BRIRs based on the informed motion-aware, blind motion-aware, and blind omnidirectional estimates. By contrast, the reference, hidden reference, and measurement-based renderings from other rooms (Part 2) were based on binaural renderings of the measured equatorial-array RIRs and were not resynthesized.

The order of the trials and of the stimuli within each trial was randomized for each participant. Participants were allowed to listen to the stimuli as often as they wanted. All stimuli in both parts of the experiment were loudness normalized using a MATLAB implementation of the algorithm in [59] and presented over AKG K702 headphones at a comfortable listening level. We used static binaural renderings to keep the listening task focused and avoid additional complexity arising from head-dependent comparisons, but added controlled perceptual variation between reference and test stimuli, by presenting reference stimuli at 30° and test stimuli (including hidden references) at −30° azimuth. Audio examples for both parts of the experiment are available online.[1]

A total of 21 participants (16 female, 5 male, average age 31 years) with self-reported normal hearing completed the experiments. Most were non-expert listeners with

[1] https://thomasdeppisch.github.io/room-acoustic-matching/.

knowledge in acoustics. The first part of the experiment took them on average 15 min, and the second part 16 min.

7.2 Part 1: within-room similarity

Setup For question (i), we employed a Multiple Stimulus with Hidden Reference and Anchor (MUSHRA) like design [60] where participants had to rate the similarity of stimuli to a reference on a scale between *identical* and *very different*. The participants were presented with one training trial for familiarization with the procedure (without any feedback on their selection), followed by eight experiment trials. Each trial presented participants with five stimuli for comparison against a reference: a hidden reference, a low-quality anchor, and three test stimuli. The test stimuli were generated from resynthesized BRIRs of the informed motion-aware method, the blind motion-aware method, and the blind omnidirectional method. The baseline method was not compared in the experiment because of its inability to estimate meaningful DRRs. All test stimuli used BRIRs synthesized via the procedure from Sect. 4. The anchor condition was an anechoic binaural rendering low-passed at 2 kHz using a second-order Butterworth filter. The reference condition was based on magnitude-least-squares binaural renderings of the measured RIRs from the equatorial microphone array, analytical radial filters, and a CH order of $N = 10$.

Each trial used a different speech sample from the Dataset of Audiovisual Speech for AR Telepresence Studies [61] as source material which was used for all stimuli within the trial. To reduce potential bias, RIR estimates obtained from female speech were tested using male speech and vice versa. The eight trials comprised one for each of the four rooms with both male and female speech.

Results Results of the MUSHRA experiment from 5 participants were removed from the analysis as their average rating of the hidden reference was below 80 MUSHRA points. Figure 10 shows violin plots of the MUSHRA ratings, with individual scores as dots, medians as white dots, and interquartile ranges as boxes. Hidden references consistently received highest median ratings and anchors lowest across all rooms. The blind motion-aware method produced lower median scores than both the blind omnidirectional method and the informed motion-aware method. In three out of four rooms, the blind omnidirectional method received higher median ratings than the informed motion-aware method. Only in the kitchen, the informed motion-aware method performed better.

The statistical analysis used participant-wise averages across male/female speech conditions. Friedman tests revealed significant effects for all rooms, with p-values below 0.001 (office: $\chi^2 = 60$, kitchen: $\chi^2 = 57$, storage: $\chi^2 = 56$, hallway: $\chi^2 = 57$). Post-hoc Wilcoxon signed rank tests with Bonferroni-Holm correction showed significant differences (p-value < 0.05) between all algorithm pairs but the pair Blind (MA)/Informed (MA) for the hallway condition (p-value $= 0.072$), with most comparisons achieving p-value < 0.01. Two comparisons showed significance for p-value < 0.05 but not for p-value < 0.01: Informed (MA)/Blind (Omni) (p-value $= 0.049$) and Informed (MA)/Blind (MA) (p-value $= 0.017$) both in the storage.

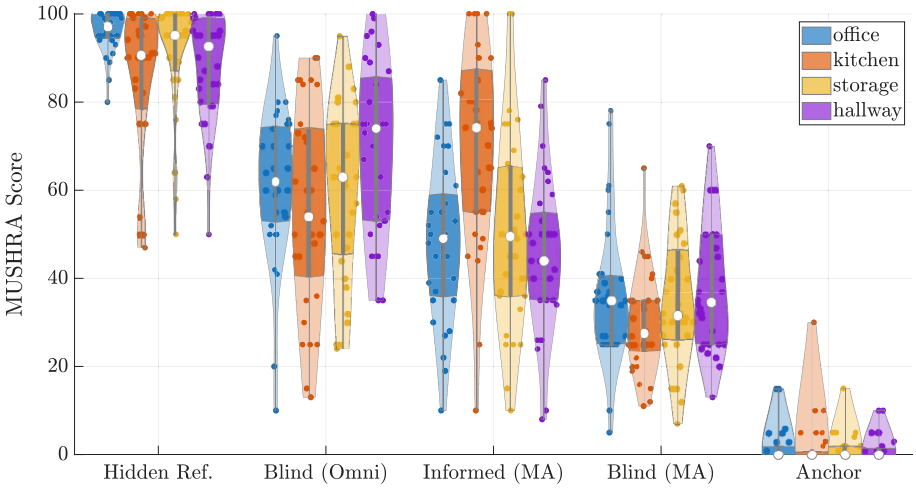

Fig. 10 Violin plots of the MUSHRA results showing individual ratings (colored dots), median ratings (white dots), and quartiles (indicated by darker-shaded boxes around the median)

In summary, the statistical analysis confirmed that the observed differences between methods were consistent across listeners. Friedman tests revealed significant effects for all rooms, indicating that the five conditions were rated differently overall. Post-hoc Wilcoxon signed-rank tests showed that the blind omnidirectional method was rated significantly higher than the blind motion-aware method and, in three out of four rooms, also higher than the informed motion-aware method.

7.3 Part 2: across-room comparison

Setup Question (ii) was addressed through a two-alternative forced choice (2AFC) paradigm similar to [23], asking participants to listen to a reference and two stimuli and select which stimulus *was recorded in the same room as the reference*. The reference binaural rendering was always based on a set of measured RIRs. The comparison stimuli contrasted binaural renderings based on resynthesized BRIRs from the blind omnidirectional estimates with binaural renderings based on measured RIRs from either the same room or a different room. To limit the complexity of the experiment, only estimates from the blind omnidirectional method from (12) applied to signals from the combined movement were used in this part of the experiment. The participants were presented with two training trials for familiarization with the procedure (without any feedback on their selection), followed by 48 experimental trials.

Unlike the MUSHRA tests, each 2AFC stimulus used different speech content that was randomly drawn for each participant and trial from a database of 8 male and 8 female speakers (a subset of [61]), each contributing 10 Harvard sentences. This design, similar to the transfer-plausibility paradigm [58], ensured that participants focused on comparing room acoustic rather than source properties.

Results The results from the second part of the experiment in Fig. 11 show the percentage of selected renderings based on resynthesized BRIRs from blind omnidirectional estimates or on measurements from the same room as the reference, over renderings based on a measurement in a different room. Significance levels for $\alpha = \{0.5, 0.05, 0.01\}$ were computed from the binomial distribution and confidence intervals for the selection percentages were calculated using the Wilson score interval [62] after averaging over repetitions for each participant.

For 7 out of 12 comparison pairs, renderings based on both the estimate and the measurement from the reference room were selected over a measurement from a different room in more than 80% of comparisons. These results are significantly different from guessing with p-value < 0.01. In four cases (office/kitchen, kitchen/office, storage/kitchen, storage/hallway), the percentage of selected measurements from the same room are lower but still significantly different from guessing with p-value < 0.01. The selection percentages for the estimates are significantly different from guessing with p-value < 0.05 in three of these cases but not in the

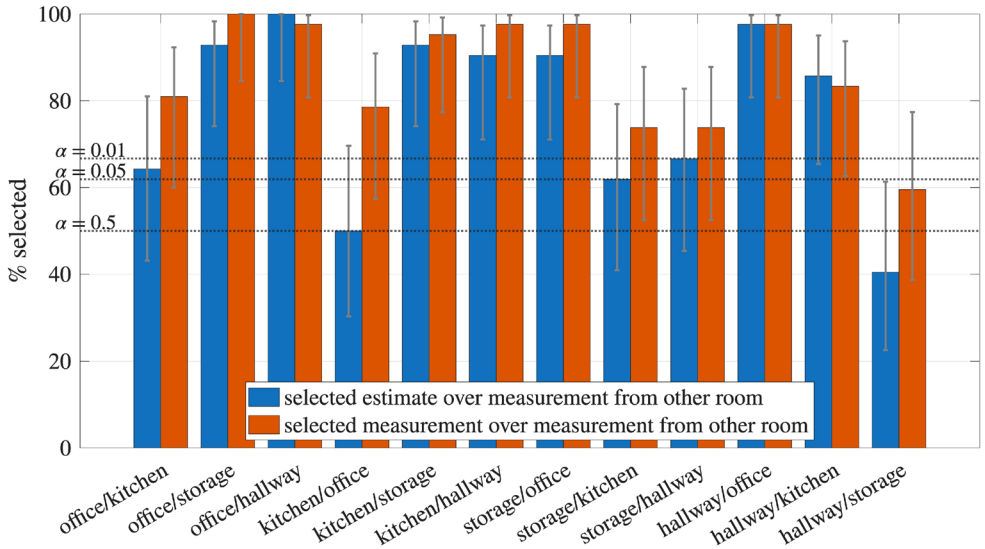

Fig. 11 Results from the 2AFC paradigm. Blue bars show the percentage of selected renderings based on resynthesized BRIRs from blind omnidirectional estimates in the reference room over measured RIRs from other rooms. Red bars show the percentage of selected renderings based on measured RIRs from the reference room over measured RIRs from other rooms

fourth case (kitchen/office). In one case (hallway/storage), neither the selection percentage of the measurement from the same room nor the estimate was found to be significantly different from a random choice.

In summary, for most room pairs, both the measured same-room renderings and the renderings based on blind omnidirectional estimates were selected significantly above chance. This indicates that the blind omnidirectional method often preserved enough room-specific information for listeners to associate the rendering with the correct room, although the measured same-room renderings were generally selected more often.

8 Discussion

The numerical results from Sect. 6 show that the motion-aware signal model achieves clear advantages in RT, DRR, and RMS energy estimation when provided with the true reference signal. However, these benefits diminish in the blind setting, where RT errors become comparable to those of the blind omnidirectional method, and RMS errors are notably higher. In other words, errors in the reference signal estimate introduced by shortcomings in dereverberation, DOA estimation, and beamforming, prevent the effective use of the motion-aware signal model. Only in DRR estimation, the motion-aware blind method outperforms the omnidirectional variant, likely due to missing energy from early reflections, as observed in Fig. 9. Across all metrics, the multichannel methods significantly surpass the single-channel baseline, whose poor results are mainly attributed to the reduced effectiveness of the WPE algorithm in the single-channel case and the missing beamforming stage.

The pronounced performance gap between the informed and blind motion-aware methods highlights the importance of accurate reference signal estimation. The dominant source of error is likely the dereverberation stage, as previous findings from static scenarios indicate that DOA estimation inaccuracies have only a minor influence on RT and DRR estimates [23]. This suggests that the overall framework could be improved by incorporating a more advanced dereverberation algorithm, for example, based on recent deep learning techniques [63, 64]. Such improvements would likely also benefit the blind omnidirectional approach and the baseline.

While the informed motion-aware method showed substantial advantages in acoustic parameter estimation, these benefits did not translate directly into perceptual quality. In the MUSHRA experiment, renderings from the informed motion-aware method were rated highest for only one of the four rooms, whereas the blind omnidirectional method performed best in the remaining cases. These findings suggest that the perceptual disadvantage of the motion-aware method may be related to errors introduced by modeling the sound field using a low-order CH representation. In principle, a higher-order CH or SH representation (requiring more microphones) could reduce such truncation errors. Whether this would improve perceptual performance in practice, however, remains a question for future work, as model mismatch may limit performance in real-world scenarios, for example due to natural movement in a 3D sound field and discrepancies between measured ATFs and the effective ATFs during human-worn operation. In practice, the estimates may also be improved by incorporating uncertainty weighting into the (recursive) averaging. For both signal models, estimation may benefit from assigning greater weight to observations close to the expansion center (the coordinate origin), or from disregarding observations that are too far away.

The 2AFC results suggest the practical viability of the blind omnidirectional approach. In the majority of room comparisons, participants were generally able to distinguish between rooms based on the rendered stimuli, with similar selection rates for estimate-based and measurement-based renderings. This suggests that the omnidirectional method in many cases captures room-specific acoustic characteristics well enough to provide a benefit for augmented reality applications and that explicit spatial modeling with the motion-aware signal model may not be required for practical applications. The cases where discrimination based on the estimates was more challenging (particularly comparisons between hallway and storage, and kitchen and office) likely arise from a combination of estimation inaccuracies and acoustic similarities between those spaces, as even renderings based on measured RIRs showed lower percentages of correctly selected stimuli.

It is noteworthy that both the numerical and the perceptual performance remained robust despite substantial array movement (up to 70 cm from the reference position) and diverse acoustic conditions across the four rooms. The combination of head rotation and translation represents a more challenging scenario than those in prior studies that considered only stationary [23] or linearly moving arrays [37]. Compared to [23], which used a microphone array integrated into smartglasses in stationary scenarios and reported average RT and DRR errors of 11% and 1.6 dB, the informed motion-aware method achieves comparable accuracy at mid frequencies, despite the additional motion.

Compared to [37], the present study considers a substantially more realistic setting, with a head-worn array undergoing both rotation and translation rather than an equatorial array moving along a straight line. Although a direct quantitative comparison is difficult because

different evaluation metrics were used, the controlled scenario in [37] appears to yield more accurate estimates overall. This suggests that the present results should be interpreted as representative for practical head-worn arrays rather than as being independent of the specific array design. Since the proposed framework is based on measured ATFs, moderate changes in shape, size, and microphone distribution across realistic wearable array designs are, however, not expected to fundamentally change the conclusions.

The present study employs CHs because they are well suited to practical head-worn arrays, whose microphones are typically distributed approximately along a horizontal contour around the head. With an array design that supports SH processing, and with a correspondingly larger number of microphones, similar or slightly improved performance could be expected because an SH-based formulation would allow the motion model to account not only for horizontal translation and yaw head rotation, but also for vertical movement and pitch/roll rotations.

The method's performance is also comparable to that of recent deep learning–based methods under static conditions [26], indicating that a more accurate reference signal estimate may close the performance gap of the proposed method in dynamic scenarios to current state-of-the-art methods in static conditions.

Nevertheless, several limitations of the present study should be acknowledged. The experiments were conducted with a single, stationary sound source in quiet environments, and only four rooms were tested due to the complexity of the setup. The evaluation focused exclusively on speech, and estimation block lengths were predefined rather than adapted dynamically. Future work should therefore focus on improving dereverberation and pseudo-reference signal estimation, introducing automatic block-length selection, and extending the evaluation to more complex scenarios.

9 Conclusion

This paper presented a method for matching the room acoustics of a virtual source to a real environment using a head-worn microphone array under realistic movement. Two signal models were evaluated: a motion-aware model that explicitly accounts for array movement and an omnidirectional model that neglects the motion and directional room acoustic properties. While the motion-aware model achieved higher accuracy in parameter estimation, the omnidirectional model outperformed it in perceptual experiments and showed robust performance in distinguishing between different rooms in a 2AFC task. These findings indicate that explicitly modeling array movement may offer limited practical benefit for head-worn arrays with few microphones, and that simpler models can provide perceptually convincing results for realistic applications. Through the analysis of the impact of different model assumptions on both parameter estimation and perceptual quality, this study provides insights that will aid the future development of robust room-acoustic matching algorithms.

Authors' contributions

T.D. contributed to conceptualization, methodology, measurements, experiments, data analysis, and writing. All authors contributed to conceptualization, manuscript revision and editing, and approved the final version of the manuscript.

Funding

Open access funding provided by Chalmers University of Technology. We thank Reality Labs Research at Meta for funding this research.

Data availability

Example code applying and comparing the methods is available at https://github.com/thomasdeppisch/room-acoustic-matching, measurement data at https://zenodo.org/records/19854597, and audio examples are served at https://thomasdeppisch.github.io/room-acoustic-matching/.

Declarations

Ethics approval and consent to participate

The study was conducted in accordance with the Declaration of Helsinki. Ethics approval was not required (Swedish Ethical Review Authority, Dnr 2025-01059-01).

Competing interests

The authors declare that they have no competing interests.

Received: 26 January 2026 Accepted: 6 June 2026
Published online: 12 June 2026

References

1. A. Neidhardt, C. Schneiderwind, F. Klein, Perceptual matching of room acoustics for auditory augmented reality in small rooms - literature review and theoretical framework. Trends Hear. **22**, 1–22 (2022). https://doi.org/10.1177/23312165221092919
2. J.C. Gil-Carvajal, J. Cubick, S. Santurette, T. Dau, Spatial hearing with incongruent visual or auditory room cues. Sci. Rep. **6**(37342), 1–10 (2016). https://doi.org/10.1038/srep37342
3. S. Werner, F. Klein, T. Mayenfels, K. Brandenburg, in *8th International Conference on Quality of Multimedia Experience*. A summary on acoustic room divergence and its effect on externalization of auditory events (IEEE, 2016), pp. 1–6. https://doi.org/10.1109/QoMEX.2016.7498973
4. A. Neidhardt, S. Kamandi, in *152nd Conv. Audio Eng. Soc.* Plausibility of an approaching motion towards a virtual sound source II: In a reverberant seminar room (Audio Engineering Society, 2022), pp. 1–13
5. S. Werner, F. Klein, T. Sporer, in *Int. Conf. on Audio for Virtual and Augmented Reality (AVAR)*. Adjustment of the Direct-to-Reverberant-Energy-Ratio to Reach Externalization within a Binaural Synthesis System (Audio Engineering Society, 2016), pp. 1–7
6. S.V. Amengual Garí, H.G. Hassager, F. Klein, J.M. Arend, P.W. Robinson, in *Int. Conf. on Immersive and 3D Audio*. Towards Determining Thresholds for Room Divergence: A Pilot Study on Perceived Externalization (IEEE, 2021), pp. 1–7

7. T.D.M. Prego, A.A. de Lima, S.L. Netto, B. Lee, A. Said, R.W. Schafer, T. Kalker, A blind algorithm for reverberation-time estimation using subband decomposition of speech signals. J. Acoust. Soc. Am. **131**(4), 2811–2816 (2012). https://doi.org/10.1121/1.3688503
8. Y. Hioka, K. Niwa, in *Proc. ACE Challenge Workshop*. PSD estimation in Beamspace for Estimating Direct-to-Reverberant Ratio from A Reverberant Speech Signal (IEEE, 2015), pp. 1–5
9. J. Eaton, N.D. Gaubitch, A.H. Moore, P.A. Naylor, Estimation of room acoustic parameters: the ACE challenge. IEEE/ACM Trans. Audio Speech Lang. Process. **24**(10), 1681–1693 (2016). https://doi.org/10.1109/TASLP.2016.2577502
10. H. Gamper, I.J. Tashev, in *IEEE Int. Workshop on Acoustic Signal Enhancement*. Blind reverberation time estimation using a convolutional neural network (IEEE, 2018), pp. 136–140
11. P. Götz, C. Tuna, A. Walther, E.A.P. Habets, Online reverberation time and clarity estimation in dynamic acoustic conditions. J. Acoust. Soc. Am. **153**(6), 3532–3542 (2023)
12. Y. Huang, J. Benesty, A class of frequency-domain adaptive approaches to blind multichannel identification. IEEE Trans. Signal Process. **51**(1), 11–24 (2003). https://doi.org/10.1109/TSP.2002.806559
13. M.A. Haque, M.K. Hasan, Noise robust multichannel frequency-domain LMS algorithms for blind channel identification. IEEE Signal Process. Lett. **308**, 305–308 (2008)
14. B. Jo, P. Calamia, in *28th European Signal Processing Conference*. Robust blind multichannel identification based on a phase constraint and different lp-norm constraints (EURASIP, 2021), pp. 1966–1970. https://doi.org/10.23919/Eusipco47968.2020.9287710
15. C.J. Steinmetz, V.K. Ithapu, P. Calamia, in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. Filtered Noise Shaping for Time Domain Room Impulse Response Estimation from Reverberant Speech (IEEE, 2021), pp. 221–225
16. S. Kitić, J. Daniel, Blind identification of ambisonic reduced room impulse response. IEEE/ACM Trans. Audio Speech Lang. Process. **458**, 443–458 (2023). https://doi.org/10.1109/taslp.2023.3332546
17. Z. Liao, F. Xiong, J. Luo, M. Cai, E.S. Chng, J. Feng, X. Zhong, in *INTERSPEECH*. Blind Estimation of Room Impulse Response from Monaural Reverberant Speech with Segmental Generative Neural Network (ISCA, 2023), pp. 2723–2727
18. A. Ratnarajah, I. Ananthabhotla, V.K. Ithapu, P. Hoffmann, D. Manocha, P. Calamia, in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*. Towards Improved Room Impulse Response Estimation for Speech Recognition (IEEE, 2023), pp. 1–5
19. S. Lee, H.S. Choi, K. Lee, in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. Yet Another Generative Model for Room Impulse Response Estimation (IEEE, 2023), pp. 1–5
20. T. Deppisch, J. Ahrens, S.V. Amengual Garí, P. Calamia, in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing Workshops*. Blind Estimation of Spatial Room Impulse Responses Using a Pseudo Reference Signal (IEEE, 2024), pp. 1–5
21. A. Perez-Lopez, A. Politis, E. Gomez, in *IEEE 22nd Int. Workshop on Multimedia Signal Processing*. Blind reverberation time estimation from ambisonic recordings (IEEE, 2020), pp. 1–6
22. N. Meyer-Kahlen, S.J. Schlecht, in *IEEE Int. Workshop on Acoustic Signal Enhancement*. Blind Directional Room Impulse Response Parameterization from Relative Transfer Functions (IEEE, 2022), pp. 1–5
23. T. Deppisch, N. Meyer-Kahlen, S.V. Amengual Garí, Blind identification of binaural room impulse responses from smart glasses. IEEE/ACM Trans. Audio Speech Lang. Process. **4065**, 4052–4065 (2024)
24. J. Su, Z. Jin, A. Finkelstein, in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. Acoustic Matching By Embedding Impulse Responses (IEEE, 2020), pp. 426–430
25. S. Saini, I. Engel, J. Peissig, An end-to-end approach for blindly rendering a virtual sound source in an audio augmented reality environment. EURASIP J. Audio Speech Music Process. **24**, 1–24 (2024)
26. P. Götz, G. D. Santo, S. J. Schlecht, V. Välimäki and E. A. P. Habets, "Matching Reverberant Speech Through Learned Acoustic Embeddings," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE, Barcelona, Spain, 2026), pp. 14577–14581. https://doi.org/10.1109/ICASSP55912.2026.11461386
27. C. Chen, R. Gao, P. Calamia, K. Grauman, in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*. Visual Acoustic Matching (2022), pp. 18836–18846. https://doi.org/10.1109/CVPR52688.2022.01829
28. A. Somayazulu, C. Chen, K. Grauman, in *37th Conference on Neural Information Processing Systems*. Self-Supervised Visual Acoustic Matching (NeurIPS Foundation, 2023), pp. 1–19
29. T. Ajdler, L. Sbaiz, M. Vetterli, Dynamic measurement of room impulse responses using a moving microphone. J. Acoust. Soc. Am. **122**(3), 1636–1645 (2007). https://doi.org/10.1121/1.2766776
30. N. Hahn, S. Spors, in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. Continuous measurement of spatial room impulse responses using a non-uniformly moving microphone (IEEE, 2017), pp. 205–208. https://doi.org/10.1109/WASPAA.2017.8170024
31. F. Katzberg, R. Mazur, M. Maass, P. Koch, A. Mertins, Sound-field measurement with moving microphones. J. Acoust. Soc. Am. **141**(5), 3220–3235 (2017). https://doi.org/10.1121/1.4983093
32. F. Katzberg, R. Mazur, M. Maass, P. Koch, A. Mertins, A compressed sensing framework for dynamic sound-field measurements. IEEE/ACM Trans. Audio Speech Lang. Process. **26**(11), 1962–1975 (2018). https://doi.org/10.1109/TASLP.2018.2851144
33. F. Katzberg, M. Maass, A. Mertins, in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*. Spherical Harmonic Representation For Dynamic Sound-Field Measurements (IEEE, 2021), pp. 426–430
34. K. MacWilliam, T. Dietzen, R. Ali, T. van Waterschoot, State-space estimation of spatially dynamic room impulse responses using a room acoustic model-based prior. Front. Signal Process. **18**, 1–18 (2024). https://doi.org/10.3389/frsip.2024.1426082
35. J. Brunnström, M.B. Møller, M. Moonen, Bayesian sound field estimation using moving microphones. IEEE Open J. Signal Process. 1–10 (2025). https://doi.org/10.1109/OJSP.2025.3526546
36. T. Deppisch, J. Ahrens, S.V. Amengual Garí, P. Calamia, in *32nd European Signal Processing Conference (EUSIPCO)*. Spatial Room Impulse Response Identification from Rotating Equatorial Microphone Arrays (EURASIP, 2024), pp. 116–120
37. T. Deppisch, S.V. Amengual Garí, P. Calamia, J. Ahrens, in *33rd European Signal Processing Conference (EUSIPCO)*. Spatial Room Impulse Response Estimation from a Moving Microphone Array (EURASIP, 2025)
38. M. Poletti, A Unified Theory of Horizontal Holographic Sound Systems. J. Audio Eng. Soc. **48**(12), 1155–1182 (2000)
39. J. Ahrens, H. Helmholz, D.L. Alon, S.V. Amengual Garí, Spherical harmonic decomposition of a sound field using microphones on a circumferential contour around a non-spherical baffle. IEEE/ACM Trans. Audio Speech Lang. Process. **3119**, 3110–3119 (2022)
40. A. Bastine, L. Birnie, T.D. Abhayapala, P. Samarasinghe, V. Tourbabin, in *European Signal Processing Conference*. Ambisonics Capture using Microphones on Head-worn Device of Arbitrary Geometry (EURASIP, 2022), pp. 309–313
41. A. Politis, H. Gamper, in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. Comparing modeled and measurement-based spherical harmonic encoding filters for spherical microphone arrays (IEEE, 2017), pp. 224–228
42. T. Yoshioka, T. Nakatani, Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening. IEEE Trans. Audio Speech Lang. Process. **20**(10), 2707–2720 (2012). https://doi.org/10.1109/TASL.2012.2210879
43. T. Yoshioka, T. Nakatani, in *21st European Signal Processing Conference (EUSIPCO)*. Dereverberation for reverberation-robust microphone arrays (EURASIP, 2013), pp. 1–5
44. S. Braun, E.A. Habets, Online dereverberation for dynamic scenarios using a Kalman filter with an autoregressive model. IEEE Signal Process. Lett. **23**(12), 1741–1745 (2016)
45. L. Drude, J. Heymann, C. Boeddeker, R. Haeb-Umbach, in *Speech Communication - 13th ITG-Fachtagung Sprachkommunikation*. NARA-WPE: A python package for weighted prediction error dereverberation in numpy and tensorflow for online and offline processing (VDE, 2020), pp. 216–220
46. C. Boedekker, T. Nakatani, K. Kinoshita, R. Haeb-Umbach, in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*. Jointly Optimal Dereverberation and Beamforming (IEEE, 2020), p. 216–220
47. S. Hashemgeloogerdi, S. Braun, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Joint Beamforming and Reverberation Cancellation Using a Constrained Kalman Filter with

Multichannel Linear Prediction (IEEE, 2020), pp. 481–485. https://doi.org/10.1109/ICASSP40776.2020.9053785
48. T. Dietzen, S. Doclo, M. Moonen, T. Van Waterschoot, Integrated sidelobe cancellation and linear prediction Kalman filter for joint multi-microphone speech dereverberation, interfering speech cancellation, and noise reduction. IEEE/ACM Trans. Audio Speech Lang. Process. **754**, 740–754 (2020). https://doi.org/10.1109/TASLP.2020.2966869
49. S. Braun, I. Tashev, Low Complexity Online Convolutional Beamforming. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics **2021**(2), 136–140 (2021), IEEE. https://doi.org/10.1109/WASPAA52581.2021.9632780
50. H.L. Van Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory* (Wiley, New York, 2002). https://doi.org/10.1002/0471221104
51. R.O. Schmidt, Multiple emitter location and signal parameter estimation. IEEE Trans. Antennas Propag. **34**(3), 276–280 (1986)
52. N. Epain, C.T. Jin, Spherical harmonic signal covariance and sound field diffuseness. IEEE/ACM Trans. Audio Speech Lang. Process. **24**(10), 1796–1807 (2016). https://doi.org/10.1109/TASLP.2016.2585862
53. C. Schörkhuber, M. Zaunschirm, R. Höldrich, in *Annual German Conference on Acoustics (DAGA)*. Binaural Rendering of Ambisonic Signals via Magnitude Least Squares (DEGA, 2018), pp. 339–342
54. M. Berzborn, R. Bomhardt, J. Klein, J.G. Richter, M. Vorländer, in *Proc. of the German Annual Conference on Acoustics (DAGA)*. The ITA-Toolbox: An Open Source MATLAB Toolbox for Acoustic Measurements and Signal Processing (DEGA, 2017), pp. 222–225
55. A. Lundeby, T.E. Vigran, H. Bietz, M. Vorländer, Uncertainties of measurements in room acoustics. Acta Acust. Acust. **81**(4), 344–355 (1995)
56. B. Bernschütz, in *Annual German Conference on Acoustics (DAGA)*. A Spherical Far Field HRIR/HRTF Compilation of the Neumann KU 100 (DEGA, 2013), pp. 592–595. https://doi.org/10.5281/zenodo.3928297
57. F. Brinkmann, A. Lindau, S. Weinzierl, On the authenticity of individual dynamic binaural synthesis. J. Acoust. Soc. Am. **142**(4), 1784–1795 (2017). https://doi.org/10.1121/1.5005606
58. N. Meyer-Kahlen, S.J. Schlecht, S. Amengual Garí, T. Lokki, Testing auditory illusions in augmented reality: plausibility, transfer-plausibility, and authenticity. J. Audio Eng. Soc. **72**(11), 797–812 (2024). https://doi.org/10.17743/jaes.2022.0178
59. International Telecommunication Union, Radiocommunication Sector, Algorithms to measure audio programme loudness and true-peak audio level, Recommendation ITU-R BS.1770-4 (ITU, Geneva, Switzerland, 2015)
60. International Telecommunication Union, Radiocommunication Sector, Method for the subjective assessment of intermediate quality level of audio systems, Recommendation ITU-R BS.1534-3 (ITU, Geneva, Switzerland, 2015)
61. N. Meyer-Kahlen, A. Hofmann, S. Schlecht, T. Lokki. Dataset of Audiovisual Speech for AR Telepresence Studies (Speech Recordings) (2025). https://doi.org/10.5281/zenodo.15034732
62. E.B. Wilson, Probable Inference, the Law of Succession, and Statistical Inference. J. Am. Stat. Assoc. **22**(158), 209–212 (1927). https://doi.org/10.1080/01621459.1927.10502953
63. T. Nakatani, N. Kamo, M. Delcroix, S. Araki, in *18th Int. Workshop on Acoustic Signal Enhancement*. Multi-Stream Diffusion Model for Probabilistic Integration of Model-Based and Data-Driven Speech Enhancement (IEEE, 2024), pp. 65–69. https://doi.org/10.1109/IWAENC61483.2024.10694664
64. E. Moliner, J.M. Lemercier, S. Welker, T. Gerkmann, V. Valimaki, in *18th Int. Workshop on Acoustic Signal Enhancement*. BUDDy: Single-Channel Blind Unsupervised Dereverberation with Diffusion Models (IEEE, 2024), pp. 120–124. https://doi.org/10.1109/IWAENC61483.2024.10694254

Publisher's Note