OPEN ACCESS Freely available online

M. B. Galindo, P. Coleman, and P. J.B. Jackson, "Microphone Array Geometries for Horizontal Spatial Audio Object Capture With Beamforming"
J. Audio Eng. Soc., vol. 68, no. 5, pp. 324–337, (2020 May.).
DOI: https://doi.org/10.17743/jaes.2020.0025

Microphone Array Geometries for Horizontal Spatial Audio Object Capture With Beamforming

MIGUEL BLANCO GALINDO[1], PHILIP COLEMAN[2], AND PHILIP J.B. JACKSON[1]
(mblancogalindo@gmail.com) (p.d.coleman@surrey.ac.uk) (p.jackson@surrey.ac.uk)

[1]*Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, UK*
[2]*Institute of Sound Recording, University of Surrey, Guildford, UK*

Microphone array beamforming can be used to enhance and separate sound sources, with applications in the capture of object-based audio. Many beamforming methods have been proposed and assessed against each other. However, the effects of compact microphone array design on beamforming performance have not been studied for this kind of application. This study investigates how to maximize the quality of audio objects extracted from a horizontal sound field by filter-and-sum beamforming, through appropriate choice of microphone array design. Eight uniform geometries with practical constraints of a limited number of microphones and maximum array size are evaluated over a range of physical metrics. Results show that baffled circular arrays outperform the other geometries in terms of perceptually relevant frequency range, spatial resolution, directivity, and robustness. Moreover, a subjective evaluation of microphone arrays and beamformers is conducted with regards to the quality of the target sound, interference suppression, and overall quality of simulated music performance recordings. Baffled circular arrays achieve higher target quality and interference suppression than alternative geometries with wideband signals. Furthermore, subjective scores of beamformers regarding target quality and interference suppression agree well with beamformer on-axis and off-axis responses; with wideband signals, the superdirective beamformer achieves the highest overall quality.

0 INTRODUCTION

Object-based audio is a spatial audio representation in which the sound field is comprised of individual objects [1]. The advantage of this paradigm over channel-based and scene-based approaches is that objects can be controlled individually before being rendered, allowing for compatibility with arbitrary reproduction systems and user personalization [1, 2]. This results in an improved listening experience, e.g., by controlling the dialogue-to-background-sound level [3], automatic optimal rendering exploiting the object metadata's semantic information [4], and customization for hearing-impaired people [5].

Sound sources composing audio objects can be captured individually with minimum spill by separate multitracked or close microphone recordings [6]. However, there are situations where close-miked recordings may not be feasible due to production constraints: insufficient resources (microphones, preamplifiers, digital converters, etc.); restricted set-up time; impractical to wear clip microphone and transmitter; and/or moving sources that cannot be followed dynamically with a microphone. In all these situations, spatial filtering (or beamforming) with a single, compact microphone array to isolate [7] or enhance [1] certain audio objects in the sound scene may be desirable.

Many of the findings from the beamforming literature apply to object capture with microphone arrays. The array output depends on the beamforming method and physical array design. Beamforming methods can be classified as [8] filter-and-sum beamformers (FSBs), differential microphones (DMs), and modal beamformers (MBs). Within these approaches, numerous contributions in filter design optimization based on different criteria have been proposed and reviewed [9–13]. However, it is not obvious how to design the microphone array to maximize the beamforming performance with respect to various metrics. The choice of microphone arrays in the literature can be to simplify the formulation, e.g., linear arrays with FSBs [9] and DMs [14, 15, 12] or circular and spherical arrays with MBs [16, 17], or to show an improved performance regarding a single physical metric of interest (e.g., resolution [18, 19] or sidelobe [20]), thus only partially rating their performance.

Most of the contributions relate to physical performance measures. While there exist some perceptual studies in

beamforming, they either relied on objective models trained on perceptual features [7, 21–23] such as PEASS [24] or performed a listening test only using speech and without stating an attribute to be rated [25].

The aim of this paper is to determine the uniform microphone array geometry that maximizes the quality of audio objects extracted from a horizontal sound field, since most sound scenes encountered in practice have much greater variation in azimuth than in elevation. The two main contributions are as follows:

1) We perform a thorough comparative evaluation of the physical beamforming performance of compact array designs. Uniform arrays are for the first time compared based on the two most practical design constraints: a given number of microphones, which impacts on the cost and processing power of the system; and a maximum array size, determining its compactness and portability. To achieve such a consistent and systematic comparison, their performance with widely used space-domain beamformers over a range of metrics is assessed through simulations, since off-the-shelf arrays do not have the same number of microphones or comparable dimensions [26]. As a result, we show which array is the optimal uniform array geometry in terms of perceptually relevant frequency range, resolution, directivity, and robustness for horizontal sound fields.

2) We conduct the first formal comparative listening evaluations of microphone array beamforming for audio applications. Two main experiments are undertaken assessing different arrays and beamformers in terms of quality of the target sound, interference suppression, and overall quality for simulated music performance recordings. Results show how critical effects on target quality and interference suppression are on the overall quality and rankings. A further listening test is employed to discriminate among the best performing arrays, obtaining statistical significance of the favored array.

The paper is structured as follows: Sec. 1 reviews the signal model, microphone array designs, and beamforming methods to assess the arrays; Sec. 2 presents the evaluation metrics, setup, and results from the physical analysis; Sec. 3 presents the methodology and results from the perceptual evaluation of beamformers and arrays; and Sec. 4 discusses the physical and perceptual results and their implications for object capture. Finally, the main conclusions are highlighted in Sec. 5.

1 BACKGROUND

This section introduces the signal model, array manifold transfer functions for open, cylindrical, and spherical baffles, and beamformers used for the array evaluation.

1.1 Signal Model

Consider a collection of S sound source signals expressed in the frequency domain as $\mathbf{s}(\omega) = [s_1(\omega), s_2(\omega), \ldots, s_S(\omega)]^T$ in the far field from an M-element microphone array at S different directions. The signal captured by the array $\mathbf{x}(\omega) = [x_1(\omega), x_2(\omega), \ldots, x_M(\omega)]^T$ can be expressed as

$$\mathbf{x}(\omega) = \mathbf{A}(\omega)\mathbf{s}(\omega) + \mathbf{v}(\omega), \tag{1}$$

where $\mathbf{A}(\omega) = [\mathbf{a}_1(\omega), \mathbf{a}_2(\omega), \ldots, \mathbf{a}_S(\omega)]$ is the array manifold steering matrix representing the transfer function between each sound source $s_s(\omega)$ and each of the microphones; $\mathbf{a}_s(\omega) = [a_{1s}(\omega), a_{2s}(\omega), \ldots, a_{Ms}(\omega)]^T$ is the equivalent vector between sound source $s_s(\omega)$ and all microphones; and $\mathbf{v}(\omega) = [v_1(\omega), v_2(\omega), \ldots, v_M(\omega)]^T$ is a noise signal with arbitrary spatial characteristics [27, 28].

The output signal of a FSB $y(\omega)$ is obtained by filtering and summing the array input $\mathbf{x}(\omega)$ with the beamformer weights $\mathbf{w}(\omega) = [w_1(\omega), w_2(\omega), \ldots, w_M(\omega)]^T$

$$y(\omega) = \mathbf{w}^H(\omega)\mathbf{x}(\omega). \tag{2}$$

The directional response $\mathbf{d}(\omega)$ can be regarded as the transfer function between a source signal at any point over the sound field considered and the array output [29]

$$\mathbf{d}(\omega) = \mathbf{w}^H(\omega)\mathbf{A}(\omega), \tag{3}$$

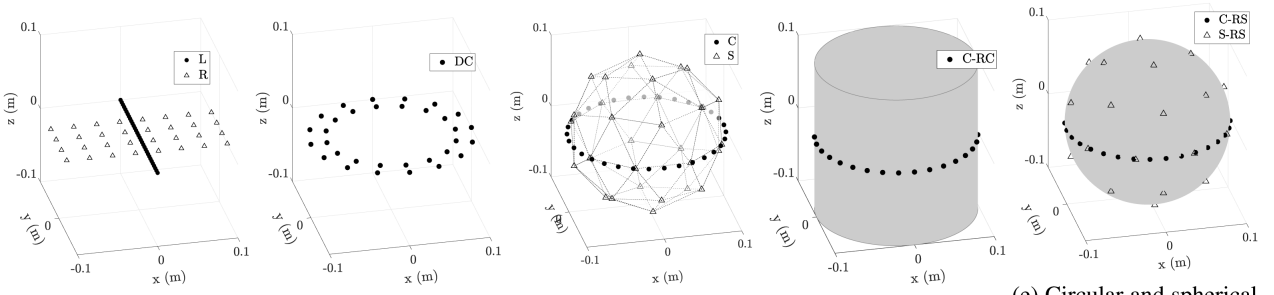
where $\mathbf{d}(\omega) = [d(\omega, \Omega_1),\ d(\omega, \Omega_2),\ \ldots,\ d(\omega, \Omega_S)]$ is the response at each angle Ω_s over Ssteering directions, with $\Omega \equiv (\theta, \phi)$ comprising the inclination and azimuth angles, respectively. The steering matrix $\mathbf{A}(\omega)$ depends, among other things, on the microphone positions and potential acoustic wave phenomena (e.g., diffraction and scattering), thus leading to different analytical expressions for the uniform array designs included in this study.

1.2 Microphone Arrays

There exist many possible designs for compact microphone arrays. Uniform linear arrays are commonly used due to their ability to simplify the formulation of a proposed beamformer or feature to be shown. However, they are unable to resolve the direction of arrival (DoA) in three dimensions (due to unavoidable front–back and elevation ambiguities). Horizontal planar arrays also feature up–down confusion. However, they have been used for noise control applications to reduce the sidelobes [20].

On the other hand, circular and spherical arrays have been widely used for 2D and 3D sound field capture in the circular/spherical harmonic domain, i.e., higher-order Ambisonic (HOA) and MB. While all circular/spherical array designs are sensitive to noise at low frequencies, their open counterparts are ill-conditioned at frequencies at which Bessel function singularities occur [17]. The latter can be remedied with dual and multiple-radius spheres/circles [30–32] or a combination of pressure and velocity microphones [33], at the cost of at least twice as many microphones, or using cardioid microphones, although their directivity is frequency dependent in practice [31, 34]. Alternatively, mounting the array on a cylindrical or spherical baffle also overcomes the robustness issue [17].

This study performs a systematic evaluation of the performance of eight of these array geometries (see Fig. 1): linear (L), rectangular (R), circular (C), dual-circular (DC), spherical (S), circular on rigid cylinder (C-RC), circular on



(a) Linear (L) and rectangular (R) (b) Dual-circular (DC) (c) Circular (C) and spherical (S) (d) Circular on rigid cylinder (C-RC) (e) Circular and spherical on rigid sphere (C-RS, S-RS)

Fig. 1. Microphone array designs under study, $M = 32$ and $r = 0.1$ m.

rigid sphere (C-RS) and spherical on rigid sphere (S-RS). These were designed to provide an unbiased comparison by setting the two most practical design factors: the number of (omnidirectional) microphones $M = 32$, impacting on the cost and processing power of the array, and their aperture limit (maximum distance between two microphones), determining its compactness and portability, by setting a maximum radius of $r = 0.1$ m. This results in different spacing Δd (minimum distance between two microphones). The inner radius of DC is 0.08 m.

Unbaffled arrays (L, R, C, DC, and S) are modeled as

$$a_m^{\text{op}}(\mathbf{k}, \mathbf{r}_m) = e^{i\mathbf{k}^T\mathbf{r}_m}, \tag{4}$$

where a_m^{op} is the open array manifold for a plane wave traveling from the sound source to the mth microphone, $\mathbf{k} = k\,[\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta]^T$ is the wavenumber vector indicating the DoA of the source in spherical coordinates for a time harmonic dependence $e^{i\omega t}$ [35], $k = \omega/c$ where c is the speed of sound, $i = \sqrt{-1}$ and $\mathbf{r}_m$ is the mth microphone position which can be expressed in Cartesian, spherical, and cylindrical coordinates: $\mathbf{r}_m^{\text{car}} = [x_m, y_m, z_m]^T$; $\mathbf{r}_m^{\text{sph}} = r_m\,[\sin\theta_m\cos\phi_m, \sin\theta_m\sin\phi_m, \cos\theta_m]^T$; $\mathbf{r}_m^{\text{cyl}} = [r_m\cos\phi_m, r_m\sin\phi_m, z_m]^T$. In cylindrical coordinates, Eq. (4) can be approximated as a Fourier series of order N_a [36]

$$a_m^{\text{op}}(\mathbf{k}, \mathbf{r}_m) = e^{ik\cos\theta z_m}\sum_{n=-N_a}^{N_a} i^n J_n(kr_m\sin\theta)e^{in(\phi_m-\phi)}, \tag{5}$$

where J_n is the Bessel function of order n. While open array manifolds can be expressed either in complex exponential (Eq. (4)) or harmonic decomposition (Eq. (5)) forms, the sound pressure on baffled arrays can only be represented via inverse cylindrical or spherical harmonic transforms.

The transfer function of a microphone array on an infinitely long rigid cylinder in a horizontal plane results in accurate approximation to its finite-length counterpart, provided its length is at least 2.8 times the radius [37]. Using this assumption, the array manifold of C-RC is [18, 38]

$$a_m^{\text{RC}}(\mathbf{k}, \mathbf{r}_m) = \frac{2e^{ik\cos\theta z_m}}{i\,\pi kr_m\sin\theta}\sum_{n=-N_a}^{N_a}\frac{i^n\,e^{in(\phi_m-\phi)}}{H_n^{(2)'}(kr_m\sin\theta)}, \tag{6}$$

where $H_n^{(2)'}$ is the derivative of the Hankel function of the second kind and $\theta \notin \{0, \pi\}$.

The plane wave transfer function for a microphone array mounted on a rigid sphere (C-RS and S-RS) is

$$a_m^{\text{RS}}(\mathbf{k}, \mathbf{r}_m) = \frac{1}{i\,(kr_m)^2}\sum_{n=0}^{N_a}\frac{i^n\,(2n+1)}{h_n^{(2)'}(kr_m)}P_n(\cos\Theta), \tag{7}$$

where $h_n^{(2)'}$ is the derivative of the spherical Hankel function of the second kind, and $\Theta = \Omega_m - \Omega$, P_n is the Legendre polynomial of order n comprising the sum over the spherical harmonics of all degrees $|p| \leqslant n$. For S and S-RS, sensors are nearly uniformly distributed [17], placed in the center of the faces of a truncated icosahedron [39].

Eqs. (4–7) assume a plane wave incidence which is valid for sources at a distance $R \geqslant 8r^2f/c$ [40], i.e., 2.3 m for $r = 0.1$ m and frequencies up to 10 kHz. This is satisfied in a practical performance capture where the sound sources will be spaced apart from each other while being evenly distant from the array as that presented in Sec. 3. Note Eqs. (5), (6), and (7) are approximations of the equivalent infinite series which result in accurate representation up to a maximum frequency $f_{\max}$, provided $N_a = \lceil 1.1k_{\max}r_m \rceil$ [41], where $f_{\max} = 20$ kHz in this study.

1.3 Beamforming

Four beamformers are used to evaluate the performance of the arrays: delay-and-sum beamformer (DSB), superdirective beamformer (SDB), minimum variance distortionless response beamformer (MVDRB), and least-squares beamformer (LSB). They are optimal in some way as reviewed below.

1.3.1 Delay and Sum

The simplest FSB is DSB, whose weights are the array manifold vector at the look direction $\mathbf{a}_l(k, \mathbf{r}) \equiv \mathbf{a}(k, \Omega_l, \mathbf{r})$,

$$\mathbf{w}_{\text{DSB}}(\omega) = \frac{1}{M}\mathbf{a}_l(k, \mathbf{r}), \tag{8}$$

to steer the array in that direction. DSB is very robust against deviations in microphone characteristics [28, 42].

1.3.2 Superdirective

SDB, also known as supergain beamformer [43] or superdirective array [27, 29], maximizes the directivity factor [29, 42] (see Sec. 2.1) by minimizing the array output power at all directions subject to a distortionless constraint in the target direction. The robust weights are [10]

$$\mathbf{w}_{\mathrm{SDB}}(\omega) = \frac{(\mathbf{\Gamma}_{\mathrm{diff}}(\mathbf{k},\mathbf{r}) + \beta(\omega)\mathbf{I})^{-1}\,\mathbf{a}_{\mathrm{l}}(k,\mathbf{r})}{\mathbf{a}_{\mathrm{l}}^{H}(k,\mathbf{r})\,(\mathbf{\Gamma}_{\mathrm{diff}}(\mathbf{k},\mathbf{r}) + \beta(\omega)\mathbf{I})^{-1}\,\mathbf{a}_{\mathrm{l}}(k,\mathbf{r})}, \tag{9}$$

where $\mathbf{I}$ is the $M\times M$ identity matrix, $\beta(\omega)$ is the regularization parameter controlling the array's sensitivity to sensor self-noise and gain, phase and positioning errors, and $\mathbf{\Gamma}_{\mathrm{diff}}(\mathbf{k},\mathbf{r})$ is the diffuse field coherence matrix:

$$\mathbf{\Gamma}_{\mathrm{diff}}(\mathbf{k},\mathbf{r}) = \frac{1}{L}\mathbf{A}(\mathbf{k},\mathbf{r})\mathbf{A}^{H}(\mathbf{k},\mathbf{r}). \tag{10}$$

1.3.3 MVDR

Unlike DSB and SDB, MVDRB [11] is a data-dependent beamformer that minimizes the array output based on the array input covariance $\mathbf{R}_{\mathrm{xx}}(\mathbf{k},\mathbf{r}) = \mathrm{E}\left[\mathbf{x}(\mathbf{k},\mathbf{r})\mathbf{x}^{H}(\mathbf{k},\mathbf{r})\right]$, subject to the distortionless constraint. The weights [10]

$$\mathbf{w}_{\mathrm{MVDRB}}(\omega) = \frac{(\mathbf{R}_{\mathrm{xx}}(\mathbf{k},\mathbf{r}) + \beta(\omega)\mathbf{I})^{-1}\,\mathbf{a}_{\mathrm{l}}(k,\mathbf{r})}{\mathbf{a}_{\mathrm{l}}^{H}(k,\mathbf{r})\,(\mathbf{R}_{\mathrm{xx}}(\mathbf{k},\mathbf{r}) + \beta(\omega)\mathbf{I})^{-1}\,\mathbf{a}_{\mathrm{l}}(k,\mathbf{r})} \tag{11}$$

resemble those for SDB. In fact, Eq. (11) simplifies to Eq. (8) in a purely diffuse field, i.e., $\mathbf{R}_{\mathrm{xx}}(\mathbf{k},\mathbf{r}) = \mathbf{\Gamma}_{\mathrm{diff}}(\mathbf{k},\mathbf{r})$.

1.3.4 Least-Squares

Since all array manifolds are frequency dependent as shown in Eqs. (4–7), so is the directional response of the above beamformers. Conversely, a particular frequency-independent desired directivity response $\mathbf{d}_{\mathrm{d}} = [d_d(\phi_1),\, d_d(\phi_2),\, \ldots,\, d_d(\phi_S)]$ can be approximated using LSB [23, 44, 45] by minimizing the error with respect to the synthesized response

$$\min_{\mathbf{w}(\omega)} \quad \|\mathbf{w}^{H}(\omega)\mathbf{A}(\mathbf{k},\mathbf{r}) - \mathbf{d}_{\mathrm{d}}\|_2^2 + \beta(\omega)\|\mathbf{w}(\omega)\|_2^2, \tag{12}$$

resulting in the following closed-form solution:

$$\mathbf{w}_{\mathrm{LSB}}(\omega) = \left(\mathbf{A}(\mathbf{k},\mathbf{r})\mathbf{A}^{H}(\mathbf{k},\mathbf{r}) + \beta(\omega)\mathbf{I}\right)^{-1}\mathbf{A}(\mathbf{k},\mathbf{r})\,\mathbf{d}_{\mathrm{d}}^{H}. \tag{13}$$

The target patterns are high-order hypercardioid, which maximize the directivity index for a given order N [14],

$$d_d(\phi_s) = \frac{1}{2N+1}\sum_{n=0}^{N} b_n \cos[n(\phi_s - \phi_{\mathrm{l}})], \tag{14}$$

where ϕ_{l} and ϕ_s are the azimuths at the look and sth steering directions, respectively, and $\mathbf{b} = [b_0, b_1, \ldots, b_N]$ are the real coefficients for natural ($n \geqslant 0$) cylindrical harmonics [13], with $b_0 = 1$ and $b_n = 2\,\forall\, n \neq 0$ [37].

The chosen target directivity patterns are similar to those designed by DMs [14, 15] and similar approaches [13]. Unlike those, LSB is regularized, stabilizing the steering matrix inversion in Eq. (13), thus limiting the array's mismatches in microphone characteristics and self-noise.

2 PHYSICAL EVALUATION

This section evaluates the objective performance of the array geometries in Fig. 1 with physical metrics by means of simulations. These are introduced below.

2.1 Evaluation Metrics

The beampattern $|\mathbf{d}(\omega)|$ is the magnitude of the directional response (Eq. (3)). It fully quantifies the array processing transfer function over steering angle and frequency. Additional metrics that summarize aspects of the beampattern are also considered:

Beam width (BW) is a measure of spatial resolution. It is defined as the angular distance between the two nulls in the beampattern delimiting the mainlobe. The sidelobe suppression level (SSL) is a measure of the minimum acoustic rejection with respect to any single direction outside of the mainlobe. It is defined as the ratio in dB of the directional response at the look direction to that given by the highest sidelobe. Similarly, the acoustic contrast (AC) is a measure of the acoustic rejection at a predefined direction (e.g., interferer direction) with respect to the look direction.

The directivity index (DI) measures the directionality of the array-beamformer as the ratio in dB of the response at the look direction to the average diffuse power [42]:

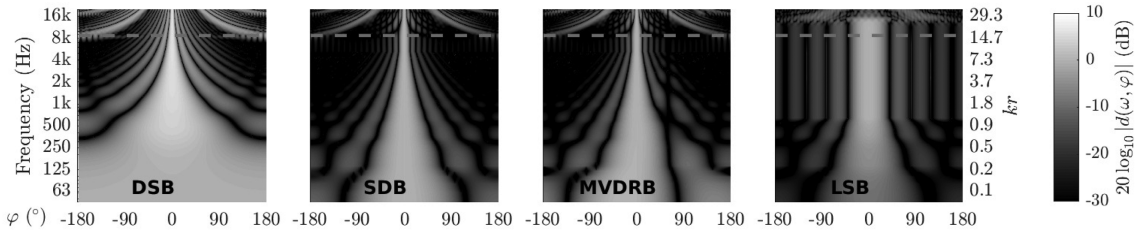
$$\mathrm{DI}(\omega) = 10\log_{10}\left(\frac{|\mathbf{w}^{H}(\omega)\mathbf{a}(k,\phi_{\mathrm{l}},\mathbf{r})|^2}{\mathbf{w}^{H}(\omega)\mathbf{\Gamma}_{\mathrm{diff}}(\mathbf{k},\mathbf{r})\mathbf{w}(\omega)}\right). \tag{15}$$

The white noise gain (WNG) is a measure of robustness of the beamforming weights against microphone self-noise and phase, gain, and positioning deviations from nominal values. It represents the gain in signal-to-noise ratio (SNR) at the beamformer output compared to a single sensor, in the presence of spatially uncorrelated noise [42]:

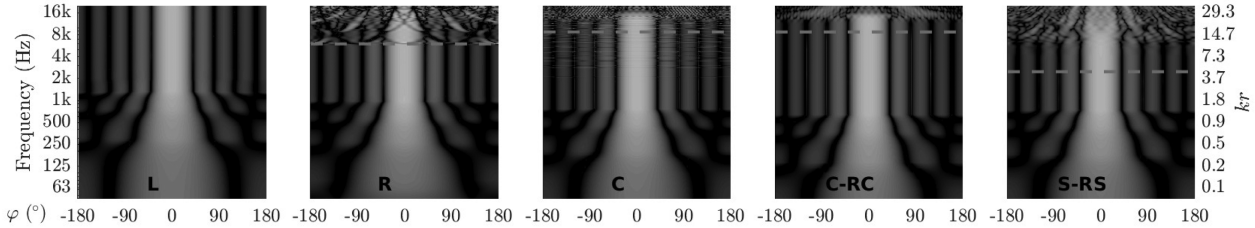
$$\mathrm{WNG}(\omega) = 10\log_{10}\left(\frac{|\mathbf{w}^{H}(\omega)\mathbf{a}(k,\phi_{\mathrm{l}},\mathbf{r})|^2}{\mathbf{w}^{H}(\omega)\mathbf{w}(\omega)}\right). \tag{16}$$

Finally, the frequency range of the array is bounded by the minimum frequency f_{min}, whose BW is smaller than 2π and the spatial aliasing frequency f_{a}, defined here as the frequency at which grating lobes due to aliasing exceed the amplitude of the sidelobes. The frequency-invariant range is set by the onset frequency f_{o} and aliasing frequency and calculated as the range within which the directional response normalized squared error $\mathrm{NSE} \leqslant -20$ dB, ensuring a minimum target response accuracy, where

$$\mathrm{NSE}(\omega) = 10\log_{10}\left(\frac{\sum_{s=1}^{S}|d(\omega,\phi_s) - d_{\mathrm{d}}(\phi_s)|^2}{\sum_{s=1}^{S}|d(\omega,\phi_s)|^2}\right). \tag{17}$$



(a) Circular rigid cylinder array with different beamformers.



(b) 4th-order hypercardioid LSB with different array geometries.

Fig. 2. Beampatterns for different beamformers and arrays from Fig. 1. Dashed lines show theoretical f_a.

2.2 Setup

The performance of the eight array geometries (L, R, C, DC, S, C-RC, C-RS, and S-RS) shown in Fig. 1 is evaluated with the beamformers introduced in Sec. 1.3 and the metrics from Sec. 2.1 over a horizontal sound field.

All beamformer weights were calculated for a look direction $\varphi_l = 0°$ (where $\varphi = 90 - \phi$), subject to a WNG constraint (WNG_{min}) of -10 dB unless otherwise stated, to limit the sensitivity to mismatches between nominal and actual array manifold responses encountered in practice. Thus, $\beta(\omega)$ is derived to meet WNG_{min}. L was pointed endfire to $\varphi_l = 0°$. MVDRB was computed as a data-independent beamformer, assuming a diffuse field with an interferer at $\varphi_i = 60°$.

2.3 Results

This section presents the results of the performance of the arrays under study evaluated in terms of the beampattern, frequency range, beamwidth, directivity, robustness, and sidelobe suppression.

2.3.1 Beampattern

The beampattern characterizes the effect of beamformer and array design choices for an arbitrary sound field. Fig. 2(a) shows the beampattern for DSB, SDB, MVDRB, and fourth-order hypercardioid LSB (shortened as LSB henceforth) with the C-RC. The shape of the beampattern changes significantly for these beamformers: DSB is the most frequency-dependent beamformer with omnidirectional response below 300 Hz, narrowing rapidly with frequency; SDB is the most directive, with gradual beam narrowing and larger attenuated region as frequency increases; MVDRB's response approaches that of SDB, with greater attenuation at the interferer; and LSB provides a fixed beampattern within the array design's operating bandwidth, while at low frequencies it becomes broader and attenuated due to the regularization to meet WNG_{min}.

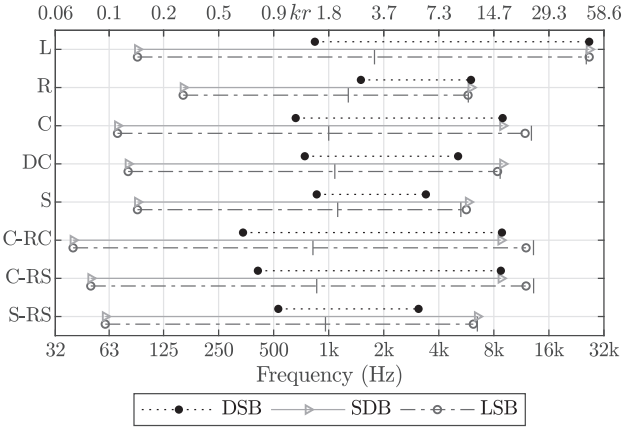


Fig. 3. Bandwidth of all arrays with DSB, SDB, LSB, and WNG_{min} = 0 dB. Frequency-independent range for LSB indicated with vertical lines.

On the other hand, the overall shape of the beampattern is more similar across different arrays with the same beamformer. An example is shown in Fig. 2(b) for LSB with different array geometries. However, the array design has significant effects in terms of frequency range, resolution, directivity, robustness, and sidelobe suppression. These are analyzed in more detailed below.

2.3.2 Frequency Range

The main effect of the array geometry is the operating frequency range. This can be seen in Fig. 2(b), where the onset and aliasing frequencies differ significantly across arrays. The operating frequency ranges of all arrays with DSB, SDB, and LSB are shown in Fig. 3. For a fixed number of sensors, the more dimensions the array spans, the smaller the operating bandwidth. In this case, with $M = 32$ and fixed maximum aperture of 0.2 m ($r = 0.1$ m), different spacing leads to different f_a, ranging with DSB from 3 kHz for S

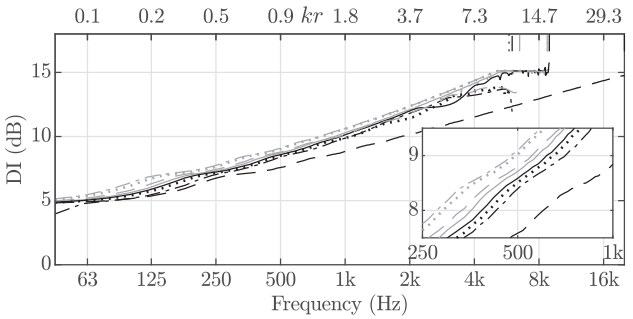


Fig. 4. DI of all arrays with SDB below f_a (marked at the top of each graph). Legend as per Fig. 5.

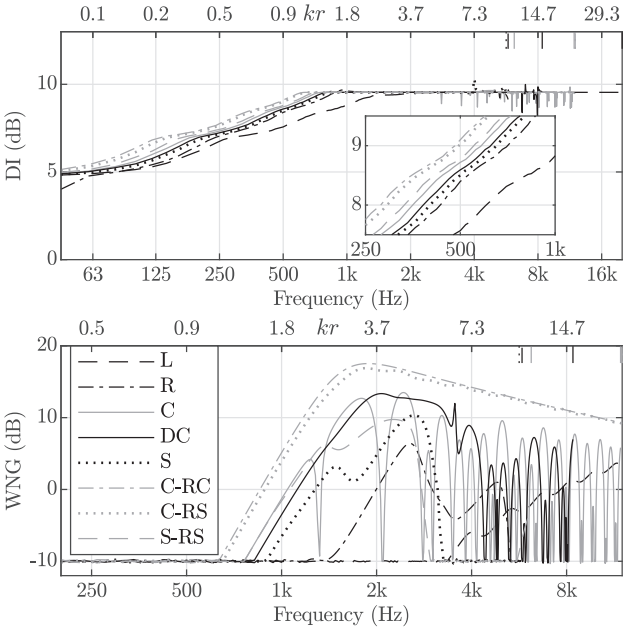


Fig. 5. DI and WNG of all arrays with LSB below f_a (marked at the top of each graph); note different frequency axes.

to over 27 kHz for L, with circular arrangements achieving the second highest value of 8.9 kHz.

On the other hand, the minimum frequency is rarely reported. Despite physically constraining the maximum aperture to a fixed size for all arrays, f_{min} varies due to their sensor phase differences. The concept of effective or virtual modal aperture was previously used to describe the effect of a baffle on circular arrays' modal response [37]. Here, the effective aperture is referred as the equivalent wave traveling distance from the microphone phase responses, which is used to show the effect of array geometry on the acoustic response (i.e., before applying the beamformer) and to explain the values of f_{min} which are the result of the acoustical and signal processing stages (see [46]). Results show R has the highest f_{min} and smallest effective aperture, due to its sensors' proximity to the origin. L follows, whose f_{min} is 25% higher than that of the highly separated circular arrangement C. With diffraction around a baffle, larger phase differences arise with the same array aperture, hence baffled arrays have larger effective apertures resulting in lower f_{min}. For C-RC and C-RS, f_{min} reduces with respect to C by factors of 2.0 and 1.5 respectively, in line with those from the effective apertures between the closest and farthest microphones from a plane wave incidence for $kr < 1$ [46]. The rigid-sphere factor of 1.5 is also derived in [40, 47]. Note that the ranking of these arrays in terms of f_{min} and f_a is consistent for the three beamformers, showing that these physical characteristics of the arrays impact on their operating bandwidth for multiple beamformers.

The beamformer, on the other hand, can further extend the arrays' operating range. SDB and LSB lower significantly f_{min} compared to DSB for all arrays. This shows that the improved low frequency performance shown in Fig. 2(a) has the equivalent effect of extending the minimum frequency of directionality. Some configurations also extend f_a beyond the theoretical values ($c/(2\Delta d)$) and those obtained numerically by DSB: DC, S and S-RS with SDB, and all arrays except L and R for LSB (Fig. 3). In addition, for circular arrangements with LSB f_a extends even beyond that of SDB, e.g., baffled circular arrays extend their upper limit from 8.5 kHz to 12 kHz. This is also seen in Fig. 2(a) with the red dashed lines indicating the theoretical f_a. Unlike DSB and SDB, LSB's first aliased lobes occur nearly at the same frequency at all angles, thus extending its upper limit when synthesizing low order patterns. This is in line with the findings in [23] for various hypercardioid orders.

Finally, the frequency-invariant range for the LSB is also shown in Fig. 3 with vertical lines. Its onset frequency f_o is higher than f_{min} for all arrays, yet with nearly identical ranking of arrays, with C-RC and L achieving the lowest and highest f_o, respectively. Observe that in Fig. 3, $\mathrm{WNG}_{min} = 0$ dB so the differences in f_{min} among arrays become apparent in the frequency range of interest.

Note that for a different constraint on r, the beampattern will scale inversely proportionally with frequency due to the dependence of Eqs. (4–7) with kr. Thus, the frequency ranges shown in Fig. 3 would be fixed in terms of kr while shifting with respect to the frequency axis. However for $r \gg 0.1$ m, the arrays will no longer meet the compactness requirement for a practical portable recording device. On the other hand, increasing M maintains the same ranking and extends the theoretical aliasing frequencies shown in Fig. 3 [46], where $f_a = c(M-1)/(4r)$ for L, $f_a \approx cM/(4\pi r)$ for circular arrays and $f_a \approx c\sqrt{M}/(8r)$ for spherical arrays.

Summarizing, the array geometry has a huge impact on the frequency range of the array-beamformer response which is very important in object capture. Baffled circular arrays (C-RC and C-RS) achieve the widest perceptually relevant bandwidth for all beamformers under study, with R and S having the narrowest ranges.

2.3.3 Resolution and Directivity

Beam resolution and directivity are important to improve the isolation from adjacent sources and, in addition to the beamforming method, depend on the array design. Spatial resolution is inversely proportional to frequency, and array size [48, 20]. Given a maximum aperture limit, we show how the effective aperture of the array also determines the

resolution and directivity. Due to the inverse relationship of resolution and frequency, at low frequencies, BW follows the same ranking as f_{min} (Fig. 3), in turn determining DI. The latter is shown in Fig. 4 for SDB. Baffled circular arrays perform best (66° and 10.6 dB at 1 kHz), followed by S-RS (74°, 10.1 dB). Among open arrays, circular arrangements (72°, 10.3 dB) are superior to S and R (78°, 9.9 dB), with L performing the worst ($\geqslant$98°, 8.8 dB).

Hence, while L can theoretically achieve the highest directivity ($\text{DI}_{\text{max}} = 10\log_{10}(2M - 1)$) [14, 27], this is only for unconstrained SDB or DMs, which are extremely sensitive to deviations from ideal microphone characteristics [43]. For robust beamforming required for practical recordings, baffled circular arrays achieve the highest directivity (and resolution) due to its increased effective aperture, being up to 3 dB higher than that for L with SDB. Note that increasing M will increase the maximum directivity which for SDB is $\text{DI}_{\text{max}} = 10\log_{10}(2N_{\text{max}} + 1)$, where $N_{\text{max}} = M - 1$ for L and $N_{\text{max}} = \lfloor M/2 \rfloor$ for circular arrangements [37, 49]. However, this may only be achieved at high frequencies (or not at all) given the robustness constraint, so the ranking of array performance remains unaltered for other practical choices of M.

Finally, the same ranking and DI are seen with LSB at low frequencies in Fig. 5 (top). While SDB can be regarded as an N_{max} th-order hypercardioid, given WNG_{min}, both regularized beamformers synthesize the same directivity below f_{o}, thus exhibiting the same array differences.

2.3.4 Robustness

Practical recordings with microphone arrays require the actual array response to be robust to typical deviations in microphone positioning, gain and phase and to sensor noise. While a minimum robustness constraint on the weights limits the sensitivity to these deviations at low frequencies, the array geometry impacts on the absolute robustness at mid-to-high frequencies as shown in Fig. 5 (bottom) with LSB. Baffled circular arrays feature the highest WNG whereas L achieves the lowest. WNG for C and DC shows a significant number of dips at particular frequencies. These correspond to Bessel function singular frequencies (Eq. (5)), becoming ill-conditioned when inverted. While this has been widely reported in MB/HOA [17, 32, 34], here it is shown that it also applies to FSBs relying on the array manifold inversion, including LSB, SDB, and MVDRB, thus being inherent to the open circular arrangement. Due to the robustness constraint, the WNG dips are limited to −10 dB. This constraint causes the directional response of these arrays to differ from the ideal response at those frequencies (even in ideal conditions). These manifest as dips in the response, e.g., DI for LSB in Fig. 5 (top). Unlike C, DC overcomes the singularities below 5 kHz, since it samples the sound field at different radial positions, thus avoiding the singularities to occur at the same frequencies. At high frequencies the number of modes is so large that the singularities overlap for different radii. Hence, careful choice of array radii is crucial as shown in [32].

WNG_{min} can be modified depending on the expected deviations from nominal microphone characteristics. A very low WNG_{min} will lead to significant performance degradation due to minor deviations in microphone characteristics whereas a very high WNG_{min} would result in a response close to that of DSB [23], thus exhibiting similar relative differences across arrays to those shown here in terms of frequency range, BW, DI and WNG.

2.3.5 Sidelobe Suppression

The SSL varies significantly with array geometry for DSB. A constant SSL of 13 dB is achieved by R, S and L, being only 7 dB for C. Baffled arrays have a SSL with larger attenuation in the lower range, with S-RS having the highest SSL yet over a narrow range. Conversely, SSL for SDB, MVDRB, and LSB is insensitive to the choice of array, being around 14 dB for all arrays with SDB. Thus, the effect of array geometry on SSL is not significant for beamformers with amplitude weights.

2.3.6 Summary

The array geometry has a significant impact on frequency range, resolution, directivity, and robustness, with baffled circular arrays performing best in all these attributes, which are important in object capture. R and S result in the narrowest bandwidths. L achieves the highest f_{a}, yet with the highest f_{o}, and performs the worst in resolution and directivity. Finally, open circular arrangements are less robust than their baffled counterparts.

3 PERCEPTUAL EVALUATION

This section perceptually evaluates the performance of different array designs and beamformers in terms of sound quality and interference suppression of the isolated audio object from a scene recording.

3.1 Procedure

Two listening tests comparisons were conducted: arrays and beamformers. In the array comparison, a fourth-order hypercardioid LSB-N4 was synthesized with L, R, C-RC and S-RS, since their frequency ranges vary significantly both in terms of f_{o} and f_{a} as shown in Sec. 2.3.2. The beamformer comparison used C-RC, since this was shown to perform best overall in Sec. 2.3, and included DSB, SDB, and LSB for orders 1, 4 and 8, providing different levels of on-axis and off-axis responses.

For each comparison, three different attributes were evaluated: 1) *target quality* refers to the quality of the target sound with respect to the reference; 2) *interference suppression* refers to any and all effects of interfering sources in each stimulus compared to the reference; and 3) *overall quality* refers to the combined score considering the *target quality* 1) and *interference suppression* 2).

Each comparative test was undertaken with two target sounds (vocals and drums), which were repeated to check intra-participant agreement, resulting in 4 trials per test. In each trial participants were asked to rate the stimuli (beam-

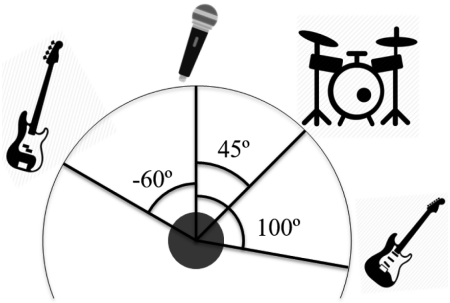


Fig. 6. Simulated music performance recording for the MUSHRA listening test. All arrays within central circle.

formed signals and hidden reference and anchors) with respect to the reference according to each of the tasks above, using a MUSHRA-style interface [50]. To familiarize with the stimuli and the interface, subjects undertook a training phase prior to the formal evaluation [50], in which they could adjust the volume of the headphones.

3.2 Stimuli

Stimuli were obtained from the Mixing Secret Dataset [51] which includes stems from professionally produced music recordings [1]. Vocals, drums, bass and guitar tracks were collected from the song "A reason to leave". Ten-second clips from these tracks were downmixed to mono and loudness normalized [52], to provide a fair comparison for different instruments. The reference signal was either the vocals or drums track for all trials. The interference task included one hidden anchor corresponding to the loudness-normalized mono mixture (Mix) of the four stems. The target quality and overall quality tasks included two hidden anchors. The low- and mid-quality anchors for the target quality task (LA and MA) were the low-pass filtered versions of the reference signal with a cut-off frequency of 3.5 kHz and 7 kHz, respectively [50]. In the overall task, equivalent low- and mid-quality anchors from the mixture were used (LAMix and MAMix).

3.3 Setup

The remaining stimuli were created from simulated microphone array beamformed signals. A sound scene comprising musical instruments was simulated (Fig. 6) by positioning them on the horizontal plane at angles that resemble a practical setup from a music performance or band practice: vocals at 0°, bass at -60°, drums at 45° and guitar at 100°. The microphone arrays were assumed to be in the center of the scene (with L pointing at 0°) and were steered towards the vocals or drums. Array transfer functions were modeled with a 1024-point FIR filter per sensor with a sampling frequency of 44.1 kHz. Microphone array and beamformed signals were calculated by filtering the stimuli as per Eqs. (1) and (2), respectively.

[1] www.cambridge-mt.com/ms-mtk.htm

3.4 Pre-Analysis

24 participants from the University of Surrey conducted the experiment, 11 of whom had formal critical listening training. Among all of them, 19 were considered in the analysis: 4 failed to rate the reference above 90 for over 85% of the items [50]; and 1 rated the interference task in terms of quality, which was confirmed by a post-test questionnaire and by the mixture (anchor) ratings above 70 for the beamformer test. Each participant's scores were normalized in each trial [50].

A repeated measures analysis of variance (RMANOVA) was performed for each attribute (target quality, interference suppression and overall quality) and comparison (beamformers and arrays) to obtain a statistical analysis of the results [50]. The multivariate normality of the residuals (differences between systems) was tested using the Henze-Zirkler's method, which failed to reject the null hypothesis (normal) for all tests, except for the overall quality task with vocals. The within factors of the two-way RMANOVAs were *system* (i.e., array or beamformer excluding reference and anchors) and *instrument*. The results from repeated tests were averaged before the analysis, as *repeat* was not a significant factor when included.

The results of the RMANOVAs for all tests are shown in Table 1 in terms of the F-statistic with significant factors in bold ($p < 0.05$). All tests showed significant deviation from sphericity using Mauchly's test and the Huynh-Feldt correction was applied [50]. RMANOVAs show significant differences within *beamformers* and *arrays* for all attributes. Moreover, the three levels are significant at least for one attribute in each comparison. Thus, post-hoc comparisons are performed to investigate the differences between the scores of SDB and the other beamformers and between the scores of C-RC and the other arrays as both performed best overall, and since comparisons of all conditions are discouraged [50]. Hochberg's sequentially acceptive step-up Bonferroni procedure was applied to control Type I error [50]. These *t*-test comparisons are described in the following and tabulated in [46].

3.5 Results

The listening test results in terms of means and 95% confidence intervals (CIs) for the two comparisons and three tasks are shown in Fig. 7, and analyzed below.

3.5.1 Beamformer Comparison

The scores of the different beamformers for the target quality task are shown in Fig. 7(a). For drums, SDB achieves the highest target score of 88, being significantly higher than those for all other methods. This is because SDB achieves a flat response compared to DSB's high frequency boost from the baffle scattering and LSB's inherent high-pass filter from regularization, as shown in Fig. 8 (left). In fact, LSB's mean scores drop from 70 to 42 when increasing the order from 1 to 8, as a result of the higher low-frequency roll-off at the look direction. On the other hand, for vocals similar target quality scores are seen with DSB, SDB, LSB-N1, and LSB-N4 with the latter having the highest yet not significant

Table 1. F-statistics of RMANOVA for each MUSHRA test. Significant factors in bold (Huynh-Feldt -corrected $p < 0.05$).

Attribute	Beamformers	Instruments	Beamformers-Instruments
Target Quality	$\boldsymbol{F(2.75, 49.59) = 15.21}$	$\boldsymbol{F(0.69, 12.40) = 27.11}$	$\boldsymbol{F(2.75, 49.59) = 10.32}$
Interference	$\boldsymbol{F(3.28, 59.11) = 102.39}$	$F(0.82, 14.78) = 0.07$	$F(3.28, 59.11) = 1.87$
Overall Quality	$\boldsymbol{F(1.56, 28.03) = 16.14}$	$F(0.39, 7.01) = 2.34$	$\boldsymbol{F(1.56, 28.03) = 6.95}$
Attribute	**Arrays**	**Instruments**	**Arrays-Instruments**
Target Quality	$\boldsymbol{F(1.82, 32.72) = 25.48}$	$\boldsymbol{F(0.61, 10.91) = 31.68}$	$\boldsymbol{F(1.82, 32.72) = 5.88}$
Interference	$\boldsymbol{F(1.98, 35.60) = 38.10}$	$F(0.66, 11.87) = 0.07$	$\boldsymbol{F(1.98, 35.60) = 29.75}$
Overall Quality	$\boldsymbol{F(2.02, 36.33) = 16.81}$	$\boldsymbol{F(0.67, 12.11) = 7.17}$	$\boldsymbol{F(2.02, 36.33) = 3.80}$

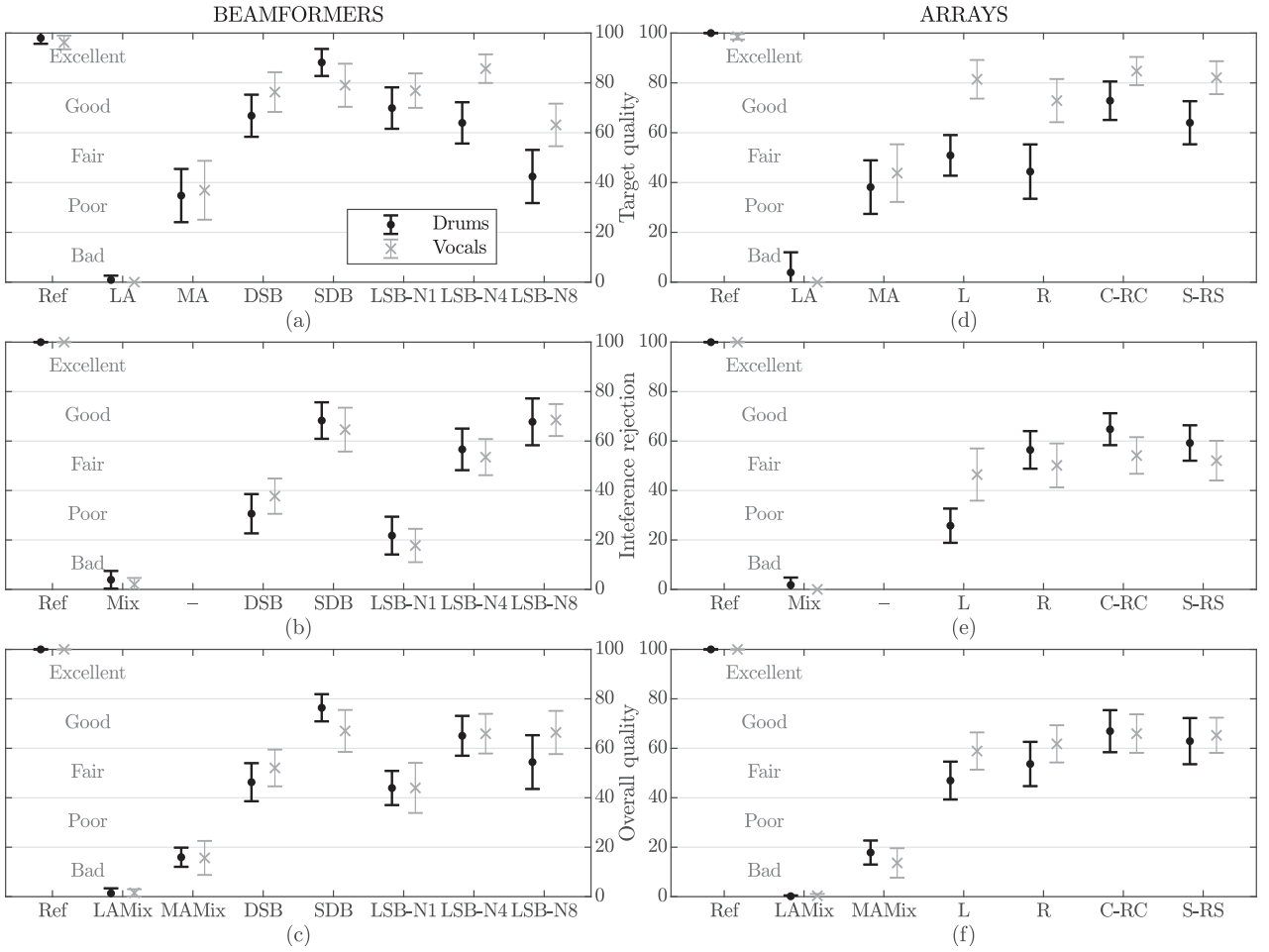


Fig. 7. MUSHRA listening test means and 95% confidence intervals for different beamformers (left) and arrays (right): target quality (top), interference rejection (middle) and overall quality (bottom).

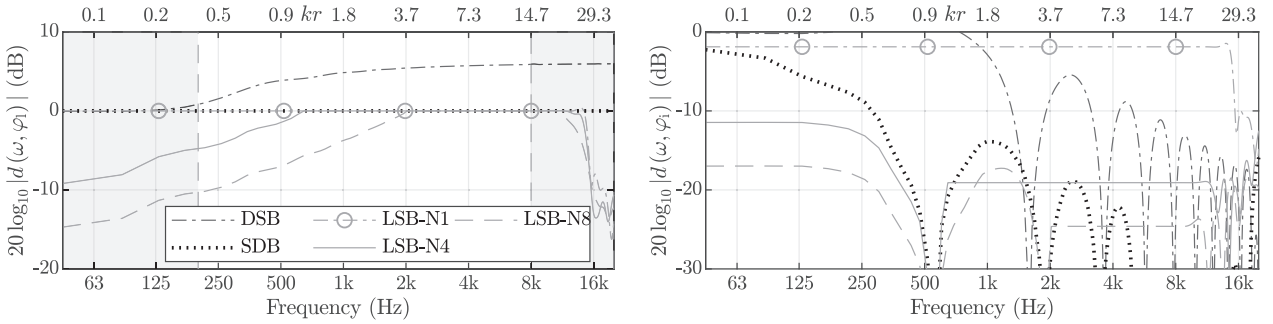


Fig. 8. Magnitude response of all beamformers at drums as target (left) and vocals as interferer (right), as per Fig. 6, with C-RC steered at drums. Shaded area corresponds to region outside of vocal frequency range for comparison.

mean of 85. This more similar subjective performance is probably due to the reduced frequency range of the vocals (shaded area in Fig. 8 (left)), where the response difference among these beamformers is deemphasized.

The results in terms of the interference rejection are shown in Fig. 7(b). LSB-N8 and SDB perform best with statistically higher scores than for all other methods. This is because they achieve the highest attenuation at the interfering instruments as shown in Fig. 8 (right). Despite SDB's more frequency-dependent response, both beamformers achieve similar subjective interference suppression scores. On the other hand, LSB-N1 achieves the lowest scores due to its flat 1.9 dB attenuation, followed by DSB, given its omnidirectional and comb-filtering responses at low and high frequencies, respectively.

Fig. 7(c) shows the overall quality scores. For drums SDB achieves significantly higher scores than all other methods, confirming its higher combined performance from each of the previous tasks. LSB-N8 is significantly worse than LSB-N4, indicating that the target quality degradation seen in Fig. 7(a) becomes important in the overall score too. For vocals, SDB, LSB-N4 and LSB-N8 obtain very similar values with means 66–67, suggesting that the reduced vocal range flattens the differences across beamformers, as seen for the target quality.

3.5.2 Array Comparison

The array comparison is shown in Fig. 7(d-f). C-RC achieves the highest scores for all attributes and instruments, yet not necessarily significant in all cases. For the target quality (Fig. 7(d)), C-RC is significantly higher than all other arrays for drums. For vocals C-RC is only significantly higher than R, since the differences in array responses reduce within the narrower vocal range.

In terms of the interference rejection (Fig. 7(e)), C-RC is significantly better than the other three arrays for drums. The scores for the linear array are exceptionally low with a mean of 26 as a result of its reduced performance when steered to a direction other than endfire, resulting in a mirrored mainlobe with respect to the endfire direction (i.e., –45° in this case). This results in very poor attenuation of the bass guitar located at –60°. For vocals the four arrays perform similarly, including L since the vocals are located at the endfire direction.

The overall score (Fig. 7(f)) for C-RC is significantly higher than those for L and R but not S-RS with drums, and only significantly higher than that for L for vocals.

3.6 3AFC Test

The MUSHRA test revealed higher mean scores by C-RC for all tests. However, some of these could not be shown to be statistically significant with the vocals excerpt for both target quality and interference. In order to show whether C-RC consistently achieves higher scores than R and S-RS, a 3-alternative forced choice (3AFC) test was designed. L was discarded due to its notable performance drop when steered at off-axis directions, which is essential in a multisource array beamforming capture.

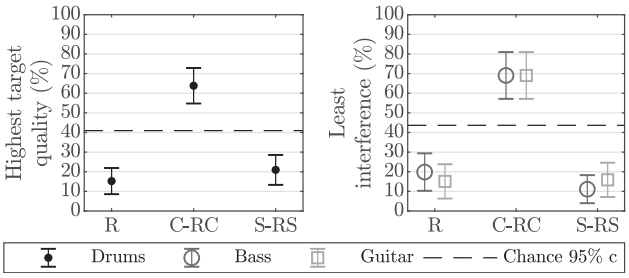


Fig. 9. 3AFC listening test percentage scores for quality of target sound (left) and interference rejection (right) tasks. Means and 95% CIs for each array and 95% critical value c for chance voting.

The 3AFC test consisted of a clean reference and three stimuli corresponding to the beamformed signals from R, C-RC and S-RS, with LSB-N4. The two tasks were to select a single stimulus that resulted in 1) *highest quality* and 2) *least interference* with respect to the reference. Since the performance of these different arrays with frequency-invariant LSB beampatterns is mainly related to the onset and aliasing frequencies, wideband signals are required. Thus, the quality of the target sound was evaluated for drums. For the interference task the drums acted as one of the interfering instruments, with the target instrument being bass or guitar. To generalize the results to multiple setups, different combinations of the angles in Fig. 6 were considered for all instruments: 5 for the quality task and 3 for each instrument in the interference task. To account for intraparticipant agreement, each trial was repeated three times, resulting in 15 and 18 trials for the quality and interference tasks, respectively.

Fourteen participants with formal critical listening conducted the experiment. All were selected for the analysis since their mean normalized mode frequency was above 2/3 (1/3 implies random scoring and 1 corresponds to fully correlated scores). Fig. 9 shows the percentage of votes for each array for the quality and interference tasks. C-RC clearly outperforms the other two arrays in both tasks with 64% and 69% of votes. To determine whether this result is statistically significant, binomial distributions of the probability of selecting any array by chance ($p_0 = 1/3$) were implemented with $t = 15 \times 14 = 210$ and $t = 18 \times 14 = 252$ trials for both tasks. The critical value c of this binomial chance probability is calculated from the cumulative distribution $F = \sum_{k=0}^{c} {}_t^i p_0^k (1 - p_0)^{t-k} \geq 1 - \alpha$ [53], where $\alpha = 0.05$ is the significance level. Since the percentage of votes from C-RC exceeds these critical values for both tasks as shown in Fig. 9, C-RC's higher quality and interference rejection is statistically significant. Moreover, since the 95% CI of the votes from C-RC does not overlap with the chance critical region, these results can be said to extrapolate to a larger population. Hence, the 3AFC test shows that C-RC achieves statistically significantly higher quality and interference rejection than R and S-RS.

4 DISCUSSION

The results from the physical and perceptual evaluations have shown evidence of the higher performance of the baffled circular arrays over the alternative array geometries considered. These are discussed in the context of desired properties of captured objects, some of which may also extrapolate to other beamforming applications.

One of the most important requirements for multisource 2D capture is to synthesize a beampattern that is independent of the steering azimuth. This is achieved by all arrays considered here except for L, whose mirrored response when steered off the endfire direction showed a significant drop in perceptual interference attenuation, making it inadequate for this application.

Another very important aspect in object capture is frequency range, since audio objects may include wideband signals such as music. Baffled circular arrays achieve the widest perceptually relevant bandwidth with the lowest onset frequency and the second highest aliasing frequency. This explains C-RC's statistically highest quality scores with drums when synthesizing a fourth-order hypercardioid pattern in both MUSHRA and 3AFC tests. On the other hand, R achieves the narrowest bandwidth, L has the highest onset frequency and S-RS has a similar onset frequency than C-RC yet with a lower aliasing frequency. The lowest quality scores achieved by R, followed by L, suggest that they are penalized by their bass drop, whereas S-RS performs better than R and L but worse than C-RC, probably due to its lower upper limit.

On the other hand, for vocals, which are not as wideband, the results for target quality and interference across arrays become much more similar. However, the 3AFC test shows the significantly higher interference suppression of C-RC evaluated with bass and guitar as target instruments, and over different relative instrument positions. This indicates that even though the differences in target quality and interference across arrays are not fully exploited with band-limited target signals, the extended frequency range of the baffled circular array may become important to attenuate low frequencies and/or aliasing effects that may be audible from the other arrays in presence of interfering wideband signals like drums.

Since the performance of the captured object also depends on the beamformer, a perceptual evaluation of different beamformers was conducted. The quality of the target sound is one of the most important aspects of object capture, with SDB achieving *excellent* quality, due to its distortionless constraint, compared to DSB's *good* quality as a result of its high-frequency boost from C-RC's baffle scattering. LSB's quality degrades as the order increases due to the higher low-frequency roll-off from its regularized response. However, this could be compensated through equalization at the look direction.

The ability to suppress other sources is important for object capture, where SDB and LSB-N8 perform best with *good* attenuation. This indicates that the overall level difference is mainly considered, compared to LSB-N4's lower yet more frequency consistent AC. However, the equivalent interference scores from SDB (which can be regarded as $N = 16$) and LSB-N8 suggests that increasing the order beyond $N = 8$ with a robustness constraint may not lead to greater perceptual attenuation.

The overall quality is highest for SDB with drums, followed by LSB-N4 and LSB-N8, showing that the low frequency roll-off from high-ordered LSB becomes detrimental in the overall quality too. On the other hand, the same overall performance is seen for these three beamformers with vocals, suggesting that LSB's high-pass filter is not as important for such band-limited signals.

For future work, target patterns other than hypercardioid, and other beamformers, may be explored to maximize the signal-to-interferer ratio for isolating the target object. The perceptual properties of these arrays may also be investigated for capturing performances in reverberant conditions.

5 CONCLUSION

This study evaluated the performance of uniform microphone array designs with the same number of microphones and maximum array size for object capture with beamforming in 2D. Simulation results show that baffled circular arrays performed best in terms of physical measures, including resolution, directivity, robustness and perceptually relevant frequency range, compared to alternative geometries. Listening tests were conducted to perceptually evaluate the performance of arrays and beamformers on simulated music performance recordings. The cylindrical array showed higher overall quality than linear, rectangular and baffled spherical arrays for a fourth-order hypercardioid LSB, yet not always significantly, especially for vocals. However, the cylindrical array showed statistically significantly higher quality of target sound and interference suppression than all other arrays in the presence of wideband signals, being confirmed by the 3AFC test. Hence, these conclusions quantitatively motivate the use of baffled circular arrays for practical horizontal source separation capture. In terms of beamformers, perceptual scores for target quality and interference suppression agreed well with beamformer on-axis and off-axis responses, respectively, with SDB achieving higher overall quality than LSB for wideband signals, potentially due to LSB's regularized high-pass response.

6 ACKNOWLEDGMENT

This work was supported by the EPSRC Programme Grant S3A: Future Spatial Audio for an Immersive Listener Experience at Home (EP/L000539/1) and the BBC as part of the BBC Audio Research Partnership.

7 REFERENCES

[1] P. Coleman, A. Franck, J. Francombe, Q. Liu, T. De Campos, R. J. Hughes, D. Menzies, M. F. Simon Galvez, Y. Tang, J. Woodcock, P. J. Jackson, F. Melchior, C. Pike, F. M. Fazi, T. J. Cox, and A. Hilton, "An Audio-Visual System for Object-Based Audio: From Recording to Listening," *IEEE Transactions on Multimedia*, vol. 20, no.

8, pp. 1919–1931 (2018 Aug.), https://doi.org/10.1109/TMM.2018.2794780.

[2] P. Coleman, A. Franck, P. J. Jackson, L. Remaggi, and F. Melchior, "Object-Based Reverberation for Spatial Audio," *J. Audio Eng. Soc.*, vol. 65, no. 1/2 (2017 Jan.), https://doi.org/10.17743/jaes.2016.0059.

[3] J. Paulus, M. Torcoli, C. Uhle, J. Herre, S. Disch, and H. Fuchs, "Source Separation for Enabling Dialogue Enhancement in Object-Based Broadcast With MPEG-H," *J. Audio Eng. Soc.*, vol. 67, no. 7-8, pp. 510–521 (2019 Jul.), https://doi.org/10.17743/jaes.2019.0032.

[4] A. Franck, J. Francombe, J. Woodcock, R. Hughes, P. Coleman, D. Menzies, T. J. Cox, P. J. Jackson, and F. M. Fazi, "A System Architecture for Semantically Informed Rendering of Object-Based Audio," *J. Audio Eng. Soc.*, vol. 67, no. 7-8, pp. 498–509 (2019 Jul.), https://doi.org/10.17743/jaes.2019.0025.

[5] L. A. Ward and B. G. Shirley, "Personalization in Object-Based Audio for Accessibility: A Review of Advancements for Hearing Impaired Listeners," *J. Audio Eng. Soc.*, vol. 67, no. 7-8, pp. 584–597 (2019 Jul.), https://doi.org/10.17743/jaes.2019.0021.

[6] F. Rumsey and T. McCormick, *Sound and Recording*, 6th ed. (Focal Press, New York, New York, 2009), https://doi.org/10.4324/9780080953960.

[7] P. Coleman, P. J. B. Jackson, and J. Francombe, "Audio Object Separation Using Microphone Array Beamforming," presented at the *138th Convention of the Audio Engineering Society* (2015 May), convention paper 9296.

[8] Y. Huang, J. Chen, and J. Benesty, "Immersive Audio Schemes," *IEEE Signal Processing Magazine*, vol. 28, no. 1, pp. 20–32 (2011 Jan.), https://doi.org/10.1109/MSP.2010.938754.

[9] B. Van Veen and K. Buckley, "Beamforming: A Versatile Approach to Spatial Filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24 (1988 Apr.), https://doi.org/10.1109/53.665.

[10] H. Cox, R. M. Zeskind, and M. M. Owen, "Robust Adaptive Beamforming," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 10, pp. 1365–1376 (1987 Oct.), https://doi.org/10.1109/TASSP.1987.1165054.

[11] J. Capon, "High-Resolution Frequency-Wavenumber Spectrum Analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418 (1969 Aug.), https://doi.org/10.1109/PROC.1969.7278.

[12] C. Pan, J. Chen, and J. Benesty, "Reduced-Order Robust Superdirective Beamforming With Uniform Linear Microphone Arrays," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 24, no. 9, pp. 1544–1555 (2016 Sep.), https://doi.org/10.1109/TASLP.2016.2568044.

[13] G. Huang, J. Benesty, and J. Chen, "On the Design of Frequency-Invariant Beampatterns With Uniform Circular Microphone Arrays," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 25, no. 5, pp. 1140–1153 (2017 May), https://doi.org/10.1109/TASLP.2017.2689681.

[14] G. W. Elko, "Differential Microphone Arrays," in Y. Huang and J. Benesty (Eds.), *Audio Signal Processing for Next-Generation Multimedia Communication Systems*, pp. 11–65 (Springer, Boston, Massachusetts, 2004), https://doi.org/10.1007/1-4020-7769-6_2.

[15] J. Benesty and J. Chen, *Study and Design of Differential Microphone Arrays* (Springer, Berlin, Germany, 2013), https://doi.org/10.1007/978-3-642-33753-6.

[16] J. Meyer and G. Elko, "A Highly Scalable Spherical Microphone Array Based on an Orthonormal Decomposition of the Soundfield," presented at the *2002 International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2002), https://doi.org/10.1109/ICASSP.2002.5744968.

[17] B. Rafaely, "Analysis and Design of Spherical Microphone Arrays," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 135–143 (2005 Jan.), https://doi.org/10.1109/TSA.2004.839244.

[18] F. Kaiser, H. Pomberger, and F. Zotter, "Investigations on Cylindrical Microphone Arrays," presented at the *Audio Engineering Society Conference: 25th UK Conference: Spatial Audio in Today's 3D World* (2012 Mar.), conference paper 02.

[19] B. Rafaely, "Plane Wave Decomposition of the Sound Field on a Sphere by Spherical Convolution," *The Journal of the Acoustical Society of America*, vol. 116, no. 4, pp. 2149–2157 (2004 Oct.), https://doi.org/10.1121/1.1792643.

[20] J. Christensen and J. Hald, "Beamforming," Tech. Rep. 1 (Brüel & Kjær, Nærum, Denmark, 2004).

[21] L. Pfeifenberger and F. Pernkopf, "Blind Source Extraction Based on a Direction-Dependent A-Priori SNR," presented at the *Fifteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 2700–2704 (2014 Sep.).

[22] L. Pfeifenberger and F. Pernkopf, "A Multi-Channel Postfilter Based on the Diffuse Noise Sound Field," presented at the *22nd European Signal Processing Conference (EUSIPCO)*, pp. 686–690 (2014 Sep.).

[23] M. Blanco Galindo, P. Coleman, and P. J. B. Jackson, "Robust Hypercardioid Synthesis for Spatial Audio Capture: Microphone Geometry, Directivity and Regularization," presented at the *2019 AES International Conference on Immersive and Interactive Audio* (2019 Mar.), conference paper 49.

[24] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and Objective Quality Assessment of Audio Source Separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 2046–2057 (2011 Sep.), https://doi.org/10.1109/TASL.2011.2109381.

[25] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A Robust Adaptive Beamformer for Microphone Arrays With a Blocking Matrix Using Constrained Adaptive Filters," *IEEE Transactions on Signal Processing*, vol. 47, no. 10, pp. 2677–2684 (1999 Oct.), https://doi.org/10.1109/78.790650.

[26] A. Farina, S. Campanini, L. Chiesi, A. Amendola, and L. Ebri, "Spatial Sound Recording With Dense Mi-

crophone Arrays," presented at the *AES 58th International Conference: 55th International Conference: Spatial Audio* (2014 Aug.), conference paper P-10.

[27] J. Bitzer and K. U. Simmer, "Superdirective Microphone Arrays," in *Microphone arrays: Signal Processing Techniques and Applications*, pp. 19–38 (Springer, Berlin, Germany, 2001), https://doi.org/10.1007/978-3-662-04619-7_2.

[28] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing* (Springer, Berlin, Germany, 2008), https://doi.org/10.1007/978-3-540-78612-2.

[29] M. R. Bai, J.-G. Ih, and J. Benesty, *Acoustic Array Systems* (John Wiley & Sons, Singapore, 2013), https://doi.org/10.1002/9780470827253.

[30] I. Balmages and B. Rafaely, "Open-Sphere Designs for Spherical Microphone Arrays," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 2, pp. 727–732 (2007 Feb.), https://doi.org/10.1109/TASL.2006.881671.

[31] B. Rafaely, "The Spherical-Shell Microphone Array," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 4, pp. 740–747 (2008 May), https://doi.org/10.1109/TASL.2008.920059.

[32] A. Kuntz and R. Rabenstein, "Wave Field Analysis Using Multiple Radii Measurements," presented at the *2009 Workshop on Applications of Signal Processing to Audio and Acoustics (WASSPA)*, pp. 317–320 (2009 Oct.), https://doi.org/10.1109/ASPAA.2009.5346537.

[33] H. Chen, T. D. Abhayapala, and W. Zhang, "Theory and Design of Compact Hybrid Microphone Arrays on Two-Dimensional Planes for Three-Dimensional Soundfield Analysis," *The Journal of the Acoustical Society of America*, vol. 138, no. 5, pp. 3081–3092 (2015 Nov.), https://doi.org/10.1121/1.4934953.

[34] A. Kuntz and R. Rabenstein, "Cardioid Pattern Optimization for a Virtual Circular Microphone Array," presented at the *EAA Symposium on Auralization*, pp. 1–4 (2009).

[35] V. Tourbabin and B. Rafaely, "On the Consistent Use of Space and Time Conventions in Array Processing," *Acta Acustica united with Acustica*, vol. 101, no. 3, pp. 470–473 (2015 May/Jun.), https://doi.org/10.3813/AAA.918843.

[36] H. Teutsch, *Modal Array Signal Processing: Principles and Applications of Acoustic Wavefield Decomposition*, vol. 348 (Springer, Berlin, Germany, 2007), https://doi.org/10.1007/978-3-540-40896-3.

[37] H. Teutsch and W. Kellermann, "Acoustic Source Detection and Localization Based on Wavefield Decomposition Using Circular Microphone Arrays," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2724–2736 (2006), https://doi.org/10.1121/1.2346089.

[38] J. Trevino, S. Koyama, S. Sakamoto, and Y. Suzuki, "Mixed-Order Ambisonics Encoding of Cylindrical Microphone Array Signals," *Acoustical Science and Technology*, vol. 35, no. 3, pp. 174–177 (2014), https://doi.org/10.1250/ast.35.174.

[39] Mh Acoustics, "EM32 Eigenmike microphone array release notes (v17.0)," Tech. rep. (2013).

[40] J. Meyer, "Beamforming for a Circular Microphone Array Mounted on Spherically Shaped Objects," *The Journal of the Acoustical Society of America*, vol. 109, no. 1, pp. 185–193 (2001), https://doi.org/10.1121/1.1329616.

[41] D. P. Jarrett, E. A. P. Habets, M. R. P. Thomas, P. A. Naylor, "Rigid sphere room impulse response simulation: Algorithm and applications," *The Journal of the Acoustical Society of America*, vol. 132, no. 3, pp. 1462–1472 (2012), https://doi.org/10.1121/1.4740497.

[42] I. A. McCowan, *Robust Speech Recognition Using Microphone Arrays*, Ph.D. thesis, Queensland University of Technology (2001).

[43] H. Cox, R. M. Zeskind, and T. Kooij, "Practical Supergain," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 3, pp. 393–398 (1986 Jun.), https://doi.org/10.1109/TASSP.1986.1164847.

[44] A. Farina, A. Capra, L. Chiesi, and L. Scopece, "A Spherical Microphone Array for Synthesizing Virtual Directive Microphones in Live Broadcasting and in Post Production," presented at the *AES 40th Intenational Conference: Spatial Audio: Sense the Sound of Space* (2010), conference paper 3-1.

[45] A. Farina, A. Amendola, L. Chiesi, A. Capra, and S. Campanini, "Spatial PCM Sampling: A New Method for Sound Recording and Playback," presented at the *AES 52nd International Conference: Sound Field Control-Engineering and Perception* (2013), conference paper 7-2.

[46] M. Blanco Galindo, *Microphone Array Beamforming for Spatial Audio Object Capture*, Ph.D. thesis, University of Surrey (2020).

[47] G. F. Kuhn, "Model for the Interaural Time Differences in the Azimuthal Plane," *The Journal of the Acoustical Society of America*, vol. 62, no. 1, pp. 157–167 (1977), https://doi.org/10.1121/1.381498.

[48] M. M. Goodwin and G. W. Elko, "Constant Beamwidth Beamforming," presented at the *1993 International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. 169–172 (1993 Apr.), https://doi.org/10.1109/ICASSP.1993.319082.

[49] J. Benesty, J. Chen, and I. Cohen, *Design of Circular Differential Microphone Arrays*, vol. 12 (Springer, London, UK) (2015), https://doi.org/10.1007/978-3-319-14842-7.

[50] ITU-R BS.1534-3, "Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems," Tech. rep. (2015).

[51] N. Ono, Z. Rafii, D. Kitamura, N. Ito, and A. Liutkus, "The 2015 Signal Separation Evaluation Campaign," presented at the *International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pp. 387–395 (2015).

[52] ITU-R BS.1770-4, "Algorithms to Measure Audio Programme Loudness and True-Peak Audio Level," Tech. rep. (2015).

[53] J. Bi, *Sensory Discrimination Tests and Measurements: Sensometrics in Sensory Evaluation* 2nd ed. (John Wiley & Sons, Chichester, UK, 2015), https://doi.org/10.1002/9781118994863.

THE AUTHORS

Miguel Blanco Galindo

Philip Coleman

Philip Jackson

Miguel Blanco Galindo received his B.Eng. degree in Telecommunications Engineering Major in Sound and Image at the Technical University of Madrid in 2011 and his M.Sc. in Engineering Acoustics at the Institute of Sound & Vibration Research (University of Southampton) in 2012. After 2.5 years working as the principal researcher at the consultancy MACH Acoustics, in 2015, he began to pursue a Ph.D. at the University of Surrey, investigating microphone array beamforming for spatial audio object capture as part of the S3A project. Miguel undertook an internship at Apple in spatial audio capturing methods in 2018. He recently completed his Ph.D. and is now working as an audio signal processing engineer at Logitech.

•

Philip Coleman is currently a Lecturer in Audio at the Institute of Sound Recording, University of Surrey, UK. Previously, he worked in the Centre for Vision, Speech and Signal Processing (University of Surrey) as a Research Fellow on the project S3A: Future spatial audio for an immersive listening experience at home. He received a Ph.D. in 2014 on the topic of loudspeaker array processing for personal audio (University of Surrey), as part of the perceptually optimized sound zones (POSZ) project. His research interests are broadly in the domain of engineering and perception of 3D spatial audio, including object-based audio, immersive reverberation, sound field control, loudspeaker and microphone array processing, and enabling new user experiences in spatial audio.

•

Philip Jackson is a Reader in Machine Audition at the Centre for Vision, Speech & Signal Processing (CVSSP, University of Surrey, UK) with an M.A. in Engineering (Cambridge University, UK) and a Ph.D. in Electronic Engineering (University of Southampton, UK). His broad interests in acoustical signals have led to research contributions in human speech production and perception, auditory localization and recognition, audio-visual machine perception, blind source separation, articulatory modeling, visual speech synthesis, sound field control, and spatial audio capture, reverberation, reproduction, and quality evaluation [Google Scholar: bit.ly/2oTRw1C]. He was research theme leader on audio-visual algorithms for object-based spatial audio in the S3A project funded in the UK by EPSRC and investigates new methods to manipulate sound spatially in current projects with BBC R&D and Bang & Olufsen.