



## Enhancing binaural rendering of head-worn microphone arrays through the use of adaptive spatial covariance matching

Janani Fernandez,<sup>1,a)</sup>  Leo McCormack,<sup>1</sup> Petteri Hyvärinen,<sup>1</sup> Archontis Politis,<sup>2</sup> and Ville Pulkki<sup>1</sup> 

<sup>1</sup>Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland

<sup>2</sup>Department of Information Technology and Communication Sciences, Tampere University, Finland

### ABSTRACT:

In this article, the application of spatial covariance matching is investigated for the task of producing spatially enhanced binaural signals using head-worn microphone arrays. A two-step processing paradigm is followed, whereby an initial estimate of the binaural signals is first produced using one of three suggested binaural rendering approaches. The proposed spatial covariance matching enhancement is then applied to these estimated binaural signals with the intention of producing refined binaural signals that more closely exhibit the correct spatial cues as dictated by the employed sound-field model and associated spatial parameters. It is demonstrated, through objective and subjective evaluations, that the proposed enhancements in the majority of cases produce binaural signals that more closely resemble the spatial characteristics of simulated reference signals when the enhancement is applied to and compared against the three suggested starting binaural rendering approaches. Furthermore, it is shown that the enhancement produces spatially similar output binaural signals when using these three different approaches, thus indicating that the enhancement is general in nature and could, therefore, be employed to enhance the outputs of other similar binaural rendering algorithms. © 2022 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.1121/10.0010109>

(Received 27 November 2021; revised 20 February 2022; accepted 18 March 2022; published online 19 April 2022)

[Editor: Efren Fernandez-Grande]

Pages: 2624–2635

### I. INTRODUCTION

The binaural reproduction of sound scenes captured using wearable microphone arrays has gained renewed interest in recent years with such arrays now being integrated into head-worn devices and used for augmented and virtual reality (AR/VR) applications.<sup>1–5</sup> In the context of hearing assistive devices, such as hearing aids, the relatively recent trend of including a data-link between devices has also prompted new proposals that take advantage of this freedom to share signals.<sup>5–7</sup> While there are similarities between the binaural rendering algorithms intended for AR/VR devices and those intended for binaural hearing aids, it should be acknowledged that there are some differing requirements. However, it should be emphasized that one important design criteria, which is relevant to all modern head-worn devices and considered in recent related research,<sup>4–9</sup> is the preservation of sound source localization cues. Furthermore, although such wearable devices have historically been limited in terms of hardware, it may be argued that with the introduction of a data-link in binaural hearing aids and as more sensors are integrated into future models, such devices are converging toward the high-sensor count microphone arrays used for high resolution spatial audio applications.

Traditionally, spherical microphone arrays (SMAs) with uniform sensor distributions have been popular for

spatial audio capturing and reproduction due to their consistent spatial resolution for all directions. SMAs also allow for convenient conversions of the microphone array signals into spherical harmonic signals with numerous signal-independent proposals available for mapping these signals to the binaural channels.<sup>10–13</sup> Other linear methods include binaural beamforming approaches.<sup>2,14</sup> As a result of the linear mapping of signals, these methods retain high signal fidelity. However, the spatial accuracy of the reproduction is inherently limited by the number of microphones in the array. Signal-dependent binaural rendering alternatives, on the other hand, have been demonstrated to surpass linear rendering methods in terms of the perceived spatial accuracy<sup>15–18</sup> when using the same number or fewer input channels. These methods are often built on perceptually motivated sound-field models and estimate the spatial parameters over time and frequency, subsequently using this information to map the input signals to the binaural channels in an adaptive and more informed manner. However, due to the nature of time-frequency processing, the signal fidelity of the output signals may be degraded. Furthermore, in practice, such processing is not always guaranteed to produce output signals that have the intended interchannel relationships dictated by the employed sound-field model. Acknowledging these issues, the concept of employing spatial covariance matching was proposed by Vilkamo *et al.*,<sup>19</sup> which may be considered as a general framework that can be used to enhance spatial audio algorithms by posing them as optimal mixing problems. This alternative approach relies on specifying the

<sup>a)</sup>Electronic mail: [janani.fernandez@aalto.fi](mailto:janani.fernandez@aalto.fi)

interchannel relationships that the output signals should exhibit and working backward to determine the suitable mixing matrices to apply to binaural signals that are produced by an existing binaural rendering method. Such spatial covariance matching based solutions have been shown to attain both high spatial accuracy and signal fidelity.<sup>20–22</sup> However, these previous works only considered the use of SMAs as input, with the application of spatial covariance matching yet to be explored in the context of microphone arrays affixed to wearable devices.

With the integration of microphone arrays becoming increasingly common in AR/VR devices and the adoption of a data-link in binaural hearing aids, there is a growing need for robust and general algorithms that can render binaural signals of high spatial accuracy for application within future devices. Given the benefits of spatial covariance matching based enhancements, as demonstrated for the capture and reproduction of sound-fields using SMAs,<sup>20–22</sup> it is postulated that similar processing may be used for head-worn arrays, which have sensors nonuniformly arranged over irregular and comparatively larger geometries and, thus, existing spherical harmonic domain solutions would be limited by a narrow operating bandwidth.

Therefore, in this article, spatial covariance matching<sup>19</sup> is explored for the task of enhancing the binaural rendering of head-worn microphone arrays. A sound-field model comprising a single source and an isotropic diffuse component per time-frequency tile, as used previously in Refs. 15, 18, and 23, serves as the foundation for this study. Three starting binaural rendering methods, which are inspired by hearing aid related literature,<sup>24–27</sup> are formulated and used to produce initial estimates of the binaural signals. These methods are based solely on signal-domain operations and may, therefore, not be able to produce binaural signals that fully conform to the employed sound-field model; due to, for example, frequency-dependent variations of the employed beamformers and/or their handling of diffuse components in the captured sound scene. It is then upon these initial estimates of the binaural signals that the proposed covariance domain enhancements are applied to obtain refined estimates of the binaural signals. These refined binaural signals aim to more closely match the employed sound-field model and should, therefore, better reproduce the intended spatial cues.

The evaluation of the proposed enhancement involved the construction of an eight-sensor microphone array, which was affixed to the temples of a pair of eyeglasses. The array was then mounted on a dummy head and subsequently measured in a free-field environment. This permitted the simulation of reference binaural signals using the head related transfer functions (HRTFs) of the dummy head, along with the array transfer functions used to simulate the corresponding array recordings to be passed through the rendering algorithms under test. Next, objective evaluations were performed based on a single source in a diffuse-field with varying ratios, followed by subjective listening tests of multisource scenarios with and without simulated room reflections. The results for both of the evaluations indicate that,

when applied to and compared with the initial binaural renders, the proposed enhancements produce binaural signals that more closely resemble the reference binaural signals in the majority of cases.

This article is organized as follows. Section II provides background literature regarding binaural rendering methods intended for hearing assistive and AR/VR devices. Section III details the sound-field model employed for this study. Section IV describes the proposed spatial covariance matching based enhancements, which may be applied to the output signals of the three suggested rendering approaches detailed in Sec. V. Information pertaining to the constructed eight-sensor microphone array is provided in Sec. VI, which is used for the evaluations described in Sec. VII. The evaluation results and discussions are given in Sec. VIII, and the article is concluded in Sec. IX.

## II. BACKGROUND

### A. Binaural rendering in hearing assistive devices

Within the long-established and vast body of literature surrounding hearing aid processing,<sup>6,28</sup> there are references to a number of proposals for rendering the signals of head-worn microphone arrays. Many of the approaches cited tend to focus only on enhancing the signal-to-noise ratio (SNR) with the primary requirements being to improve speech intelligibility<sup>29</sup> and reduce cognitive listening effort.<sup>30</sup> Blind source separation<sup>31</sup> and multichannel Wiener filtering<sup>32</sup> are examples of SNR enhancing algorithms that are well established in practice for monaural or bilateral hearing aid devices. These algorithms, however, have also been shown to lead to degradations in signal quality<sup>28</sup> and often do not seek to preserve the spatial attributes of the original sound scene.<sup>33</sup> In the context of monaural and bilateral hearing aids, however, the benefits arising due to the improved SNR are generally deemed to outweigh these drawbacks.

However, owing to the introduction of a data-link between a pair of modern hearing aid devices, collectively referred to as a binaural hearing aid, the primary design goals for newer devices have gravitated more toward enhancing the SNR and preserving the localization cues.<sup>7,8,24,34,35</sup> Many of the algorithms employed are based on the use of relative-transfer functions (RTFs),<sup>36</sup> which, in the free-field case, refer to the array steering vectors aligned to two reference sensor positions located near the left and right ears. Spatial filters (also known as beamformers) may be steered toward sound sources from the perspective of each reference sensor and routed to the respective ear canals of the listener. The binaural minimum variance distortionless response (MVDR) algorithm<sup>25</sup> is one example of an applicable beamformer for this task. Not only is the SNR enhanced by such processing, but the interaural time difference (ITD) cues are inherently preserved due to the physical location of the two reference sensors. Additionally, many of the interaural level difference (ILD) cues that arise as a result of head-shadowing effects are preserved. The application of binaural linearly constrained minimum variance

(LCMV) beamformers may then extend this cue preservation to also encompass interfering sound sources,<sup>5,37</sup> which can lead to improved speech intelligibility in multi-speaker scenarios due to the spatial release from masking.<sup>38–41</sup> Other localization cue preserving proposals include those based on multichannel Wiener filtering.<sup>9</sup>

Many of these aforementioned approaches, however, do not preserve the monaural localization cues unless the reference sensors are located near the entrance of the listener's ear canals or HRTFs are included as gain constraints during beamforming.<sup>42</sup> In practice, the preservation of monaural cues may be considered less important in the context of assistive hearing devices since meaningful pinna interactions occur above 6 kHz,<sup>43</sup> which may be above the detection threshold of the hearing impaired listener. The spatial attributes of other components in the sound scene, such as reverberation and weakly directional sounds, are also rarely addressed as they directly conflict with the SNR enhancement requirement. Furthermore, when rendering the output binaural signals, retaining a high sound quality is still considered to be less important in the hearing aid processing context when compared to improving speech intelligibility. Although, the addition of more microphones, and/or of higher quality, can help alleviate such issues. The importance of producing spatially accurate auralisations of sound scenes with hearing aid devices was highlighted by Best *et al.*,<sup>44</sup> who drew specific attention to the fact that sound externalization has been overlooked in the hearing aid research literature. This study aims to contribute to this discussion by offering a formal computational framework for rendering sound scenes for hearing aid users, which is easily augmentable for future perceptual studies.

## B. Binaural rendering in AR/VR devices

Another area that has received little scientific attention is the binaural rendering of sound scenes captured by microphone arrays integrated within AR/VR devices; this is likely because commercial devices<sup>45</sup> have only become widely available in recent years. Nonetheless, the recent release of datasets intended for developing algorithms for such devices<sup>46</sup> does highlight that there is growing interest in this area. Contrary to the requirements of binaural hearing aids, the retention of high signal quality is often an important requirement in the AR/VR context along with the preservation of localization cues. Additionally, appropriately reproducing the spatial attributes of reverberation present in the sound scene may be favored over increased SNR. Therefore, while binaural hearing aid algorithms could conceivably be used in the context of AR/VR systems, most proposals have relied on linear signal-independent processing<sup>2,3,14</sup> to forgo the need for source separation and retain high signal quality. However, the spatial accuracy attained through purely linear processing is inherently limited by the number of microphones in the array.

Considering other options, one may look to parametric signal-dependent alternatives,<sup>15–18,47–49</sup> which have been

demonstrated to yield higher perceived spatial accuracy compared to their linear counterparts when using either the same number, or fewer microphones. It is noted, however, that such time-varying processing can introduce signal fidelity degradations. Parametric methods based on the use of spatial covariance matching,<sup>20–22</sup> on the other hand, have been shown to largely address such problems. These solutions rely on computing mixing matrices, which, when applied to an initial estimate of the binaural signals, aim to optimally produce output signals that conform to the specified spatial characteristics while constraining the solutions to retain high signal fidelity.<sup>19</sup> Therefore, contrary to SNR enhancements, which are often sought after in the field of hearing processing, these spatial covariance matching solutions aim, instead, to *spatially enhance* the existing binaural signals. In Ref. 20, a modelless approach was proposed predicated on rendering loose approximations of binaural beamformers derived from SMAs such that they resemble, instead, the spatial selectivity of much sharper but also noisier binaural beamformers while retaining much of the original signal fidelity. The spherical harmonic domain proposals outlined in Refs. 21 and 22 instead used the approach to spatially enhance binaural Ambisonic decoders based on the use of parametric sound-field models. In Ref. 21, the model involved applying sector based processing to softly mix between multiple source estimates and an anisotropic diffuse-field. Whereas, in Ref. 22, the focus was on the application of post-filters to improve the spatial segregation achieved through source and ambient beamforming, which without a constrained spatial covariance matching solution would, otherwise, lead to a reduction in signal fidelity.

Spatial covariance matching based enhancements, however, have not yet been explored within the context of head-worn devices, where the microphones are typically mounted on nonspherical geometries with nonuniform sensor placements. Given additional practical limitations regarding the number of available sensors, which are also spaced more widely apart relative to compact SMAs sensor arrangements, existing spherical harmonic domain solutions would be heavily bandwidth limited, and the patterns of space-domain beamformers may vary greatly with the direction. Therefore, this article differentiates from the aforementioned past works through the formulation of a spatial enhancement that is specifically intended for head-worn devices. Three binaural rendering methods are devised, which are inspired by hearing aid related literature,<sup>24–27</sup> and used to acquire initial estimates of binaural signals based on head-worn microphone array signals. It is demonstrated how space-domain spatial parameter analysis may be conducted to construct target binaural covariance matrices corresponding to a sound-field model comprising a single source mixed with an isotropic diffuse-field. The study also involves an in-depth objective and subjective evaluation of the approach in the context of using a makeshift head-mounted microphone array comprising eight sensors, which represents a potential configuration for future devices.

### III. SOUND-FIELD MODEL

It is assumed that the sound-field is captured via a head-mounted array of  $M$  microphones worn by the listener. The array signals are then transformed into the time-frequency domain  $\mathbf{x}(t, f) \in \mathbb{C}^{M \times 1}$ , where  $f$  denotes the frequency and  $t$  is the down sampled time index. In practice, a short-time Fourier transform (STFT) or a perfect/near-perfect reconstruction filterbank may be employed for this task. For each time-frequency tile, it is assumed that the sound-field may comprise a single dominant source component,  $s$ , an ambient component encapsulating isotropic diffuse noise and reverberation  $\mathbf{d}(t, f) \in \mathbb{C}^{M \times 1}$ , or a combination of the two. The array signal vector may, therefore, be expressed as

$$\mathbf{x}(t, f) = \mathbf{a}(\gamma, f)s(t, f) + \mathbf{d}(t, f), \quad (1)$$

where  $\mathbf{a} \in \mathbb{C}^{M \times 1}$  is the array steering vector for a sound source incident from the direction  $\gamma$ . Note that the array steering vectors may be obtained through free-field measurements or simulations of the array while it is worn by the listener/manikin or modeled analytically by approximating the listener's head as a sphere.<sup>50,51</sup> It is, henceforth, assumed that array steering vectors,  $\mathbf{A} = [\mathbf{a}(\gamma_1), \dots, \mathbf{a}(\gamma_K)] \in \mathbb{C}^{M \times K}$ , are available for a dense grid of  $K$  directions  $\mathbf{\Gamma}_K = [\mathbf{\Gamma}_1, \dots, \mathbf{\Gamma}_K]$ .

Assuming that the source signals are uncorrelated with the diffuse noise and reverberation, the array signal statistics may be expressed via their spatial covariance matrix (SCM) as

$$\begin{aligned} \mathbf{C}_x(f) &= \mathbb{E}[\mathbf{x}(t, f)\mathbf{x}^H(t, f)] \\ &= \mathbf{a}(\gamma, f)\mathbf{a}^H(\gamma, f)\mathbb{E}[|s(t, f)|^2] \\ &\quad + \mathbb{E}[\mathbf{d}(t, f)\mathbf{d}^H(t, f)], \end{aligned} \quad (2)$$

where  $\mathbb{E}$  denotes the expectation operator.

Note that this assumption of a single source mixed with diffuse sound, although simplistic, is often met during practical scenarios provided that the frequency resolution of the transform is sufficiently high and the sound sources present in the scene are sufficiently sparse in frequency and/or over time.

### IV. PROPOSED SPATIAL COVARIANCE MATCHING BASED ENHANCEMENT

In this section, the spatial covariance matching framework under consideration is formulated and applied for the task of enhancing an initial estimate of binaural signals, henceforth referred to as baseline binaural signals, which are obtained with

$$\mathbf{y}_{bl}(t, f) = \mathbf{Q}(f)\mathbf{x}(t, f), \quad (3)$$

where  $\mathbf{Q} \in \mathbb{C}^{2 \times M}$  is the baseline mixing matrix. Suitable candidates for this matrix will be introduced in Sec. V and evaluated later (both with and without the proposed enhancement) in Sec. VII. A block diagram of the overall method is also provided in Fig. 1.

The proposed enhancement is based on the idea that the narrow-band SCMs of the output binaural signals should ideally match those of the target SCMs, which are derived directly through the employed sound-field model. Continuing from the assumptions laid down thus far and by describing the balance between direct and diffuse components using a diffuseness term  $\psi \in [0, 1]$ , the target narrow-band binaural SCMs are given as

$$\begin{aligned} \mathbf{C}_y(f) &= (1 - \psi(t, f))P_{total}(f)\mathbf{h}(\gamma, f)\mathbf{h}(\gamma, f)^H \\ &\quad + \psi(t, f)P_{total}(f)\mathbf{D}_{bin}(f), \end{aligned} \quad (4)$$

where  $P_{total} = \text{tr}[\mathbf{C}_x]$  is the total input signal power;  $\mathbf{h} \in \mathbb{C}^{2 \times 1}$  is the HRTF corresponding to the source direction;  $\mathbf{D}_{bin} \in \mathbb{C}^{2 \times 2} = \mathbf{H}\mathbf{W}\mathbf{H}^H$  is a binaural diffuse coherence matrix (DCM), which is derived from a dense grid of HRTF measurements  $\mathbf{H} = [\mathbf{h}(\gamma_1), \dots, \mathbf{h}(\gamma_K)] \in \mathbb{C}^{2 \times K}$ ; and  $\mathbf{W} \in \mathbb{R}^{K \times K}$  is an optional diagonal matrix of integration weights to account for a nonuniform measurement grid. Note that the inclusion of the binaural DCM serves to enforce the diffuse isotropic properties of the nondirect sounds by imposing the appropriate interaural coherence (IC) cues that would be experienced by the listener while under such conditions.<sup>52</sup> Furthermore, it is noted that the direct-to-diffuse ratio (DDR), which is more commonly used in the signal enhancement literature, is directly related to the employed diffuseness measure as  $\psi = (1 + 10^{\text{DDR}/20})^{-1}$ . The time and frequency indices are also omitted henceforth for the brevity of notation.

Depending on the choice of the baseline mixing matrix, the narrow-band SCMs of the baseline binaural signals  $\mathbf{C}_{bl} = \mathbb{E}[\mathbf{y}_{bl}\mathbf{y}_{bl}^H] \in \mathbb{C}^{2 \times 2}$  may deviate from their respective target narrow-band SCMs. For example, such scenarios may arise due to beamformers encapsulating not only direct

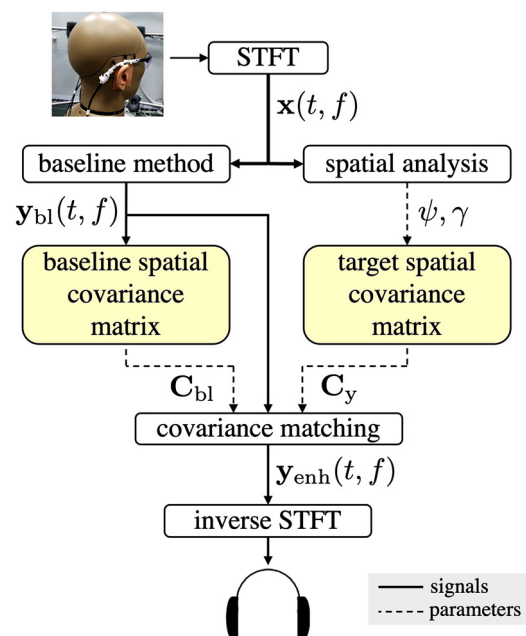


FIG. 1. (Color online) A block diagram of the proposed processing.

sounds but also other sound components (such as reflections) or by rendering the residual components of the scene in a manner that deviates from the isotropic and diffuse characteristics dictated by the employed model. The proposed enhancement approach is, therefore, principally tasked with determining the mixing matrices  $\mathbf{M} \in \mathbb{C}^{2 \times 2}$  to apply to the baseline binaural signals such that the resulting signals directly match the target SCMs and, consequently, also match the employed model

$$\mathbf{y}_{\text{enh}} = \mathbf{M}\mathbf{y}_{\text{bl}} = \mathbf{M}\mathbf{Q}\mathbf{x}, \quad (5)$$

where

$$\begin{aligned} \mathbb{E}[\mathbf{y}_{\text{enh}}\mathbf{y}_{\text{enh}}^H] &= \mathbf{M}\mathbf{C}_{\text{bl}}\mathbf{M}^H = \mathbf{M}\mathbf{Q}\mathbf{C}_x\mathbf{Q}^H\mathbf{M}^H \\ &\approx \mathbf{C}_y. \end{aligned} \quad (6)$$

One option for solving this problem is to first decompose the target and baseline covariance matrices as  $\mathbf{C}_y = \mathbf{K}_y\mathbf{K}_y^H$  and  $\mathbf{C}_{\text{bl}} = \mathbf{K}_{\text{bl}}\mathbf{K}_{\text{bl}}^H$  using, for example, the eigenvalue or Cholesky decomposition, and computing

$$\mathbf{M} = \mathbf{K}_y\mathbf{K}_{\text{bl}}^{-1}. \quad (7)$$

However, although the solution described by Eq. (7) will produce signals that conform to the employed sound-field model, it will not necessarily do so with any consistency across frequency. Therefore, the time-domain representation of this matrix of filters may be ill-conditioned, which would subsequently result in signal fidelity degradations. However, it is also highlighted that these decompositions are not unique, since

$$\begin{aligned} \mathbf{C}_{\text{bl}} &= \mathbf{K}_{\text{bl}}\mathbf{P}_{\text{bl}}\mathbf{P}_{\text{bl}}^H\mathbf{K}_{\text{bl}}^H, \\ \mathbf{C}_y &= \mathbf{K}_y\mathbf{P}_y\mathbf{P}_y^H\mathbf{K}_y^H \end{aligned} \quad (8)$$

hold true for any unitary matrixes  $\mathbf{P}_y$  and  $\mathbf{P}_{\text{bl}}$ . Therefore, it is clear that additional degrees of freedom exist, which may be used to optimize the solution, and it is upon this principle that the covariance domain framework proposed in Ref. 19 aims to fulfill the SCM matching task while also optimally constraining the solution to preserve the high signal fidelity.

In this study, this optimized solution was employed as

$$\mathbf{M}_{\text{opt}} = \mathbf{K}_y\mathbf{V}\mathbf{U}^H\mathbf{K}_{\text{bl}}^{-1}, \quad (9)$$

where  $\mathbf{U}$ ,  $\mathbf{V}$  are obtained from the singular value decomposition  $\mathbf{U}\mathbf{S}\mathbf{V}^H = \mathbf{K}_{\text{bl}}^H\mathbf{G}\mathbf{K}_y$ , where

$$\mathbf{G} = (\text{Diag}[\mathbf{C}_y]\text{Diag}[\mathbf{C}_{\text{bl}}]^{-1})^{-1/2} \quad (10)$$

is a nonnegative diagonal matrix, which is used to normalize the channel energies.

## V. BASELINE BINAURAL RENDERING APPROACHES

With the target sound-field model and SCM matching framework now outlined, three suitable candidate

approaches for the baseline mixing matrix,  $\mathbf{Q}$ , which will be later employed for the evaluations in Sec. VII, are now described.

### A. Using reference sensor signals as baseline signals

The simplest baseline approach applicable to this study is to select two reference microphone signals, which are ideally located nearest to the left,  $\mathbf{x}_l$ , and right,  $\mathbf{x}_r$ , ear canals of the listener, and route the signals directly as  $\mathbf{y}_{\text{bl}} = [\mathbf{x}_l, \mathbf{x}_r]$ . Note that this represents bilateral or binaural hearing aids set to low-power/pass-through modes.<sup>24</sup> Here, the elements of the baseline mixing matrix should be zero, except for the indices mapping the left and right reference sensors to the respective binaural channels, which should be one. For example, if  $M = 8$  and the reference sensors are index one for the left ear and index five for the right ear, the baseline mixing matrix is expressed as

$$\mathbf{Q}^{(\text{basic})} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}. \quad (11)$$

Note that if these reference microphone sensors are located inside the ear canals of the listener, then their signals will capture both the binaural and monaural localization cues and, thus, the computed spatial enhancement mixing matrix will tend toward an identity matrix provided that the captured sound-field conforms to the assumed sound-field model. However, in practice, the reference sensors will likely be situated away from the ear canals. For example, binaural hearing aids will often indicate the top forwardmost sensors as the reference sensors; in which case, the ITD and much of the head-shadowing related ILD cues will be preserved by the baseline signals, and the SCM matching will mainly seek to introduce the missing monaural cues at higher frequencies where pinna interactions are more prevalent. Whereas for an augmented reality device, which may have the sensors located much further away from the listener's ears, the SCM matching solution may require more severe mapping of the input signals to fulfill the target inter-channel dependencies.

### B. Baseline signals using spatial analysis and beamforming

Alternative suitable baseline candidates include those based on beamformers informed by direction-of-arrival (DoA) estimates, which may provide a starting point that is closer to the assumed model. In this work, the filter-and-sum (FaS) beamformer<sup>53</sup> and the binaural MVDR beamformer<sup>25,54</sup> are explored for obtaining the source signal estimates. Then, by assuming that the reference microphones selected by Eq. (11) may be used to approximate diffuse binaural signals when the listener is under such conditions, the source signals and these assumed diffuse signals can be mixed based on the diffuseness parameter. Therefore, the baseline mixing matrix in the case of FaS beamformers is obtained as

$$\mathbf{Q}^{(\text{FaS})} = (1 - \psi)\mathbf{h}(\gamma)\frac{\mathbf{a}(\gamma)^H}{\mathbf{a}^H(\gamma)\mathbf{a}(\gamma)} + c\psi\mathbf{Q}^{(\text{basic})}, \quad (12)$$

where, because it is assumed that the reference sensor signals correspond to diffuse components, an equalization term is also included:

$$c = \sqrt{\frac{\text{tr}[\mathbf{D}_{\text{bin}}]}{\text{tr}[\mathbf{D}_{\text{array}}]}}, \quad (13)$$

where  $\mathbf{D}_{\text{array}} = \mathbf{A}\mathbf{W}\mathbf{A}^H$  is the DCM of the array, which serves to bring the diffuse-field spectral response of the microphone array to reflect, instead, that of the diffuse-field response of the employed HRTFs. Note that balancing between the binaural beamformer and reference signals has been formulated previously in Refs. 27 and 28, through a user-controllable parameter as opposed to the time-frequency-dependent diffuseness term employed in this study.

The third baseline mixing approach explored in this study alternatively involves the use of binaural MVDR beamformers, which are popularly employed in binaural hearing aid device studies,<sup>5,8,24</sup> and is given as

$$\mathbf{Q}^{(\text{MVDR})} = (1 - \psi)\mathbf{h}(\gamma)\frac{\mathbf{a}^H(\gamma)\mathbf{C}_x^{-1}}{\mathbf{a}(\gamma)^H\mathbf{C}_x^{-1}\mathbf{a}(\gamma)} + c\psi\mathbf{Q}^{(\text{basic})}, \quad (14)$$

where it is noted that  $\mathbf{C}_x$  should, in theory, be replaced with an estimate of the array noise covariance matrix to achieve higher noise suppression. However,  $\mathbf{C}_x$  was selected for this study to eliminate problems that may arise due to erroneous source-activity detection.

Note that FaS and binaural MVDR should be able to better capture and represent the source components, provided that the DoA estimates are correct, as both beamformers have the unity gain constraint, and the HRTF directivities are then imposed onto the signals they capture. The target of the SCM matching, therefore, is to bring the interaural cues delivered by the diffuse/reference signals to be more in line with the binaural DCM. However, overall, it is expected that these beamforming based baseline alternatives will produce signals that are closer to the assumed model than those in the basic case represented by Eq. (11) and, thus, will require less severe

corrections to match the SCMs. To illustrate this, a metric describing the deviation of the calculated mixing matrix from an identity matrix was derived as  $\text{tr}[(\mathbf{I} - \mathbf{M})(\mathbf{I} - \mathbf{M})^H]$ . This metric was then computed for two different source directions under diffuse conditions, averaged over time, and plotted over frequency for all three baseline cases, as depicted in Fig. 2. Here, it is evident that both of the beamforming based baselines require less drastic modification to produce output signals with the target interchannel dependencies, especially for high frequencies, although the effect of having a baseline that is closer to the assumed model is shown to be minimal in the later evaluations.

## VI. IMPLEMENTATION

To investigate the performance of the proposed SCM matching based enhancement for the binaural rendering of wearable microphone arrays, eight DPA IMK4060 microphones (DPA Microphones, Denmark) were first affixed to a pair of safety glasses, as depicted in Fig. 3. The safety glasses were mounted onto a KEMAR 45 BC dummy head (GRAS, Denmark), which was placed in an anechoic chamber, with the array directional responses subsequently measured for every 1° on the horizontal plane using the swept-sine technique.<sup>55</sup> An omnidirectional microphone in the same location as the dummy head was used to create a compensation equalization curve to mitigate colorations incurred by the measurement loudspeaker. For processing the signals, the alias-free STFT design, as described in Ref. 57, was selected and configured to employ a window size of 256 samples (sample rate 48 kHz) with 90% window overlap. The three baseline binaural rendering approaches were implemented as described in Sec. V and subjected to the SCM matching based enhancement as detailed in Sec. IV.

To offer further insights into the practical application of the proposed SCM matching solution, the objective evaluation described in Sec. VII A was conducted both with known/Oracle spatial parameters and estimated spatial parameters. Additionally, since the subjective evaluation described in Sec. VII B involved multiple simultaneous sound sources, and due to the single-source assumption of the employed sound-field model, processing based on known parameters would not be meaningful. Therefore, the spatial parameter estimators, which are used to inform the

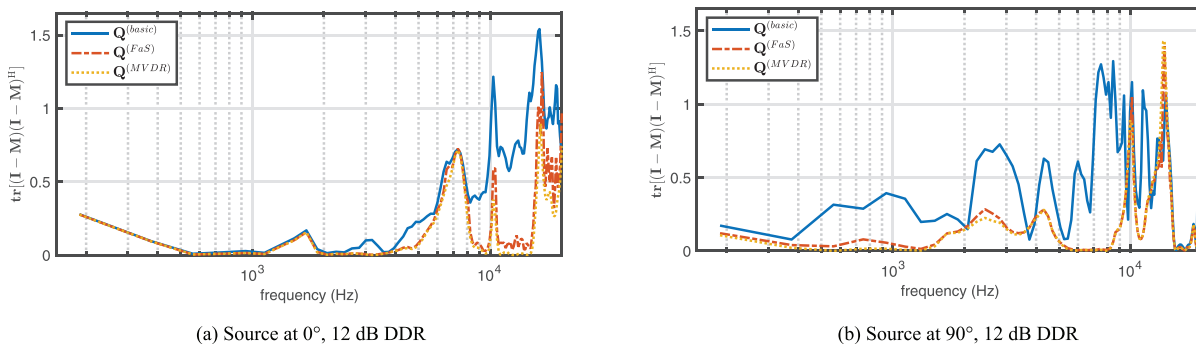


FIG. 2. (Color online) A measure that indicates how much the spatial covariance matrix (SCM) matching mixing matrix deviates from an identity matrix, which is plotted over the frequency for a single-source scenario when using all three of the tested baseline approaches. The lower the deviation, the closer the baseline signals are to the target binaural SCM and, thus, less drastic changes are conducted by  $\mathbf{M}_{\text{opt}}$ .

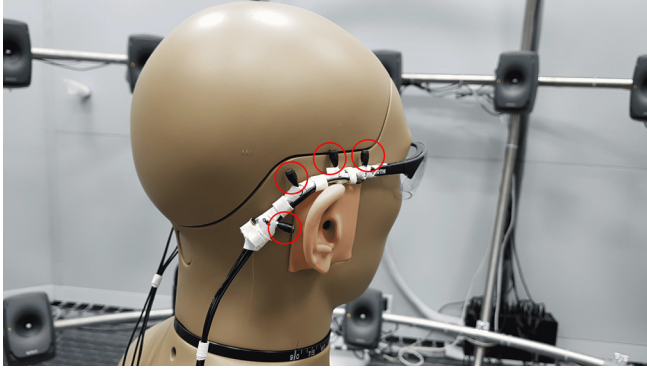


FIG. 3. (Color online) The wearable microphone array employed for the study, viewed from the right side, which is approximately mirrored for the other side of the head. The sensor locations have been highlighted with red circles and had an approximate separation of 2.5 cm between adjacent sensors on each temple and a distance of 14 cm between the respective sensors on each temple.

processing of the adaptive binaural rendering algorithms explored in this study, are now described.

### A. Spatial parameter estimation

The input SCM is first spatially whitened to ensure that it exhibits an identity-like structure when the microphone array is placed under isotropic diffuse-field conditions. The whitened input SCM, thus, conforms to

$$\mathbf{C}_x^{(w)} = \mathbf{T}\mathbf{C}_x\mathbf{T}^H, \quad (15)$$

where  $\mathbf{T} = \mathbf{\Lambda}^{-1/2}\mathbf{R}^H$  given the eigenvalue decomposition of the array DCM,  $\mathbf{D}_{\text{array}} = \mathbf{R}\mathbf{\Lambda}\mathbf{R}^H$  with  $\mathbf{T}\mathbf{D}_{\text{array}}\mathbf{T}^H = \mathbf{I}_M$ . The subspace decomposition with the employed single-source assumption is then applied as

$$\mathbf{C}_x^{(w)} = \mathbf{V}\mathbf{\Sigma}\mathbf{V}^H = \sigma_1\mathbf{v}_1\mathbf{v}_1^H + \sum_{m=2}^M \sigma_m\mathbf{v}_m\mathbf{v}_m^H, \quad (16)$$

where the eigenvalues  $\sigma$  are given in descending order and correspond to their respective eigenvectors  $\mathbf{v}$ . The eigenvectors corresponding to the  $M - 1$  smallest eigenvalues make up the noise subspace  $\mathbf{V}_n \in \mathbb{C}^{M \times (M-1)}$ .

The employed diffuseness parameter estimation, which is based on the COMEDIE algorithm,<sup>57</sup> is then determined through observing the variance of the eigenvalues

$$\psi = 1 - \frac{\beta}{\beta_0}, \quad (17)$$

where the normalization  $\beta_0 = 2(M - 1)$ , the deviation  $\beta = (1/\langle\sigma\rangle) \sum_{m=1}^M |\sigma_m - \langle\sigma\rangle|$ , and the mean  $\langle\sigma\rangle = (1/M) \sum_{m=1}^M \sigma_m$ .

For the DoA estimation, the multiple-signal classification (MUSIC) approach<sup>58</sup> was used, which, given the single-source assumption, is formulated as

$$P_{\text{MUSIC}}(\gamma) = \frac{1}{\|\mathbf{V}_n^H \mathbf{T} \mathbf{a}(\gamma)\|^2} \quad \text{for } \gamma \in \Gamma_K. \quad (18)$$

A peak-finding exercise is then conducted to numerically extract the DoA estimate from the resulting pseudospectrum per frequency.

## VII. EVALUATION

### A. Objective evaluation

The proposed SCM solution was first evaluated objectively in the context of binaural cue preservation. Here, 360 single-source reference binaural signals (1 for each degree on the azimuthal plane) were simulated using whitenoise stimuli and then mixed with cylindrically isotropic diffuse noise to obtain the following DDRs:  $[-60, -6, 0, 6, 12, \text{Inf}]$  dB. Note that the gains required to attain these DDRs were determined based on an omnidirectional receiver, i.e., without the presence of the array. Next, the binaural reference signals and microphone array recordings of these simulated scenarios were obtained by convolving incident plane-waves with either HRTFs or the measured array steering vectors, respectively. The array recordings were then rendered to the binaural channels using the three baseline approaches formulated in Sec. V,  $\mathbf{Q}^{(\text{basic})}$ ,  $\mathbf{Q}^{(\text{FaS})}$ , and  $\mathbf{Q}^{(\text{MVDR})}$ , with and without the proposed SCM matching (CM) enabled, as described in Sec. IV.

The binaural covariance matrix, based on the estimated binaural signals  $\hat{\mathbf{y}}$ , is then given by

$$\hat{\mathbf{C}}_y(f) = \begin{pmatrix} c_{y_{1,1}}(f) & c_{y_{1,2}}(f) \\ c_{y_{2,1}}(f) & c_{y_{2,2}}(f) \end{pmatrix} = \mathbb{E}[\hat{\mathbf{y}}(t,f)\hat{\mathbf{y}}^H(t,f)], \quad (19)$$

and the ILD, interaural phase difference (IPD), IC, and binaural coloration metrics may be computed as<sup>11,56</sup>

$$\text{ILD}(f) = 10 \log_{10} [c_{y_{1,1}}(f)/c_{y_{2,2}}(f)], \quad (20)$$

$$\text{Coloration}(f) = 10 \log_{10} [c_{y_{1,1}}(f) + c_{y_{2,2}}(f)], \quad (21)$$

$$\text{IC}(f) = \frac{\text{real}[c_{y_{1,2}}(f)]}{\sqrt{c_{y_{1,1}}(f)c_{y_{2,2}}(f)}}, \quad (22)$$

$$\text{IPD}(f) = \arg [c_{y_{1,2}}(f)]. \quad (23)$$

Since past studies have demonstrated that binaural hearing aid algorithms may perform differently with known or estimated DoAs,<sup>7,59</sup> the above objective perceptual metrics were computed for all three baselines with and without CM using both known/Oracle parameters and those estimated through the parameter analysis described in Sec. VI A.

### B. Subjective evaluation

A multiple stimulus test was conducted to evaluate the proposed SCM matching solution for the task of binaurally reproducing the microphone signals in more realistic multi-source scenarios. To create the listening test scenes, three source stimuli were placed on the horizontal plane at positions directly to the left, in front, and to the right of the

listener in the simulation. Three different sets of simultaneously played source stimuli were selected, which represent a diverse range of different time-frequency content, (1) a shaker, bass guitar, and strings; (2) a male English speaker, a female English speaker, and a male Danish speaker; and (3) cicadas, a dog barking, and birds tweeting. The stimuli durations were between 13 and 16 s. Two different acoustic settings were selected: anechoic (*dry*) and a moderately reverberant medium sized class room (*rev*). The reference scenarios for the anechoic cases were created by directly convolving the source stimuli with HRIRs in the directions  $[-90 \ 0 \ 90]$  degrees on the horizontal plane. The eight-channel array responses for these same directions were also convolved with the same stimuli to create a synthetic microphone array recording of the same anechoic scene. To reduce the number of test cases for the listening test, and because it is later shown in Sec. VIII that the FaS and MVDR produced similar results in terms of the objective evaluations, only the  $Q^{(basic)}$  and  $Q^{(MVDR)}$  baselines were selected. These array recordings were then rendered using these two baseline methods with the CM enhancement either enabled or disabled.

Reverberant counterparts for the above scenarios were then created using a shoebox room simulator based on the image-source method. The simulator<sup>60</sup> was configured for room dimensions  $10 \times 7 \times 4$  m and had the wall absorption coefficients tuned to resemble a moderately reverberant environment (a broadband T60 of approximately 0.5 s). The listener position was situated directly in the center of the room with the source positions set to 1 m away from the listener position. The simulated direct path and image-source reflections were then convolved with the nearest HRTFs to create the reference reverberant test cases, whereas the nearest microphone array steering vectors were convolved to create synthetic microphone array recordings of the same reverberant scene. These were rendered using the same two baselines, and the CM enhancement was either enabled and disabled, as with the anechoic cases.

In total, there were six test scenes, as summarized in Table I, and five test cases, as summarized in Table II. The listening test was conducted in three parts:

- *Spatial*: In this part of the evaluation, all of the test cases were equalized to the reference case. This was conducted by passing the reference case through the same STFT that was used by the methods under test and determining the reference,  $E_{ref}$ , and test case,  $E_{tc}$ , energies, which were averaged over the whole stimuli duration, followed by computing the equalization gains as  $c_{spatial} = \sqrt{E_{ref}/E_{tc}}$  separately for each bin. This served to mitigate the timbral differences while still retaining the spatial differences between the renderings. The participants were instructed to assess the test cases on a scale from 0 to 100 based on their spatial similarity to the reference (with respect to the source localization, externalization, and reverberation characteristics) and ignore any timbral differences that remained.

TABLE I. The listening test scenes.

Name	Room	Source stimuli
Band_dry	Anechoic	Shaker, bass guitar, strings
Band_rev	Reverberant	Shaker, bass guitar, strings
Speech_dry	Anechoic	Two male and one female speakers
Speech_rev	Reverberant	Two male and one female speakers
Mix_dry	Anechoic	Cicadas, a dog barking, bird calls
Mix_rev	Reverberant	Cicadas, a dog barking, bird calls

- *Timbre*: Here, the reference case was instead duplicated and equalized by each test case,  $c_{timbre} = \sqrt{E_{tc}/E_{ref}}$ , to obtain spatial equivalence across all of the test cases while retaining any timbral colorations that may be introduced by the processing operations associated with the methods under test. The listening subjects were instructed to rate the cases on a scale of 0 to 100 based on their timbral similarity with the reference. It was emphasized that any spatial differences that the listeners perceived should be ignored as equalization can change one’s perception of the spatial cues.<sup>43,61</sup>
- *Overall*: For this part of the listening test, the test cases were simply normalized to the reference based on the broadband root mean squares of the signals, which were averaged across the whole stimuli duration and the binaural channels. The listening subjects were asked to rate the test cases based on their personal preference on a scale of 0 to 100.

In total, 15 test subjects participated in the study, all of whom reported having normal hearing and were naive as to the hypothesis of the study.

### VIII. RESULTS AND DISCUSSION

The results of the objective evaluations are depicted in Fig. 4. The root mean square error (RMSE), computed with respect to the reference binaural cues and averaged over the frequency and following the perceptually motivated equivalent rectangular bandwidths (ERB) scale, is plotted along the y axis with the error bars denoting the standard deviations. The plots given in Fig. 4(a) were calculated based on the known spatial parameters, whereas Fig. 4(b) used the estimated parameters. It can be seen that the results of the IC and IPD evaluations show significant improvements in the RMSE when CM is enabled with known and estimated spatial parameters. On the other hand, while the application of CM reduces the RMSE of the ILD cue when using a known DoA,

TABLE II. The listening test cases.

Name	Rendering
Hidden ref	Ideal binaural receiver
MVDR CM	$Q^{(MVDR)}$ baseline with SCM matching
Basic CM	$Q^{(basic)}$ baseline with SCM matching
MVDR	$Q^{(MVDR)}$ baseline without SCM matching
Basic	$Q^{(basic)}$ baseline without SCM matching



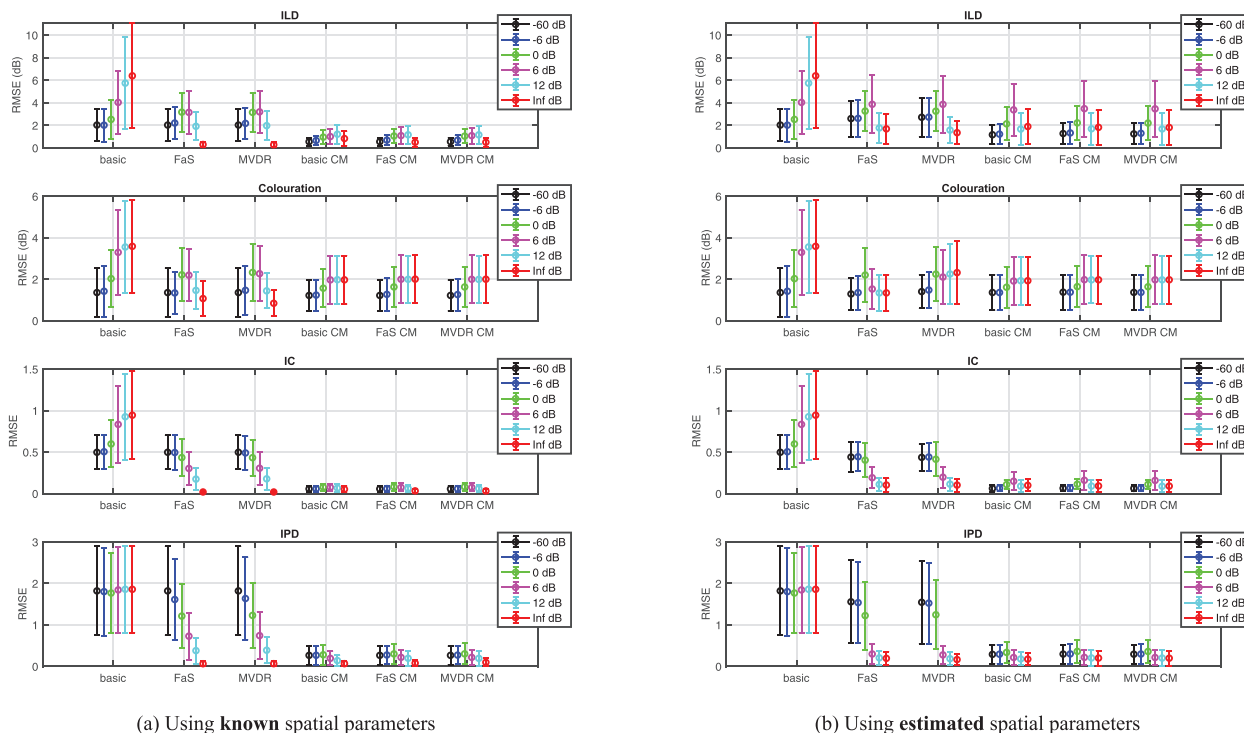


FIG. 4. (Color online) The perceptual metrics results for different DDRs when using either known (a) or estimated (b) spatial parameters.

this reduction in error is not as prevalent when using the estimated DoA. It follows that in this respect, CM is sensitive to errors in the DoA estimation but no more so than the FaS and MVDR baselines. Finally, the RMSE for the coloration does not show significant improvement for the FaS and MVDR baseline methods but does show some improvement for the basic baseline method, although it is noted that colorations of 2 dB are not easily perceptible. Furthermore, this metric does not appear to be affected by the DoA estimation errors.

The results of the subjective evaluation are provided in Fig. 5. It can be observed that, in the majority of cases, the test cases where CM was enabled were rated higher and closer to the reference than when CM was disabled. To provide further insight, statistical analyses were also performed on the data with the exception of the results for the reference stimuli. The analyses were performed using functions from MATLAB’s Statistics and Machine learning toolbox, version 12.1 (The MathWorks, Natick, MA) with the alpha-error significance level set to 0.05 for all of the tests. The Friedman tests (MATLAB function Friedman) were performed separately for each stimulus on both the spatial and timbral ratings. The  $\chi$  square results are listed in Table III, wherein the analyses resulting in  $p$ -values lower than 0.01 are denoted by the symbol “\*\*.” It can be seen that the Friedman tests revealed statistically significant differences between the processing methods in terms of the spatial and timbral subjective evaluation outcomes for all of the test scenes. Subsequently, *ad hoc* multiple comparison tests (MATLAB function multcompare) were performed with a Tukey honest significance difference (HSD) criterion to establish which methods differed significantly from the others.

Statistically significant differences in the spatial ratings were found between the basic and basic CM conditions for the band\_rev ( $p < 0.01$ ), band\_dry ( $p < 0.01$ ), mix\_rev ( $p < 0.01$ ), and mix\_dry test scenes ( $p < 0.01$ ) but did not reach significance for the speech\_rev ( $p > 0.05$ ) and speech\_dry test scenes ( $p > 0.05$ ). Similarly, a statistically significant difference was found between the MVDR and MVDR CM methods for the band\_rev ( $p = 0.05$ ), band\_dry ( $p < 0.01$ ), mix\_rev ( $p < 0.01$ ), and mix\_dry test scenes ( $p < 0.01$ ) but did not reach significance for the speech\_rev ( $p > 0.05$ ) and speech\_dry test scenes ( $p > 0.05$ ). Between the basic CM and MVDR CM methods, only the speech-dry stimulus ( $p < 0.04$ ) spatial ratings case achieved statistical significance. For the timbre section of the tests, the *ad hoc* multiple comparison test with a Tukey HSD criterion applied showed significant differences in the comparison of the ratings for the basic and basic CM methods for all of the test scenes (i.e., band\_rev,  $p < 0.01$ ; band\_dry,  $p < 0.01$ ; mix\_rev,  $p < 0.01$ ; mix\_dry,  $p < 0.01$ ; speech\_rev,  $p = 0.017$ ; and speech\_dry,  $p = 0.032$ ). Meanwhile the comparison of the MVDR and MVDR CM methods only reached statistical significance for the speech-rev scene ( $p < 0.01$ ). There were no statistically significant differences between the ratings for the basic CM and MVDR CM methods.

It should be noted that statistical analyses were not applied to the overall section of the subjective evaluation as the results were considered to be highly subjective, with the ratings of the listener dependent on whether they valued spatial accuracy over timbral fidelity or vice versa. However, a positive trend can be seen in Fig. 5 wherein the majority of the subjects preferred spatial covariance matching solutions over the baseline methods.

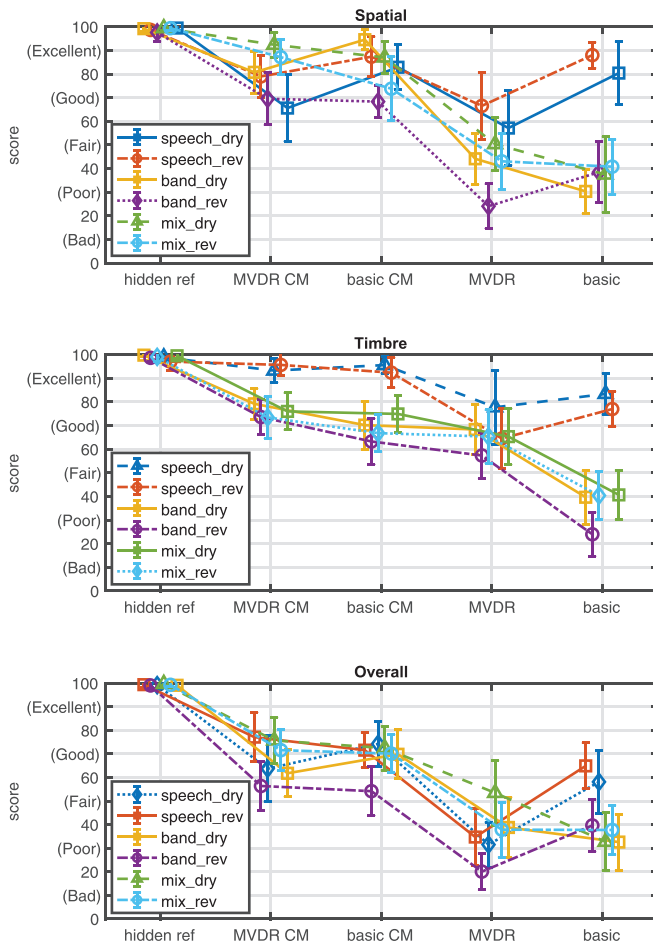


FIG. 5. (Color online) The listening test results, based on 15 test subjects, display the medians and 95% confidence intervals.

The objective and subjective evaluations imply that the application of spatial covariance matching leads to a greater preservation of the spatial cues in comparison to the use of only the baseline techniques. The objective metrics are clearly improved with the application of spatial covariance matching, while the ratings in the subjective evaluation also indicate the improvement is perceptually meaningful given more practical multisource input scenes. The participants in the listening test reported that the spatial accuracy improved with respect to the reference for both the basic and MVDR baseline conditions in the case of the band and mix stimuli. Additionally, whereas the improvement for the speech stimuli was not found to be statistically significant, it can be

TABLE III. The Friedman test results. (\*\*,  $p < 0.01$ ).

Stimulus	$\chi^2_f(3)$	
	Spatial	Timbre
Band_dry	39.36**	21.02**
Band_rev	28.53**	29.03**
Speech_dry	18.81**	11.56**
Speech_rev	12.94**	34.63**
Mix_dry	33.50**	29.30**
Mix_rev	28.00**	25.06**

seen in Fig. 5 that the subjects in the listening test rated the basic and MVDR baseline methods to already be more spatially accurate with respect to the reference for the speech stimuli than for the band and mix stimuli. It follows that, while the application of spatial covariance matching may have improved the spatial accuracy, the improvement was not sufficiently large enough to be statistically significant.

The results of the timbral part of the subjective evaluation show that the use of the spatial covariance matching together with the baseline method reproduces the scene spectral information more faithfully than using the baseline techniques alone. The subjective timbral fidelity appears to be better than that achieved during the objective evaluation. This may be because the objective metrics were calculated using single-source white noise scenarios, whereas the subjective evaluations used three more spectrally diverse sound sources. In the objective evaluation, the beamformers used to generate the baseline prototype signals should have encapsulated the single source with minimal colorations as there were no interferers overlapped by the sidelobes of the beamformers. Additionally, the DoA error rate was presumably lower in the case of a single sound source scene. It follows, therefore, that the coloration for the baseline technique was not noticeably lower in comparison to the coloration for the proposed spatial covariance matching method. On the other hand, during the generation of the listening test stimuli, it is expected that the beamformers will encapsulate some of the signals of the interferers due to a combination of the beamformer sidelobes and DoA estimation errors during periods where the single-source assumption for each time-frequency tile was not met. The spatial covariance matching solution mitigates some of these timbral coloration issues as the target powers  $P_{total}$  are not affected by the DoA estimation errors. Hence, although it is not made clear in the single-source objective evaluation, it is expected that the CM will introduce less timbral coloration, which is apparent in the multiple source subjective evaluation.

Aside from the improved spatial and timbral accuracy of the binaural rendering, it is highlighted that the proposed CM method is still based on the parameterization of the sound scene from the point of view of the listener. Therefore, sound-field modifications may be realized in a computationally efficient way by simply manipulating the spatial parameters prior to reproducing the scene. The sound-field modifications may include rotations, direction-dependent loudness manipulations, and exaggeration of the direct components located in front of the device wearer. Spatial audio effects may also be realized through simple parameter manipulations,<sup>62</sup> which may be desirable for AR/VR applications. Potential avenues to explore in future work, therefore, include an investigation into the effect of manipulating the parameters involved, such as the diffuseness parameter (which has a directly proportional impact on the DDR), or the application of different signal manipulation techniques, such as dynamic range compression, independently applied to the direct and diffuse streams of audio.

## IX. CONCLUSION

This article investigated the application of spatial covariance matching applied to head-worn microphone arrays as a means of enhancing the spatial accuracy when binaurally reproducing the sound scenes that they capture. The sound-field model employed for this study assumes a single sound source per time-frequency index accompanied by an isotropic diffuse component with the proposed enhancements imposed via the spatial covariance matching framework established in Ref. 20. During the study, an eight-sensor microphone array was attached to the temples of a pair of eyeglasses, which was then placed on a dummy head to enable measurements to be taken in a free-field environment to obtain array steering vectors for many directions. These vectors, in conjunction with the HRTFs of the dummy head, were used by the rendering algorithms to produce output binaural signals that may be directly compared with the reference binaural signals. This provided a robust framework for the evaluation of different binaural rendering approaches both with and without the proposed spatial enhancements applied.

In the objective evaluation, the ILDs, IPD, IC, and binaural coloration errors were calculated for renderings of the simulated array recordings in which three baseline techniques were used in isolation and in conjunction with the proposed spatial covariance matching technique. It was found that the application of spatial covariance matching greatly reduced the RMSE for the IC and IPD metrics. The ILD error was also minimized with the application of the proposed enhancement when using known spatial parameters, although this improvement diminished when using the estimated spatial parameters. In the subjective evaluation, a listening test was conducted wherein 15 participants rated multisource sound scenes based on the spatial and timbral similarity with a reference sound scene, as well as overall preference. The results for the listening test indicated that the spatial accuracy of the stimuli significantly improved with the application of spatial covariance matching for the majority of the sound scenes simulated for the test. The timbral attributes were found to be significantly improved over the basic baseline method for all of the stimuli, whereas the improvement for the baseline method that incorporated the MVDR beamforming was found to be statistically significant for only the reverberant speech scenario.

In conclusion, this study demonstrates that spatial covariance matching can be efficiently formulated for application in sound reproduction using head-worn microphone arrays and, on application, produces binaural signals that more closely match those that would, otherwise, have been captured at the ear canals of the listener. In addition, spatial covariance matching improved the spatial attributes and timbral quality of the resultant sound scenes for the basic baseline technique as well as the more complex baseline techniques considered for this study. Although it is noted that the processing does not seek to enhance the SNR and speech intelligibility, with such enhancements, instead,

required to be applied by the baseline method prior to applying the spatial enhancements of the proposed approach. Finally, because the proposed spatial enhancements are still based on the parameterization of the captured sound scene, it is noted that aspects of the rendering may be easily augmented, such as manipulating the direct-to-diffuse balance or applying direction-dependent gains to only the sound sources in the scene.

- <sup>1</sup>P. Calamia, S. Davis, C. Smalt, and C. Weston, "A conformal, helmet-mounted microphone array for auditory situational awareness and hearing protection," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New York (2017), pp. 96–100.
- <sup>2</sup>L. Madmoni, J. Donley, V. Tourbabin, and B. Rafaely, "Beamforming-based binaural reproduction by matching of binaural signals," in *Audio Engineering Society Conference: 2020 AES International Conference on Audio for Virtual and Augmented Reality*, Redmond (2020).
- <sup>3</sup>J. Ahrens, H. Helmholz, D. Lou Alon, and S. Amengual Gari, "Spherical harmonic decomposition of a sound field based on microphones around the circumference of a human head," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY (2021).
- <sup>4</sup>V. Pulkki, L. McCormack, and R. Gonzalez, "Superhuman spatial hearing technology for ultrasonic frequencies," *Sci. Rep.* **11**(1), 1–10 (2021).
- <sup>5</sup>E. Hadad, S. Gannot, and S. Doclo, "Binaural linearly constrained minimum variance beamformer for hearing aid applications," in *IWAENC 2012; International Workshop on Acoustic Signal Enhancement*, VDE, Aachen, Germany (2012), pp. 1–4.
- <sup>6</sup>G. Courtois, P. Marmaroli, M. Lindberg, Y. Oesch, and W. Balade, "Implementation of a binaural localization algorithm in hearing aids: Specifications and achievable solutions," in *Audio Engineering Society Convention 136*, Berlin, Germany (2014).
- <sup>7</sup>H. As'ad, M. Bouchard, and H. Kamkar-Parsi, "A robust target linearly constrained minimum variance beamformer with spatial cues preservation for binaural hearing aids," *IEEE/ACM Trans. Audio, Speech, Lang. Process.* **27**(10), 1549–1563 (2019).
- <sup>8</sup>E. Hadad, D. Marquardt, S. Doclo, and S. Gannot, "Theoretical analysis of binaural transfer function MVDR beamformers with interference cue preservation constraints," *IEEE/ACM Trans. Audio, Speech, Lang. Process.* **23**(12), 2449–2464 (2015).
- <sup>9</sup>D. Marquardt, E. Hadad, S. Gannot, and S. Doclo, "Theoretical analysis of linearly constrained multi-channel Wiener filtering algorithms for combined noise reduction and binaural cue preservation in binaural hearing aids," *IEEE/ACM Trans. Audio, Speech, Lang. Process.* **23**(12), 2384–2397 (2015).
- <sup>10</sup>M. A. Gerzon, "Periphony: With-height sound reproduction," *J. Audio Eng. Soc.* **21**(1), 2–10 (1973).
- <sup>11</sup>M. Zauschirm, C. Schörkhuber, and R. Höldrich, "Binaural rendering of ambisonic signals by head-related impulse response time alignment and a diffuseness constraint," *J. Acoust. Soc. Am.* **143**(6), 3616–3627 (2018).
- <sup>12</sup>C. Schörkhuber, M. Zauschirm, and R. Höldrich, "Binaural rendering of ambisonic signals via magnitude least squares," in *Proceedings of the DAGA* (2018), Vol. 44, pp. 339–342.
- <sup>13</sup>T. McKenzie, D. T. Murphy, and G. Kearney, "Interaural level difference optimization of binaural ambisonic rendering," *Appl. Sci.* **9**(6), 1226 (2019).
- <sup>14</sup>E. Rasumow, M. Blau, S. Doclo, S. van de Par, M. Hansen, D. Püschel, and V. Mellert, "Perceptual evaluation of individualized binaural reproduction using a virtual artificial head," *J. Audio Eng. Soc.* **65**(6), 448–459 (2017).
- <sup>15</sup>V. Pulkki, A. Politis, M.-V. Laitinen, J. Vilkamo, and J. Ahonen, "First-order directional audio coding (DirAC)," in *Parametric Time-Frequency Domain Spatial Audio*, edited by V. Pulkki, S. Delikaris-Manias, and A. Politis (Wiley, Newark, NJ, 2017), pp. 89–138.
- <sup>16</sup>S. Berge and N. Barrett, "High angular resolution planewave expansion," in *Proceedings of the 2nd International Symposium on Ambisonics and Spherical Acoustics*, Paris, France (2010), pp. 6–7.
- <sup>17</sup>A. Politis, S. Tervo, and V. Pulkki, "COMPASS: Coding and multidirectional parameterization of ambisonic sound scenes," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Alberta, Canada (2018), pp. 6802–6806.

- <sup>18</sup>C. Schörkhuber and R. Höldrich, “Linearly and quadratically constrained least-squares decoder for signal-dependent binaural rendering of ambisonic signals,” in *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*, York, UK (2019).
- <sup>19</sup>J. Vilkamo, T. Bäckström, and A. Kuntz, “Optimized covariance domain framework for time–frequency processing of spatial audio,” *J. Audio Eng. Soc.* **61**(6), 403–411 (2013).
- <sup>20</sup>S. Delikaris-Manias, J. Vilkamo, and V. Pulkki, “Parametric binaural rendering utilizing compact microphone arrays,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia (2015), pp. 629–633.
- <sup>21</sup>A. Politis, L. McCormack, and V. Pulkki, “Enhancement of ambisonic binaural reproduction using directional audio coding with optimal adaptive mixing,” in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New York (2017), pp. 379–383.
- <sup>22</sup>L. McCormack and S. Delikaris-Manias, “Parametric first-order ambisonic decoding for headphones utilizing the cross-pattern coherence algorithm,” in *EAA Spatial Audio Signal Processing Symposium*, Paris, France (2019), pp. 173–178.
- <sup>23</sup>S. Braun, M. Torcoli, D. Marquardt, E. A. Habets, and S. Doclo, “Multichannel dereverberation for hearing aids with interaural coherence preservation,” in *2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC)* Juan-les-Pins, France (2014), pp. 124–128.
- <sup>24</sup>S. Doclo, S. Gannot, M. Moonen, and A. Spriet, “Acoustic beamforming for hearing aid applications,” in *Handbook on Array Processing and Sensor Networks*, edited by S. Haykin and K. J. Ray Liu (Wiley-IEEE Press, Hoboken, NJ, 2010), pp. 269–302.
- <sup>25</sup>S. M. Golan, S. Gannot, and I. Cohen, “A reduced bandwidth binaural MVDR beamformer,” in *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Tel-Aviv, Israel (2010).
- <sup>26</sup>T. Van den Bogaert, S. Doclo, J. Wouters, and M. Moonen, “Speech enhancement with multichannel wiener filter techniques in multimicrophone binaural hearing aids,” *J. Acoust. Soc. Am.* **125**(1), 360–371 (2009).
- <sup>27</sup>R. M. Corey, “Microphone array processing for augmented listening,” Ph.D. thesis, University of Illinois at Urbana-Champaign, 2019.
- <sup>28</sup>M. A. Akeroyd, “An overview of the major phenomena of the localization of sound sources by normal-hearing, hearing-impaired, and aided listeners,” *Trends Hear.* **18**, 2331216514560442 (2014).
- <sup>29</sup>Y. Wang, A. Narayanan, and D. Wang, “On training targets for supervised speech separation,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.* **22**(12), 1849–1858 (2014).
- <sup>30</sup>A. Bronkhorst and R. Plomp, “The effect of head-induced interaural time and level differences on speech intelligibility in noise,” *J. Acoust. Soc. Am.* **83**(4), 1508–1516 (1988).
- <sup>31</sup>H. Buchner, R. Aichner, and W. Kellermann, “Blind source separation for convolutive mixtures: A unified treatment,” in *Audio Signal Processing for Next-Generation Multimedia Communication Systems* (Springer, Boston, MA, 2004), pp. 255–293.
- <sup>32</sup>A. Spriet, M. Moonen, and J. Wouters, “Spatially pre-processed speech distortion weighted multi-channel Wiener filtering for noise reduction,” *Signal Process.* **84**(12), 2367–2387 (2004).
- <sup>33</sup>T. Van den Bogaert, T. J. Klases, M. Moonen, L. Van Deun, and J. Wouters, “Horizontal localization with bilateral hearing aids: Without is better than with,” *J. Acoust. Soc. Am.* **119**(1), 515–526 (2006).
- <sup>34</sup>T. Van den Bogaert, S. Doclo, J. Wouters, and M. Moonen, “The effect of multimicrophone noise reduction systems on sound source localization by users of binaural hearing aids,” *J. Acoust. Soc. Am.* **124**(1), 484–497 (2008).
- <sup>35</sup>X. Leng, J. Chen, and J. Benesty, “On the compromise between noise reduction and speech/noise spatial information preservation in binaural speech enhancement,” *J. Acoust. Soc. Am.* **149**(5), 3151–3162 (2021).
- <sup>36</sup>S. Gannot, D. Burshtein, and E. Weinstein, “Signal enhancement using beamforming and nonstationarity with applications to speech,” *IEEE Trans. Signal Process.* **49**(8), 1614–1626 (2001).
- <sup>37</sup>A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, J. Jensen, and M. Guo, “Binaural beamforming using pre-determined relative acoustic transfer functions,” in *2017 25th European Signal Processing Conference (EUSIPCO)*, Kos, Greece (2017), pp. 1–5.
- <sup>38</sup>M. L. Hawley, R. Y. Litovsky, and J. F. Culling, “The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer,” *J. Acoust. Soc. Am.* **115**(2), 833–843 (2004).
- <sup>39</sup>A. W. Bronkhorst, “The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions,” *Acta Acust. Acust.* **86**(1), 117–128 (2000).
- <sup>40</sup>R. Beutelmans and T. Brand, “Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners,” *J. Acoust. Soc. Am.* **120**(1), 331–342 (2006).
- <sup>41</sup>R. Y. Litovsky, “Spatial release from masking,” *Acoust. Today* **8**(2), 18–25 (2012).
- <sup>42</sup>S. Goetze, T. Rohdenburg, V. Hohmann, B. Kollmeier, and K.-D. Kammeyer, “Direction of arrival estimation based on the dual delay line approach for binaural hearing aid microphone arrays,” in *2007 International Symposium on Intelligent Signal Processing and Communication Systems*, Xiamen, China (2007), pp. 84–87.
- <sup>43</sup>J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization* (MIT Press, Cambridge, MA, 1997).
- <sup>44</sup>V. Best, R. Baumgartner, M. Lavandier, P. Majdak, and N. Kopčo, “Sound externalization: A review of recent research,” *Trends Hear.* **24**, 2331216520948390 (2020).
- <sup>45</sup>For example, the Ray-Ban Stories sunglasses, developed in collaboration with Facebook Reality Labs. Details can be found at the press release available at <https://tech.fb.com/ray-ban-and-facebook-introduce-ray-ban-stories-first-generation-smart-glasses/> (Last viewed February 15, 2022).
- <sup>46</sup>J. Donley, V. Tourbabin, J.-S. Lee, M. Broyles, H. Jiang, J. Shen, M. Pantic, V. K. Ithapu, and R. Mehra, “Easycorn: An augmented reality dataset to support algorithms for easy communication in noisy environments,” *arXiv:2107.04174* (2021).
- <sup>47</sup>O. Thiergart, M. Taseska, and E. A. Habets, “An informed parametric spatial filter based on instantaneous direction-of-arrival estimates,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.* **22**(12), 2182–2196 (2014).
- <sup>48</sup>J. Nikunen, A. Diment, T. Virtanen, and M. Vilermo, “Binaural rendering of microphone array captures based on source separation,” *Speech Commun.* **76**, 157–169 (2016).
- <sup>49</sup>V. Gunnarsson and M. Sternad, “Binaural auralization of microphone array room impulse responses using causal Wiener filtering,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.* **29**, 2899–2914 (2021).
- <sup>50</sup>E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography* (Elsevier Science & Technology, San Diego, 1999).
- <sup>51</sup>H. Teutsch, *Modal Array Signal Processing: Principles Applications of Acoustic Wavefield Decomposition* (Springer, New York, 2007), Vol. 348.
- <sup>52</sup>A. Politis, “Diffuse-field coherence of sensors with arbitrary directional responses,” *arXiv:1608.07713* (2016).
- <sup>53</sup>J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, “Robust localization in reverberant rooms,” in *Microphone Arrays*, edited by M. Brandstein and D. Ward (Springer, Berlin, Heidelberg, 2001), pp. 157–180.
- <sup>54</sup>O. L. Frost, “An algorithm for linearly constrained adaptive array processing,” *Proc. IEEE* **60**(8), 926–935 (1972).
- <sup>55</sup>A. Farina, “Simultaneous measurement of impulse response and distortion with a swept-sine technique,” in *Audio Engineering Society Convention 108*, Paris, France (2000).
- <sup>56</sup>J. Vilkamo and T. Bäckström, “Time-frequency processing: Methods and tools,” in *Parametric Time-Frequency Domain Spatial Audio*, edited by V. Pulkki, S. Delikaris-Manias, and A. Politis (Wiley, Newark, NJ, 2017), pp. 1–24.
- <sup>57</sup>N. Epain and C. T. Jin, “Spherical harmonic signal covariance and sound field diffuseness,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.* **24**(10), 1796–1807 (2016).
- <sup>58</sup>R. Schmidt, “Multiple emitter location and signal parameter estimation,” *IEEE Trans. Antennas Propag.* **34**(3), 276–280 (1986).
- <sup>59</sup>A. Kuklasinski and J. Jensen, “Multichannel Wiener filters in binaural and bilateral hearing aids—speech intelligibility improvement and robustness to doa errors,” *J. Audio Eng. Soc.* **65**(1/2), 8–16 (2017).
- <sup>60</sup>The employed image-source based shoebox room simulator is available at <https://github.com/polarch/shoebox-roomsim> (Last viewed February 11, 2022).
- <sup>61</sup>J. Blauert, “Sound localization in the median plane,” *Acta Acust. Acust.* **22**(4), 205–213 (1969).
- <sup>62</sup>L. McCormack, A. Politis, and V. Pulkki, “Parametric spatial audio effects based on the multi-directional decomposition of ambisonic sound scenes,” in *Proceedings of the 24th International Conference on Digital Audio Effects (DAFx20in21)*, Vienna, Austria (2021), pp. 214–221.