

Perceived Quality of Binaural Rendering From Baffled Microphone Arrays Evaluated Without an Explicit Reference

HANNES HELMHOLTZ,^{1*} *AES Student Member*, JEFFERY CRUKLEY,^{2,3}
 (hannes.helmholtz@chalmers.se) (crukley@mcmaster.ca)

SEBASTIÀ V. AMENGUAL GARÍ,⁴ *AES Member*, ZAMIR BEN-HUR,⁴ AND
 (samengual@meta.com) (zamirbh@meta.com)

JENS AHRENS,¹ *AES Member*
 (jens.ahrens@chalmers.se)

¹*Division of Applied Acoustics, Chalmers University of Technology, Gothenburg, Sweden*

²*Data Science and Statistics, Toronto, ON, Canada*

³*Department of Psychology, Neuroscience, and Behaviour, McMaster University, Canada*

⁴*Reality Labs Research, Meta, Redmond, WA, USA*

The authors present a perceptual evaluation of the binaural rendering quality of signals from several types of baffled microphone arrays. They employ the multi-stimulus category rating (MuSCR) paradigm that does not require a reference stimulus. The tested conditions also comprise a very high numerical accuracy stimulus, given the highest quality rating in approximately half of the multi-stimulus trials. A comparison with the literature on spherical microphone arrays (SMAs) shows that MuSCR allows for drawing representative conclusions regarding the dependency of the perceived quality on the array and rendering parameters as in previous experiments with an explicit high-fidelity reference. The authors applied the MuSCR paradigm to evaluate the perceived reproduction quality of equatorial microphone arrays (EMAs) with microphones only along the equator of the spherical baffle and of equatorial arrays with a nonspherical baffle (XMAS). The results endorse the observations from SMAs that an increasing spherical harmonic order leads to improved perceived quality. The authors also confirm that EMAs lead to the same perceived quality as SMAs despite the substantial difference in the number of microphones. Magnitude equalization of artifacts from spatial undersampling can be very effective for XMAS the raw solution deviates significantly from the other array types at high frequencies.

0 INTRODUCTION

The accurate capture and reproduction of existing acoustic events and environments are of relevance to many immersive experiences and virtual reality applications. A spatial reproduction of the audio content allows interactions with the presented auditory scene with increasing *degrees of freedom* (DOF; e.g., changes in listener orientation or position through sound field rotation or translation, respectively). *Spherical harmonics* (SHs) [1] are an efficient and flexible mathematical framework for the storage and subse-

quent auralization of the captured audio content. Although a broad range of reproduction setups are possible, the authors target the application of binaural rendering to provide a first-person experience for a listener with headphones. The rendering utilizes *head-related transfer functions* (HRTFs) as an acoustic representation of the physiological shape of the listener's ears, head, and torso.

Spherical microphone arrays (SMAs) employ microphones distributed over a spherical surface of an open or rigid baffle and are an established tool for the capture and subsequent analysis or auralization of spatial sound fields (e.g., [2]). A convenient property of SMAs is that their capture properties are independent of the direction of sound incidence. *Equatorial microphone arrays* (EMAs) also employ a spherical baffle, whereby they achieve the same SH

*To whom correspondence should be addressed, hannes.helmholtz@chalmers.se. Last updated: August 7, 2024.

order N with a significantly lower number of $2N + 1$ microphones [3] compared to at least $(N + 1)^2$ for SMAs. This comes at the price of partially incorrect elevation cues in the binaural signals for sounds impinging from outside the equatorial plane while interaural time delays are preserved [4]. Possible perceptual implications of this have not been documented formally, which is one of the purposes of the present article. It was demonstrated for room environments that EMAs provide similar spectral accuracy as SMAs of the same SH order so long as the sound source is close to the equatorial plane [5].

The requirement of a spherical baffle was removed in [4] for both SMAs and EMAs, which led to the new category of *nonspherical microphone arrays* (XMAS). XMAS require a calibration measurement under anechoic conditions, as the acoustic influence of the baffle cannot be modeled analytically. The authors only consider equatorial XMAS in the present work because of their practical relevance.

Practical implementations of either SMAs, EMAs, or XMAS require using a finite number of microphones, which leads to spatial undersampling of the captured sound field. This produces a loss of information in the form of SH order truncation [6] and ambiguities in spatial information in the form of spatial aliasing [7]. The spatial undersampling manifests as deviations in the magnitude and time composition of the binaural output signals, which occur primarily at frequencies above the spatial aliasing frequency, which in turn depends on the size of the array and the number of microphones employed [8]. Perceptually, the loss of spatial information can lead to reduced externalization and localizability of virtual sound sources as well as impairments in the perceived location or timbre during head movements (see, e.g., [7, 9]).

SMAs have been evaluated extensively in the literature [7, 10–12]. Usually, a variant of the *multi-stimulus test with hidden reference and anchor* (MUSHRA) [13] paradigm was used in which the subjects compared a given condition to a reference with high accuracy. This reference are typically binaural signals from a mannequin measurement or an emulation of a very high-accuracy SMA achieved through sequential measures. The most important conclusion is that differences toward the reference become minor for higher SH orders and frontal sound incidence. With state-of-the-art equalization of the binaural signals, diminishing returns are observed from approximately SH order 7, above which some difference remains detectable and does not diminish further [12, 10, 11]. EMAs and XMAS have not been evaluated formally.

The disadvantages of the MUSHRA paradigm are 1) the circumstance that it requires a high-accuracy reference, which is not available in many relevant situations, and 2) the fact that it measures a deviation from the reference without providing information about what the deviation entails. It is important to note that a perceptual difference from a reference in a MUSHRA-like experiment does not necessarily mean a quality reduction [14].

To overcome the disadvantages, the authors use a variant of *multi-stimulus category rating* (MuSCR) [15] in which the subjects rate the overall quality on a continu-

ous scale without a reference stimulus. MuSCR can help identify the most favorable condition, but the ability to conclude what property makes a stimulus receive a higher rating is minimal. Each trial in this experiment comprises a high-accuracy stimulus, which enables the authors to compare the results from their experiment to those from the MUSHRA-like experiments in the literature. The authors first demonstrate the effectiveness of the MuSCR paradigm in the present context with SMAs and then evaluate EMAs and XMAS with the same method.

The structure of this article is as follows. The methods for the binaural rendering of the different microphone arrays are summarized in SEC. 1, followed by a description of the employed acoustic data sets in SEC. 2. Subsequently, the setup of the implemented evaluation procedure is specified in SEC. 3, and the perceptual results are summarized and discussed in SEC. 4. The article is concluded in SEC. 5.

1 PROCESSING METHOD

For SMAs, the captured sound pressure field, including scattering on the surface of the baffle, is decomposed in the SH domain [1]. The resulting coefficients are a complete representation of the incident sound field. In practice, this transformation is order-limited, leading to the impairments from spatial undersampling due to the finite number of microphones employed. In the case of SMAs, the SH coefficients allow for removing the influence of the rigid baffle with analytic solutions by so-called radial filtering [2].

The solution for EMAs performs a circular harmonic decomposition of the captured sound field from which an SH representation is computed [3]. This yields a horizontally propagating sound field (i.e., a projection of the impinging sound field on the array's equatorial plane). Like with SMAs, the effect of the baffle is removed utilizing radial filters that are computed analytically. EMAs and SMAs of identical size and SH order exhibit similar properties regarding magnitude deviations due to spatial undersampling and during 3-DOF head rotations [5]. In either case, equalizing the binaural signals is useful [12]. The authors' equalized SMA and EMA conditions employ *end-to-end magnitude least squares* (eMagLS) rendering as described in [16] and [5].

XMAS perform SH decomposition and removal of the baffle's effect based on linear filters obtained from a least-squares fit on anechoic calibration data in the frequency domain [4]. Approximately spherical body shapes, such as a human head, have been shown to work well. Even convex body shapes with corners are possible [17]. The binaural signals from XMAS exhibit significant deviations from the ground truth above the spatial aliasing frequency. This is equalized via a MagLS-based filter applied in the SH domain.

In the rendering, the listener's HRTFs are decomposed in the SH domain to compute the virtual left and right ear signals that arise as if the listener were exposed to the captured sound field. The authors implement the rendering into *binaural room impulse responses* (BRIRs) by multiplying

the sound field and HRTF coefficients directly in the SH domain, referred to as *modal rendering* [18]. An alternative method exists via spatially sampling the HRTF coefficients to a discrete grid at the SH order of the microphone array as *virtual loudspeakers* [19]. A summary of the mathematical pitfalls for binaural rendering in the SH domain and compatibility to established *ambisonic* frameworks is presented in [20].

Note that encoding the captured sound field as described above is signal-independent. The alternative is parametric encoding that separates direct sound field components from diffuse ones (e.g., for the binaural reproduction from SMAs [21] and arbitrary array geometries [22]). For this experiment, the authors perform the analytic SH transformations for SMAs and HRTFs by discretizing the transformation integral and including the specific quadrature weights of the sampling grids.

The SMA and EMA renderings without equalization employ modal radial filters with a length of 4,096 samples and Tikhonov regularization [23] of the filter magnitude at 18 dB or equivalently 40 dB according to the convention for EMA radial filters [3, Eqs. (18) and (20)]. The eMagLS filters are generated with identical parameters to [16] and [5] (i.e., a target length of 512 samples and the regularization weight of 0.01). Thereby, the transition frequency of the eMagLS processing is adjusted to the array order by $N \cdot 500$ Hz resulting in 1, 2, and 4 kHz for the equalized SMAs and EMAs, respectively. The XMA calibration and equalization filters for each SH order are generated with a target length of 1,024 samples via a least-squares fit in the frequency domain. For all array and equalization configurations, the delays arising from different processing filter lengths are compensated for by time alignment. Further implementation-specific parameters and their adopted settings are available in the MATLAB rendering scripts¹ and the supplementary materials.²

2 DATA

An overview of all array/rendering configurations and acoustic scenarios in this experiment is given in Table 1. The abbreviations are used throughout the manuscript to encode the employed study conditions and trials. All scripts for the capture, post-processing, and the resulting acoustic impulse responses are based on open-source software and available as a self-contained dataset.³ The MATLAB scripts for BRIR rendering based on the above data set are published in a code repository.¹ The resulting prerendered BRIRs are available as part of the supplementary materials² (see SEC. 3).

An anechoic percussion stimulus (bass drum, snare, and hi-hat) of 7.7-s length serves as the program material termed *drums*. It provides a spectrally broad and temporally transient signal, which was shown to be suitable for the perceptual evaluation of binaural rendering from SMAs [7, 11, 9]. Depending on the acoustic scenario, percussive signals

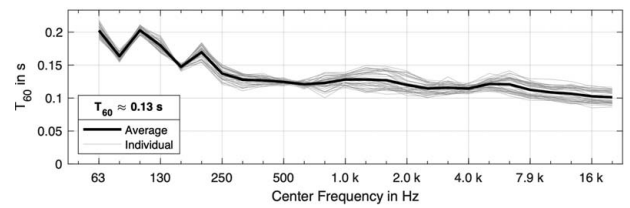


Fig. 1. Reverberation decay time in third-octave bands of the auralized source/room environment as estimated from averaging the 42 individual impulse responses (gray lines) of the SMA for SH4.

yielded similar to slightly more critical results for revealing differences between array conditions. The authors, therefore, limit this investigation to the drums audio source signal in order to limit the required experimental time and number of trials.

2.1 Room Environment

In this experiment, the subjects evaluate the static acoustic scenario shown in Fig. 2(a). The room constitutes a laboratory environment at Chalmers University of Technology, Gothenburg, Sweden, with the dimensions of approximately $4.0 \times 6.8 \times 2.4$ m (width, length, height) and a volume of ≈ 66 m³. The sound source is a *Genelec 8030A* loudspeaker placed on a speaker stand. The receiver position is approximately in the middle of the room, and the source is at a distance of 2.6 m along the long axis and 1.2 m to the back wall of the room. The center points of the acoustic source and the receivers are placed at a height around 1.2 m, the typical ear height of a sitting adult listener. The source elevation is therefore labeled *ear* in Table 1.

The room [see Fig. 2(a)] exhibits an acoustically dry environment due to a multilayer construction with an air gap behind absorptive panels at the walls and a thick layer of porous absorber material at the ceiling. The concrete floor is intentionally left without absorption in order to present a challenging scenario for EMAs and XMA with nonhorizontal sound incidence from a floor reflection that may be present in many practical recording scenarios. The room exhibits an average reverberation decay time of 0.13 s with individual third-octave band values shown in Fig. 1.

2.2 Array Data

High-resolution SMA measurements are captured with a *DPA 4060* microphone flush-mounted on the spherical surface of a rigid wooden baffle of 8.5-cm radius, as shown in Fig. 2(b). Thereby, 2,702 points on a Lebedev grid are captured sequentially by rotating the baffle on a *VariSphear* turntable similar to the one used in [24]. In order to minimize measurement time, the authors generate arbitrary lower-resolution microphone grids from this data by spatially resampling the 44th-order data in the SH domain. The method yields accurate results (i.e., the post hoc resampled data was confirmed to be equivalent to a sequential SMA room measurement at a grid suitable for decomposition at SH order 8). For the experiment configurations in Table

¹<https://github.com/HaHeho/baffled-arrays-to-binaural>.

²<https://doi.org/10.5281/zenodo.10901444>.

³<https://doi.org/10.5281/zenodo.8206571>.

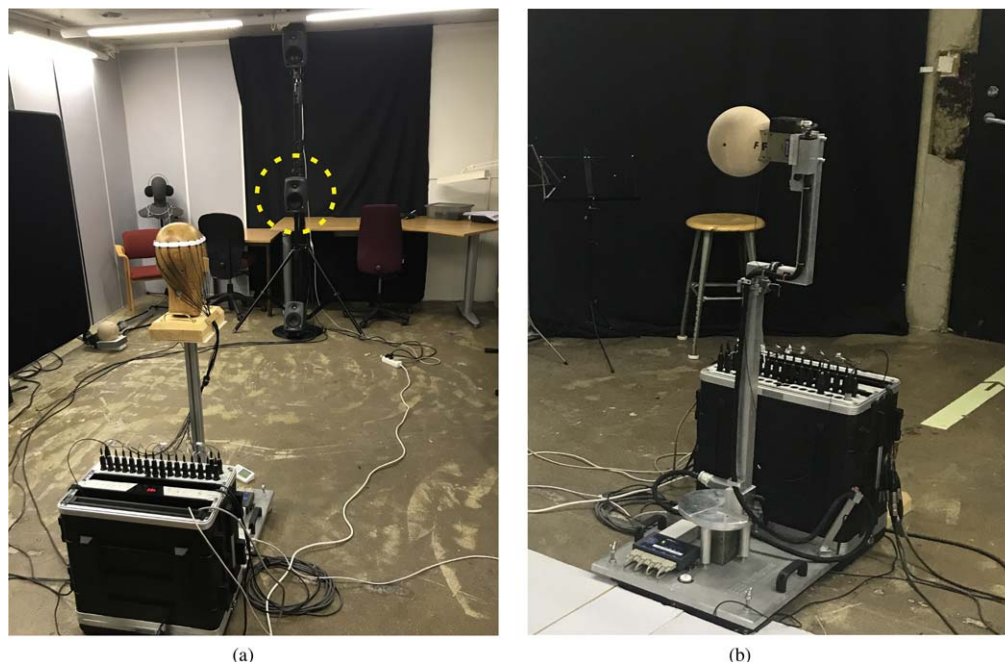


Fig. 2. Capture of the sound source and room environment with different microphone arrays auralized in this experiment. (a) Utilized sound source at ear height (dashed circle) and head-mounted XMA, which is spatially subsampled to lower-order XMAs. (b) *VariSphear* turntable to capture a high-accuracy SMA, which is spatially resampled to lower-order SMA and EMA grids.

1, SMA data are generated on a t-design grid⁴ for SH1 (6 points), SH2 (14 points), SH4 (42 points), and SH8 (146 points). EMA data is generated by spatial resampling to an equiangular grid for SH2 (5 points), SH4 (9 points), and SH8 (17 points). The number in the configuration name represents the SH order that the respective microphone array provides.

XMA measurements are captured with 18 microphones of the type *Rode Lavalier GO* mounted in an elastic band on a wooden head-shaped baffle shown in Fig. 2(a). The resulting microphone arrangement is an ellipse with 7.5-cm and 10.5-cm semiaxis length. The authors use the room measurement with the nose of the baffle facing the acoustic source. Subsets of every third and every second microphone are taken from each measurement to yield lower-resolution XMA data for the respective orders SH2 (6 points), SH4 (9 points), and SH8 (18 points).

Anechoic measurements of the XMA setup are required to generate the array calibration and equalization filters (see SEC. 1). Array transfer functions for 360° incidence directions in the equatorial plane are captured in 1°-steps. These measurements demand careful postprocessing of the impulse responses because of scattering off the measurement apparatus. Therefore, the authors applied frequency-dependent truncation based on [25]⁵ with long time windows to maintain frequency resolution at low frequencies and short time windows to remove spurious scattering contributions at medium to high frequencies.

The time and frequency response of the arrays' microphones was fully compensated by deconvolution with their on-axis free-field response. Despite using different microphones, this yields equivalent spectral responses from SMA/EMA and XMA renderings. The authors refer the reader to the documentation of the measurement dataset³ for a detailed description.

2.3 HRTF Data

The authors employ the HRTFs of a *G.R.A.S KEMAR* acoustic mannequin for the binaural rendering of all array configurations. The HRTFs were modified from [26], which were measured on the authors' unit of KEMAR in three ways. Firstly, the original data set with 64,800 measurement points was spatially resampled in the SH domain to 2,702 points on a Lebedev grid supporting SH order 44. This was done to lower computation resources for all binaural renderings without a loss in spatial resolution. Secondly, the HRTFs were delayed so that the signals started with the lowest energy sample out of all incidence directions and both ears, which amounted to a circular shift of 92 samples. Lastly, the authors applied cosine-squared windows with a one-sided length of 16 samples to fade the start and end of the HRTFs to further limit artifacts from time-discontinuities.

2.4 BRIR Rendering

The binaural rendering from the array room impulse responses (see SEC. 2.2) is performed in the MATLAB programming language and software environment according to the methods described in SEC. 1. Thereby, rendering configurations either do or do not employ the described

⁴<https://web.maths.unsw.edu.au/~rsw/Sphere/EffSphDes/sf.html>.

⁵<https://github.com/floriandenk/FDT>.

Table 1. Each subject of the listening experiment is presented 48 stimuli, consisting of six trials with randomized acoustic scenarios (left) and eight conditions with randomized array rendering configurations (right).

Trials (6)					Conditions (8)		
Source Signal	Source Elevation	Source Direction	Room	Array Type	Array Order	Equalization	
drums	× ear	× {frontal, lateral}	× Lab	SMA	SH44	EQ off	High anchor (1)
				{SMA, EMA, XMA}	{SH8, SH4, SH2}	× {EQ on, EQ off}	Combinations (6)
				SMA	SH1	EQ off	Low anchor (1)

equalization method (EQ on, EQ off) for the respective array type. Each set of BRIRs is precomputed for 1-DOF reproduction with 360° head orientations in 1° steps in the horizontal plane. This is done for each combination of array type and rendering condition according to Table 1, excluding the two attributes source signal (drums) and relative source incidence direction (frontal, lateral), which are implemented during runtime of the user study (see SEC. 3.1).

After rendering, the impulse response set is truncated at the last sample to reach 80 dB under the global BRIR peak and faded out by a one-sided cosine-squared window of 64 samples. This is done to save storage and computational resources during the real-time auralization in the user study (see SEC. 3.1). The authors found the truncation imperceptible when convolving the BRIRs with a source signal. The truncation results in BRIR lengths between 0.19 and 0.25 s for the acoustic environment employed in this investigation (see SEC. 2.1).

3 PERCEPTUAL EVALUATION

The authors implement a perceptual user study with a complete factorial-within-subjects design (i.e., all subjects were presented with stimuli in all conditions). Static binaural listening examples of all presented study conditions are available with annotations.⁶ The auralized BRIRs (see SEC. 2) and the tools to recreate the listening test infrastructure (code and configuration files; see SEC. 3.1) are available as supplementary materials.² This dataset also contains the resources of the statistical analysis (see SEC. 3.4).

3.1 Setup

The subjects' task is to rate the quality of the conditions on a quasi-continuous scale, which allows for ratings between the labels “very high,” “high,” “medium,” “low,” and “very low,” as shown in Fig. 3. The study instructions encouraged the subjects to use the full range of the scale. The graphical user interface is based on open-source software⁷ designed initially to conduct MUSHRA experiments, which

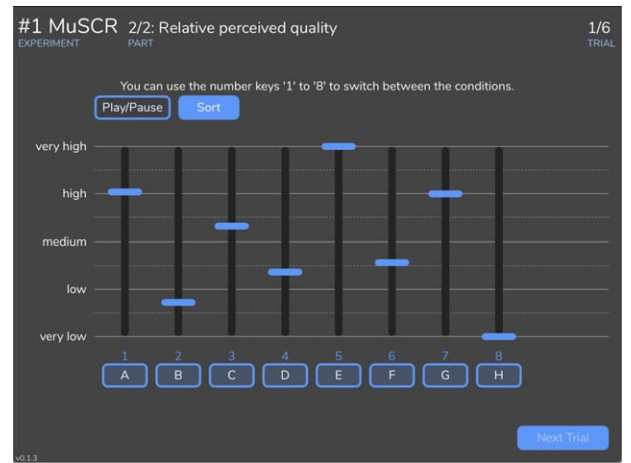


Fig. 3. The study interface allows for rating the relative perceived quality of eight conditions in a MuSCR paradigm without an explicit reference.

the authors modified for the MuSCR paradigm by removing the explicit reference stimulus. The given listeners' ratings are converted to an integer scale from 100 to 0 and are exported accompanied by metadata that describes the trial and condition configurations and the presentation order. Upon user interaction, the application sends network messages via the Open Sound Control protocol to signal respective events and progression of the study procedure (e.g., to the audio playback and processing applications).

The pre-computed BRIR sets (see SEC. 2.4) are loaded by the *SoundScape Renderer* (SSR) [27]. This application performs real-time auralization by convolving an input signal with a pair of BRIRs selected based on the instantaneous head orientation of the listener. The auralization tool allows for tracking the listener's head orientation around the vertical axis (i.e., only 1-DOF yaw rotations are considered in this experiment). This is a common limitation in the SMA literature, and it has been shown to be sufficient for providing an indistinguishable reproduction compared to a binaural mannequin reference when presenting sources close to the horizontal plane [28]. The head-tracking data are read with low latency from a *Polhemus Fastrak* receiver attached to the listener's headphones. For all arrays, the two nominal sound incidence directions, *frontal* and *lat-*

⁶<http://www.ta.chalmers.se/research/audio-technology-group/audio-examples/jaes-2024a/>.

⁷<https://git.iem.at/rudrich/mushra>.

eral (see Table 1), were realized using an offset on the head tracking data. The SSR receives network messages from the user interface regarding what stimulus condition to play. The source signal plays continuously, and the transition between stimuli occurs through a cross-fade.

The authors manually equalize the loudness of all stimuli because they are not interested in loudness differences between rendering configurations. The authors are unaware of an implementation to automate the loudness equalization of multidirectional BRIRs. Thus, the replay level of each stimulus is adjusted manually in 1-dB steps by comparative listening of the head-tracked reproduction in the SSR.

A separate *Pure Data* (Pd)⁸ instance is used to transform Open Sound Control messages from the study interface into network messages for the SSR, which uses a different format, and to relay signals to other Pd patches via built-in objects. Furthermore, Pd provides the playback of the source signal to be auralized with the SSR. This experiment uses an anechoic drum signal for all presented conditions. The source signal is looped if not intentionally paused by the subject (see Fig. 3). The *JACK Audio Connection Kit* (JACK)⁹ is utilized to route all audio signals between Pd and SSR instances and as an interface to the output hardware for headphone playback. All real-time processing is performed at a JACK block size of 256 samples, which is short enough to prevent audible latency of the head tracking.

Sennheiser HD650 headphones driven by a *Focusrite Scarlett 2i2* were used by all subjects. Accurate headphone compensation is essential for spectrally balanced reproduction of binaural signals [29]. Therefore, the authors averaged ten measurements of the transfer functions between the headphones and their mannequin head. The equalization filters are generated by an inversion method including phase-linearization based on [30] with a length of 2,048 samples. The filters are inserted into the processing pipeline after the output of the SSR by convolution in Pd.

3.2 Procedure

The experiment was conducted in a quiet, undisturbed office environment at Chalmers University of Technology, Gothenburg, Sweden. All subjects were provided with written and verbal instructions before the experiment. Thereby, descriptions for the perceptual attributes “timbre,” “temporal behavior,” “spatial impression,” and “general quality” were provided to reduce the uncertainty due to the interpretation of the terms. The instructions also highlighted that head-tracking was applied to encourage the subjects to use this feature.

The experiment started with a familiarization phase in which the subjects were asked to take as much time as needed to become familiar with the study interface, acoustic scenario, range in quality between the different rendering configurations, and perceptual evaluation task itself. Eight conditions were presented, manually picked to include representative examples from the entire range of qualities accompanied by the two anchor configurations according to

Table 1. This was performed for two trials for the frontal and lateral relative source directions.

After completion of the familiarization, the headphone playback volume was fixed to a comfortable listening level. Every subject was presented with eight conditions in six trials according to Table 1 presented in random order. Each condition could be listened to and repeated without a time limit. During a trial, the order of the conditions could be sorted by the subject by their current rating to allow for easier comparisons between stimuli of similar quality (see the button *Sort* in Fig. 3). After the experiment, the subjects were asked to briefly describe their response strategy and answer a questionnaire on their experience with spatial audio.

3.3 Subjects

The authors collected data from 20 subjects (6 female, 14 male) aged between 23 years and 59 years (median 31 years). Participants completed the main part of the experiment in a duration between 9 and 37 min (median 20 min). Four subjects each reported “none,” “<1 year,” “1 to 3 years,” “3 to 5 years,” and “>5 years” of experience with listening to binaural audio. Although recommended differently in the description of the MuSCR method [15], the authors chose to keep inexperienced listeners in the final data analysis to investigate differences in response behavior optionally. 14 subjects reported having participated at least more than twice in perceptual listening experiments. All subjects reported no known hearing impairment except one; the oldest participant described having mild age-induced hearing loss. The responses of this subject are qualitatively similar to those of others, which is why they were kept in the analysis.

3.4 Statistical Analysis

The authors conducted their analyses and constructed their models under a Bayesian framework using the *CmdStanR* package¹⁰ (version 0.7.1) as an interface to the Stan programming language (version 2.34.1) [31] in the R statistical computing software (version 4.3.3) [32]. The raw response data, the R scripts to reproduce the statistical analysis and plotting, and additional analysis resources (plots and tables) are available as part of the supplementary materials.²

Bayesian regression models provide distributions of parameter estimates for each predictor on each outcome, which affords direct quantification of uncertainty in the estimates through credible intervals, which represents a range (percentage of the posterior distribution) containing the most probable/compatible parameter estimates conditioned on the data, priors, and likelihood. Because Bayesian regression treats data as fixed and parameters as random, Bayesian analyses do not require the calculation of *p* values or their associated confidence intervals. Bayesian regression produces a posterior probability distribution of parameter estimates, which the authors use to draw infer-

⁸<https://puredata.info/>.

⁹<https://jackaudio.org/>.

¹⁰<https://mc-stan.org/cmdstanr/>.

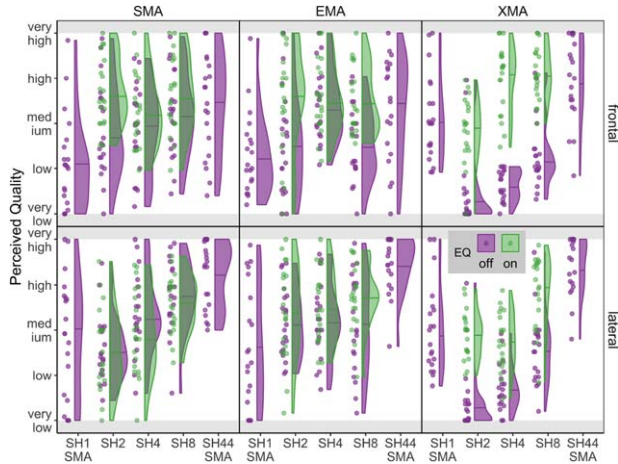


Fig. 4. Anterior observed subject ratings of perceived quality for each array order by array type (horizontal), sound incidence direction (vertical), and equalization (dark purple/light green areas). Points present individual subject ratings, and violin plots with lines representing the median rating provide an alternative illustration of the data distribution.

ences, including direct uncertainty quantification in their estimates [33].

First, the data were converted from quality ratings to a condition rank order (from 1 to 8, with 8 being the condition with the highest rating) within each presented trial. The authors will motivate this choice in SEC. 4.1. Conditions receiving the same rating in a given trial were randomly assigned either the lower or higher rank. Contrary to other draw resolution strategies, this has the advantage of yielding uniform distribution of ranks over all data points.

The resulting ranks were then modeled as ordered discrete responses using a hierarchical generalized linear regression (cumulative) model. The authors specified seven thresholds corresponding to the eight available ranks and allowed the threshold positions on the latent scale to vary. The model includes subject rating, array type (SMA, EMA, XMA), array order (SH1, SH2, SH4, SH8, SH44), equalization (EQ on, EQ off), source incidence direction (frontal, lateral), and interaction terms as population-level effects and effects for subjects, trial, and slope of individual ratings as group-level (varying) effects. To account for possible variance differences between experience levels, the authors estimated variance as a function of reported experience, time spent on the task, and the interaction term, as laid out by Eqs. (1)–(10) in APPENDIX A.1.

The authors used Pareto smoothed importance sampling leave-one-out cross-validation [34] to estimate the expected log predictive density and to evaluate the fit of their model to the data. All estimated generalized Pareto distribution shape parameters κ were below 0.7, suggesting no overly influential observations (outliers) in the data. The model provides a good fit to the data as visualized by contrasting observed and predicted response ranks in the results section in Fig. 6.

The analysis produced an estimate of the posterior distribution of joint parameter estimates [35]. This estimate

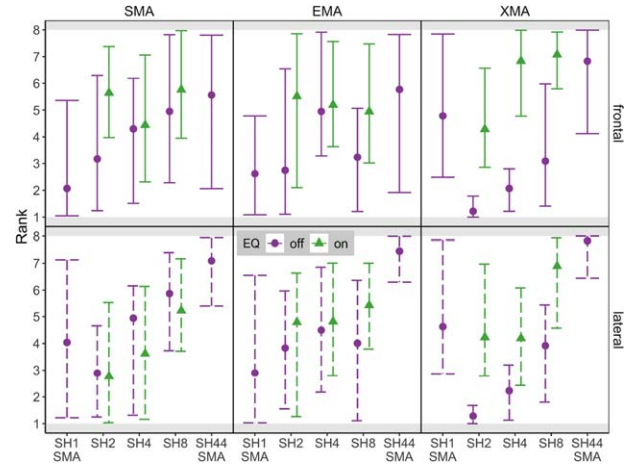


Fig. 5. Posterior predicted distribution of ranks for each array order by array type (horizontal), sound incidence direction (vertical and solid/dashed lines), and equalization (dark purple/light green). Points represent median rank, and error bars depict the 89% highest density credible interval.

included the probability of each rank for each combination of subject, array order, equalization, array type, source direction, and trial for each Markov-chain Monte Carlo iteration. To derive a single estimate of ranks for each subject, array order, equalization, array type, source direction, and trial, the authors calculated a weighted sum of ranking (sum-score) $\Psi_i = \sum_{k=1}^8 p(i)k$, where p is the expected probability of rank k at each iteration i . The resulting sum scores can be interpreted as mean ranks.

4 RESULTS AND DISCUSSION

4.1 Observations From the MuSCR paradigm

Fig. 4 summarizes the perceptual ratings from the 20 subjects described in SEC. 3.3. The discussion starts with general observations regarding the potential influences of the employed study design. In [14], the MuSCR paradigm was directly compared to, among others, a MUSHRA design for evaluating the overall audio quality of binaural rendering in a virtual reality experiment. The results showed similar trends regarding median quality but notable differences regarding the observed ratings.

Firstly, utilization of the scale range was more narrow in MuSCR compared to MUSHRA, which was attributed to the subject's interpretation of the quality labels or potential response scaling biases in the absence of an explicit reference [14]. An investigation of the individual responses from this experiment reveals that half of the subjects made use of the full range of the perceptual scale (expressed as integers from 100 to 0). The smallest range from one subject pooled over all six trials is 68. Ratings from the eight conditions per trial vary regarding their submitted minimum (0 to 86, mean 16.0, SD 19.1) and maximum (51 to 100, mean 89.9, SD 11.4), resulting in a broad use of the scale range (12–100, mean 73.9, SD 25.1) in most trials. This data instead exhibit a range equalizing bias (i.e., the utilization of the

entire scale regardless of the actual range of the stimuli [36], instead of the limited range observed in [14]).

Secondly, a larger variability of responses across different acoustic scenes was observed in MuSCR compared to MUSHRA, resulting in fewer significant differences between conditions or decreased discernability/sensitivity of the paradigm [14]. In the results from this experiment, five subjects showed a clear tendency to space their ratings in fixed equal intervals. The nine observed levels coincide with the five qualities and the respective semi-steps between the labels shown in the graphical study interface (see Fig. 3). This is termed stimulus spacing bias (i.e., differences between ratings are compressed or expanded regardless of the stimuli distribution in the perceptual domain [36]). This is one of the potential scaling biases that may be more prevalent in MuSCR when no explicit reference is available [14]. A few significant differences between reproduction conditions due to the increased variance in the response data will also be observed (e.g., see SEC. 4.4).

Lastly, although the data suggest representing an absolute quality scale, the meaning behind scale labels is difficult to quantify due to the scaling biases in the subject's responses mentioned before [14]. This and the resulting large data variances motivate the transformation into a condition rank order before statistical modeling as described in SEC. 3.4. An overview of the results from all study conditions and trials according to Table 1 is shown in Fig. 5.

4.2 Comparison to MUSHRA Results for SMAs

SMAs were evaluated extensively in the literature, often using a MUSHRA-like paradigm in which the quality was rated compared to a high-fidelity reference (e.g., the ground truth ear signals). The condition SH44 in this experiment can be interpreted as such a ground truth signal so that the results for SMAs can be compared with previous studies from the literature. The unequalized (EQ off) SMA conditions serve as systems of well-understood quality to allow for contextualization of the results according to the MuSCR recommendation [15].

It is expected that variances in the response data from the MuSCR paradigm are larger compared with MUSHRA due to various potential scaling biases discussed in SEC. 4.1. Previous experiments in the literature often employed frequentist analysis (e.g., ANOVA) and effect reporting methods. The authors will compare the relative magnitude and differences between their median sum-scores and means/medians provided in the literature. The comparison of result variances from their Bayesian credible intervals and frequentist confidence intervals is not meaningful.

These results can be directly compared to the evaluations performed in [11, 12, 28] since they used theoretically identical rendering methods (there may be slight differences in implementation and parametrization) and an SMA capture device of similar radius (8.75 cm instead of 8.5 cm in this case). When compared with the respective mannequin reference, SMAs achieve a perceptually high reproduction quality (based on mean difference ratings better than “small”) from SH order 7 [12] or 8 [11]. The data in Fig. 5 similarly

show that the perceived quality increases monotonically with SH order (see SEC. 4.4). Very high-resolution SMAs of SH orders 13–20 for frontal and 15–22 for lateral sources were required in order to achieve a reproduction indistinguishable from a mannequin reference [28, Fig. 6]. These results confirm that a rendering at SH8 remains distinguishable from and rated at slightly lower quality than the high-accuracy condition (see SEC. 4.3). Similarly, qualitative differences are observed to be more apparent for lateral sound incidence (see SEC. 4.6).

The eMagLS equalization method (EQ on) can be compared only indirectly because it has not yet been employed in a perceptual study. eMagLS is an improvement of the MagLS equalization method as presented and evaluated in [10]. From anechoic SMA data, MagLS yielded significantly improved reproduction compared with the unequalized rendering [10, Fig. 10], whereas the observed quality from eMagLS in Fig. 5 similarly increases with SH order although with a smaller magnitude over EQ off (see SEC. 4.5). The remaining differences from MagLS toward a mannequin reference were smaller for frontal than lateral sound incidence, which is also observed in the data with eMagLS (see SEC. 4.6). Overall, the authors find similar improvements from the equalization methods, although they remain at a level of “small” differences toward the high-accuracy conditions (see MagLS at SH order 7 in [12, Fig. 9] and eMagLS at SH8 in Fig. 5).

4.3 Selection of the Highest-Quality Condition

The main benefit of the MuSCR paradigm is omitting an explicit reference and, therefore, not preconditioning the subjects on what should be considered the highest rendering quality. Fig. 4 shows that almost all conditions may be rated with very high quality by individual subjects in single trials, except for the unequalized XMA configurations and, in case of lateral incidence, also some spherical array conditions (SMA at SH2; SMA with EQ on at SH4; EMA with EQ off at SH2 and SH4). However, the distribution over the entire quality scale differs between conditions, as visualized by median ratings and violin plots.

The predicted ranks in Fig. 5 show similar trends in medians and credible intervals. In the case of frontal incidence (top), the rank distribution suggests that for all array types, multiple configurations compete for the highest quality condition. Considerable overlaps of the credible intervals suggest that differences between the favored conditions are statistically insignificant, as confirmed in SEC. 4.4. For lateral source directions (bottom), SH44 is more often the highest rated configuration, as further discussed in SEC. 4.6.

Fig. 6 shows an alternative analysis of the rank distribution per configuration. The colored bars present the directly observed percentage of chosen presentations per rank, and the error bars represent the predicted data with median and uncertainty leading to the beforehand investigated posterior rank distributions in Fig. 5. The number of presentations in each subplot per array type, array order, and equalization add up to 100%. The data are combined for all subjects

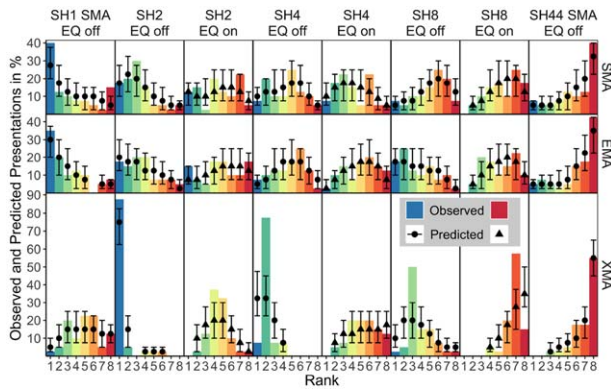


Fig. 6. Anterior observed (shaded/colored bars) and posterior predicted (error bars) percentages of presentations per rank by array order and equalization (horizontal) and array type (vertical). Bars represent the observed percentage, points represent the median predicted percentage, and error bars depict the 89% highest density credible interval.

as well as frontal and lateral source directions, because this is more representative of assessing the overall quality of a configuration when exposed to dynamic acoustic scenarios.

For all array types, SH44 is the condition most often resulting in the highest rank (8) by being given the highest (or tied) rating within a trial. According to Fig. 6, this occurs in around 40% of all observed cases with SMAs (top) and EMAs (middle) and in 55% of the XMA (bottom) trials. Anyone but the highest rank (1 to 7) was given to this condition in as much as $\approx 20\%$ of trials. SH44 received the cumulative sum of the lower-half ranks (1–4) in around 25% of SMA, 20% of EMA, and 5% of XMA trials.

The second-highest rank (7), and therefore best among the practical array configurations excluding the high-accuracy stimulus, was given most often to SH8 with EQ on in 55% of XMA trials and was less distinct from other conditions in around 25% of SMA and EMA trials. To compare, this condition was given the highest rank (8) in around 15% of all trials. The follow-up configurations regarding perceived quality are SH8 with EQ off for SMA, SH2 and SH4 with EQ on for EMA, and with larger perceptual distance SH4 with EQ on for XMA.

4.4 Influence of the Array Order

As discussed for SMA conditions in SEC. 4.2, the authors observe a positive trend in perceived reproduction quality with increasing array order, which aligns with the existing SMA literature. For EMA and XMA, similar monotonically increasing ranks are seen when comparing medians and credible intervals for $SH1 < SH2 < SH4 < SH8 < SH44$ in Fig. 5. Two exceptions arise where changes with increasing array order are nonmonotonic: firstly, for frontal incidence from an SMA with equalization (SH2 vs. SH4) and an EMA without equalization (SH4 vs. SH8), and secondly, for lateral incidence from an SMA without equalization (SH1 vs. SH2). The authors validated the data rendering and rating acquisition multiple times and could not ultimately clarify

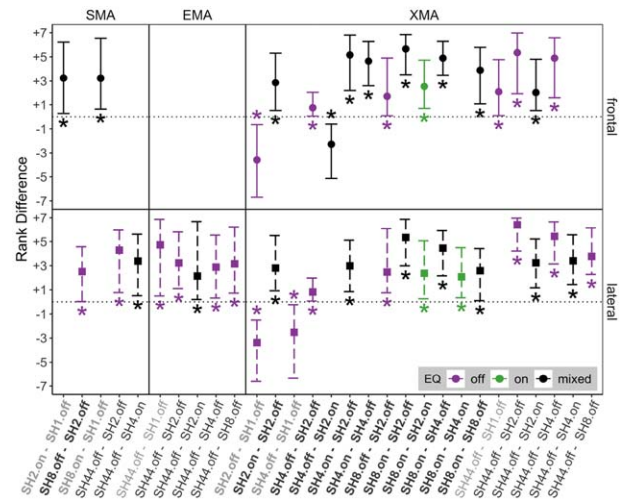


Fig. 7. Posterior distribution of rank differences between all combinations of array orders and equalizations by array type (horizontal) and sound incidence direction (vertical and solid/dashed lines). Points represent median rank, and error bars depict the 89% highest density credible interval. Only combinations with statistically significant differences are shown and indicated by asterisks.

the cause for this observed tendency. Remember that SH1 and SH44 always employ SMAs as identical anchor conditions between trials. Hence, the decrease in quality from the SMA at SH1 to the XMA configurations without equalization is adequate, as discussed in SEC. 5.

The authors also tested for the significance of differences between all combinations of array order and equalization more closely. The significant pairs per array type and source direction are shown in Fig. 7.

For the XMA, around half of all comparisons (31 of 56) are significant from both directions of sound incidence in most cases. When omitting pairs that involve the SH1 and SH44 anchor conditions, most of the remaining comparisons (19 of 30) show significant differences. As supported by Fig. 5, substantial improvements in quality are seen with higher-order XMAs when not employing equalization. With equalization, an increase in XMAs order yields clear quality improvements from either frontal or lateral directions of sound incidence.

For SMA and EMA, only a few comparisons (5 of 56, respectively) are significant with respect to array order and equalization. Differences primarily arise for lateral sound incidence (i.e., significantly lower quality than SH44; see SEC. 4.3 for the comparison against the high-accuracy condition). With and without equalization, the improvements in perceived reproduction quality are only incremental. This may be attributed to the fact that spectral impairments of the binaural output signals through spatial aliasing and SH order truncation are minor even for low-order SMA and EMA configurations in this experiment (see SEC. 4.5 for the discussion of the array radius).

4.5 Influence of the Equalization Method

Fig. 8 summarizes differences in rank due to the employment of equalization. For the XMA, improvements from

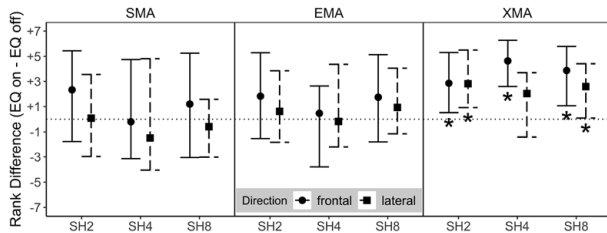


Fig. 8. Posterior distribution of rank differences from EQ off vs. EQ on equalization for each array order by array type (horizontal) and sound incidence direction (solid/dashed lines). Points represent median rank, and error bars depict the 89% highest density credible interval. Asterisks indicate statistically significant differences.

EQ on are directly apparent in Fig. 5 (right) and significant at all investigated array orders. The perceived reproduction quality from the unequalized XMAs is clearly below the provided anchor condition with an SMA at SH1. This can be attributed to the strong attenuation of high frequencies, which is compensated by the employed equalization method (see [4, Fig. 5]). The low variance in the perceptual ratings suggests that the high-pass behavior of the EQ off conditions consistently affects the perceived quality. Therefore, suitable equalization is confirmed to be crucial for the reproduction of XMA captures with perceptual fidelity comparable to the other array types.

For SMAs and EMAs, the contribution of eMagLS equalization is generally positive but statistically insignificant for the presented array configurations. An instrumental investigation of the BRIRs shows that the potential for improvement of the diffuse-field response of the unequalized rendering is small (see [5, Fig. 3] for similar SMA and EMA configurations with 8.75-cm radius). With the present array radii of 8.5 cm, opposing influences on the overall magnitude response at high frequencies (i.e., attenuation from SH order truncation of the HRTFs [6] and amplification from spatial aliasing in the decomposed sound field [7]) are in balance. Especially arrays with pronounced spectral errors above the spatial aliasing frequency can benefit substantially from equalization, as apparent for XMAs in this experiment. This applies to most commercially available SMAs (and equivalent hypothetical EMAs) of smaller size (e.g., the *Eigenmike em32* with 4.2-cm radius showing a significant improvement of the overall spectral balance at high frequencies in [16]).

For SMAs and lateral source incidence, all comparisons show that EQ on (median 2.8, 3.6, 5.2) was rated slightly below EQ off (median 2.9, 4.9, 5.9) at SH2, SH4 and SH8, respectively. The degradation in perceived reproduction quality is statistically insignificant and of small magnitude, as seen in the raw response data in Fig. 4. Notably, the effect is not observed for the EMA. Instrumental investigation of the BRIRs shows that, although of small magnitude, eMagLS improves the spectral accuracy at both frontal and lateral directions over the EQ off condition. Therefore, changes in the perceived spatial compo-

sition and source externalization may have influenced the rendering quality of SMAs and lateral source directions.

4.6 Influence of the Source Incidence Direction

The authors investigated the influence of the nominal direction of direct sound incidence (frontal or lateral) on the perceived reproduction quality. Recall that this is implemented by offsetting the listener's orientation in the room. None of the comparisons for combinations of array type, array order, and equalization were significant.

Noteworthy is that the high-accuracy configuration is selected much more consistently as the highest-quality condition for lateral (median 7.1, 7.4, 7.8) compared with frontal (median 5.6, 5.8, 6.8) directions for trials involving SMAs, EMAs, and XMAs, respectively, as already discussed in SEC. 4.3. This is observed by fewer outliers with low ratings for SH44 in Fig. 4 resulting in narrower credible intervals compared with frontal directions in Fig. 5. Consequently, lower-order SMA, EMA, and XMA renderings with EQ on are rated insignificantly lower for lateral directions. The latter effect is similarly insignificant but varies in direction for renderings without equalization. SH1 as the low-accuracy anchor is not influenced by the prevalent direction of sound incidence.

Overall, the results indicate that the reproduction quality from order-limited microphone arrays is more impaired for lateral source directions due to truncation effects of the HRTF in the SH domain (see SEC. 0). Thereby, previous studies involving comparisons against an explicit reference have shown an apparent degradation in the perceived quality of order-limited SMAs from lateral incidence [10, 11, 28]. Here, this effect is small and statistically insignificant from the MuSCR paradigm. Instead, the high-accuracy condition (SH44), which would serve as the *hidden reference* in a MUSHRA paradigm, is more consistently assigned the highest quality.

5 CONCLUSION

The authors evaluated the quality of the binaural reproduction of signals from baffled microphone arrays. All resources are publicly available, including the utilized array impulse responses,³ the rendering scripts,¹ the resulting BRIRs, the tools to recreate the listening experiment, and the statistical analysis scripts.²

This experiment confirmed the observations made in previous studies that comprised an explicit reference stimulus, which validates the MuSCR paradigm that was employed for the first time in this context. MuSCR is more practical than the previous MUSHRA-like paradigms because it does not require a high-fidelity reference against which the conditions under test are evaluated. Such a reference stimulus is often not available.

This experiment comprised a stimulus with a numerical accuracy significantly higher than what can be achieved in practice. This condition was rated highest in quality in most cases (40%–55%, depending on the investigated array type). This provides new evidence for the assumption made

in previous studies that high accuracy is closely related to high perceived reproduction quality, even if no reference is available. However, depending on array type in cumulatively 25% to 5% of the trials, that same stimulus also received any of the lower (half) ranks. This affirms that rating the quality without an explicit reference is a highly complex task. Thereby, listeners must resort to an inner reference, which may lead to response biases, the increased influence of the acoustic scenario and program material, and, consequently, larger variances in the subjective ratings.

This experiment did not comprise a direct comparison of array types. This information has to be deduced from comparison across trials, which comes with some uncertainty but may be tolerable given that trials comprised identical low and high anchors. This data suggest that all array types (SMA, EMA, XMA) provide a comparable reproduction quality for a given SH order if state-of-the-art equalization (EQ on) of the binaural signals is applied. This is remarkable because EMAs and XMAs require significantly fewer microphones than SMAs and can be deployed much easier because of their equatorial layout. Improvements in reproduction quality from high-order arrays (in this experiment up to SH8) yield diminishing returns from frontal sources but can be significant for lateral directions of nominal sound incidence.

Future work covers the direct comparison of array types for different program materials containing speech and more reverberant acoustic environments. This also includes the evaluation of EMAs and XMAs for nonequatorial incidence of the primary source where their properties depart more from those of SMAs than for equatorial sound incidence. Furthermore, evaluating perceptual attributes beyond the general quality (e.g., sound localization and perceived source distance) may be relevant to future applications.

6 ACKNOWLEDGMENT

The authors thank Reality Labs by Meta for funding the research. Thanks to David Lou Alon for providing expertise and tools during the design and implementation of the project. The authors thank all subjects for participating in the listening study and the editor and reviewers for their helpful comments and suggestions. The authors declare that the research was conducted without any commercial or financial relationships that could be construed as a potential conflict of interest.

7 REFERENCES

- [1] N. A. Gumerov and R. Duraiswami, *Fast Multipole Methods for the Helmholtz Equation in Three Dimensions*, Elsevier Series in Electromagnetism (Elsevier, Amsterdam, The Netherlands, 2005). <https://doi.org/10.1016/B978-0-08-044371-3.X5000-5>.
- [2] B. Rafaely, "Analysis and Design of Spherical Microphone Arrays," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 135–143 (2005 Jan.). <https://doi.org/10.1109/TSA.2004.839244>.
- [3] J. Ahrens, H. Helmholtz, D. L. Alon, and S. V. Amengual Garí, "Spherical Harmonic Decomposition of a Sound Field Based on Observations Along the Equator of a Rigid Spherical Scatterer," *J. Acoust. Soc. Am.*, vol. 150, no. 2, pp. 805–815 (2021 Aug.). <https://doi.org/10.1121/10.0005754>.
- [4] J. Ahrens, H. Helmholtz, D. L. Alon, and S. V. Amengual Garí, "Spherical Harmonic Decomposition of a Sound Field Using Microphones on a Circumferential Contour Around a Non-Spherical Baffle," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 3110–3119 (2022 Sep.). <https://doi.org/10.1109/TASLP.2022.3209940>.
- [5] H. Helmholtz, T. Deppisch, and J. Ahrens, "End-to-End Magnitude Least Squares Binaural Rendering for Equatorial Microphone Arrays," in *Proceedings of the Fortschritte der Akustik (DAGA)*, pp. 1679–1682 (Hamburg, Germany) (2023 Mar.).
- [6] Z. Ben-Hur, F. Brinkmann, J. Sheaffer, S. Weinzierl, and B. Rafaely, "Spectral Equalization in Binaural Signals Represented by Order-Truncated Spherical Harmonics," *J. Acoust. Soc. Am.*, vol. 141, no. 6, pp. 4087–4096 (2017 Jun.). <https://doi.org/10.1121/1.4983652>.
- [7] B. Bernschütz, *Microphone Arrays and Sound Field Decomposition for Dynamic Binaural Recording*, Ph.D. thesis, Technische Universität Berlin, Berlin, Germany (2016 Feb.). <https://doi.org/10.14279/depositonce-5082>.
- [8] B. Rafaely, B. Weiss, and E. Bachmat, "Spatial Aliasing in Spherical Microphone Arrays," *IEEE Trans. Signal Process.*, vol. 55, no. 3, pp. 1003–1010 (2007 Mar.). <https://doi.org/10.1109/TSP.2006.888896>.
- [9] T. Lübeck, J. M. Arend, and C. Pörschmann, "Spatial Upsampling of Sparse Spherical Microphone Array Signals," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 1163–1174 (2023 Mar.). <https://doi.org/10.1109/TASLP.2023.3250841>.
- [10] M. Zaunschirm, 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.*, vol. 143, no. 6, pp. 3616–3627 (2018 Jun.). <https://doi.org/10.1121/1.5040489>.
- [11] J. Ahrens and C. Andersson, "Perceptual Evaluation of Headphone Auralization of Rooms Captured With Spherical Microphone Arrays With Respect to Spaciousness and Timbre," *J. Acoust. Soc. Am.*, vol. 145, no. 4, pp. 2783–2794 (2019 Apr.). <https://doi.org/10.1121/1.5096164>.
- [12] T. Lübeck, H. Helmholtz, J. M. Arend, C. Pörschmann, and J. Ahrens, "Perceptual Evaluation of Mitigation Approaches of Impairments due to Spatial Under-sampling in Binaural Rendering of Spherical Microphone Array Data," *J. Audio Eng. Soc.*, vol. 68, no. 6, pp. 428–440 (2020 Jun.). <https://doi.org/10.17743/jaes.2020.0038>.
- [13] International Telecommunication Union, "Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems," *Recommendation ITU-R BS.1534-3* (2015 Oct.).
- [14] T. Robotham, O. S. Rummukainen, M. Kurz, M. Eckert, and E. A. Habets, "Comparing Direct and Indirect Methods of Audio Quality Evaluation in Virtual Reality Scenes of Varying Complexity," *IEEE Trans. Vis. Com-*

put. Graph., vol. 28, no. 5, pp. 2091–2101 (2022 May). <https://doi.org/10.1109/TVCG.2022.3150491>.

[15] International Telecommunication Union, “Method for the Subjective Quality Assessment of Audible Differences of Sound Systems Using Multiple Stimuli Without a Given Reference,” *Recommendation ITU-R BS.2132-0* (2019 Oct.).

[16] T. Deppisch, H. Helmholtz, and J. Ahrens, “End-to-End Magnitude Least Squares Binaural Rendering of Spherical Microphone Array Signals,” in *Proceedings of the Immersive and 3D Audio: From Architecture to Automotive (I3DA)*, pp. 173–179 (Bologna, Italy) (2021 Sep.). <https://doi.org/10.1109/I3DA48870.2021.9610864>.

[17] J. Ahrens and Z. Hu, “Evaluation of Non-Spherical Scattering Bodies for Ambisonic Microphone Arrays,” in *Proceedings of the AES International Conference on Audio for Virtual and Augmented Reality* (2022 Aug.), paper 7.

[18] B. Rafaely and A. Avni, “Interaural Cross Correlation in a Sound Field Represented by Spherical Harmonics,” *J. Acoust. Soc. Am.*, vol. 127, no. 2, pp. 823–828 (2010 Feb.). <https://doi.org/10.1121/1.3278605>.

[19] B. Bernschütz, A. Vázquez Giner, C. Pörschmann, and J. M. Arend, “Binaural Reproduction of Plane Waves With Reduced Modal Order,” *Acta Acust. united Acust.*, vol. 100, no. 5, pp. 972–983 (2014 Sep./Oct.). <https://doi.org/10.3813/AAA.918777>.

[20] J. Ahrens, “Binaural Audio Rendering in the Spherical Harmonic Domain: A Summary of the Mathematics and its Pitfalls,” *arXiv preprint arXiv:2202.04393* (2022 Sep.). <https://doi.org/10.48550/arXiv.2202.04393>.

[21] A. Politis, S. Tervo, and V. Pulkki, “COMPASS: Coding and Multidirectional Parameterization of Ambisonic Sound Scenes,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6802–6806 (Calgary, Canada) (2018 Apr.). <https://doi.org/10.1109/ICASSP.2018.8462608>.

[22] L. McCormack, A. Politis, R. Gonzalez, T. Lokki, and V. Pulkki, “Parametric Ambisonic Encoding of Arbitrary Microphone Arrays,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 2062–2075 (2022 Jun.). <https://doi.org/10.1109/TASLP.2022.3182857>.

[23] S. Moreau, J. Daniel, and S. Bertet, “3D Sound Field Recording With Higher Order Ambisonics – Objective Measurements and Validation of Spherical Microphone,” presented at the *120th Convention of the Audio Engineering Society* (2006 May), paper 6857.

[24] P. Stade, B. Bernschütz, and M. Rühl, “A Spatial Audio Impulse Response Compilation Captured at the WDR Broadcast Studios,” in *Proceedings of the 27th Tonmeister-tagung – VDT International Convention*, pp. 551–567 (Cologne, Germany) (2012 Nov.).

[25] F. Denk, B. Kollmeier, and S. D. Ewert, “Removing Reflections in Semianechoic Impulse Responses by Frequency-Dependent Truncation,” *J. Audio Eng. Soc.*, vol. 66, no. 3, pp. 146–153 (2018 Mar.). <https://doi.org/10.17743/jaes.2018.0002>.

[26] H. Braren and J. Fels, “A High-Resolution Head-Related Transfer Function Data Set and 3D-Scan of KEMAR,” *Tech. Rep. 11307* (2020 Nov.). <https://doi.org/10.18154/RWTH-2020-11307>.

[27] M. Geier, J. Ahrens, and S. Spors, “The SoundScape Renderer: A Unified Spatial Audio Reproduction Framework for Arbitrary Rendering Methods,” presented at the *124th Convention of the Audio Engineering Society* (2008 May), paper 7330.

[28] T. Lübeck, J. M. Arend, and C. Pörschmann, “Binaural Reproduction of Dummy Head and Spherical Microphone Array Data—A Perceptual Study on the Minimum Required Spatial Resolution,” *J. Acoust. Soc. Am.*, vol. 151, no. 1, pp. 467–483 (2022 Jan.). <https://doi.org/10.1121/10.0009277>.

[29] A. Lindau and F. Brinkmann, “Perceptual Evaluation of Individual Headphone Compensation in Binaural Synthesis Based on Non-Individual Recordings,” in *Proceedings of the 3rd International Workshop on Perceptual Quality of Systems (PQS)*, pp. 126–131 (Bautzen, Germany) (2010 Sep.). <https://doi.org/10.21437/PQS.2010-23>.

[30] J. Gomez-Bolaños, A. Mäkilvirta, and V. Pulkki, “Automatic Regularization Parameter for Headphone Transfer Function Inversion,” *J. Audio Eng. Soc.*, vol. 64, no. 10, pp. 752–761 (2016 Oct.). <https://doi.org/10.17743/jaes.2016.0030>.

[31] B. Carpenter, A. Gelman, M. D. Hoffman, et al., “Stan: A Probabilistic Programming Language,” *J. Stat. Softw.*, vol. 76, no. 1, pp. 1–32 (2017 Jan.). <https://doi.org/10.18637/jss.v076.i01>.

[32] R Core Team, “R: A Language and Environment for Statistical Computing,” <https://www.r-project.org/> (accessed Aug. 6, 2024).

[33] J. K. Kruschke and T. M. Liddell, “The Bayesian New Statistics: Hypothesis Testing, Estimation, Meta-Analysis, and Power Analysis From a Bayesian Perspective,” *Psychon. Bull. Rev.*, vol. 25, no. 1, pp. 178–206 (2018 Feb.). <https://doi.org/10.3758/s13423-016-1221-4>.

[34] A. Vehtari, A. Gelman, and J. Gabry, “Practical Bayesian Model Evaluation Using Leave-One-Out Cross-Validation and WAIC,” *Stat. Comput.*, vol. 27, no. 5, pp. 1413–1432 (2017 Sep.). <https://doi.org/10.1007/s11222-016-9696-4>.

[35] A. Gelman, J. B. Carlin, H. S. Stern, et al., *Bayesian Data Analysis* (Chapman and Hall/CRC, New York, NY, 2013), 3rd ed. <https://doi.org/10.1201/b16018>.

[36] S. Zielinski, F. Rumsey, and S. Bech, “On Some Biases Encountered in Modern Listening Tests,” *J. Audio Eng. Soc.*, vol. 56, no. 6, pp. 427–451 (2008 Jun.).

A.1 STATISTICAL MODEL

Here is the hierarchical generalized linear regression (cumulative) model introduced in SEC. 3.4 and applied in SEC. 4. The model employs the resulting ranks including subject rating, array order (SH1, SH2, SH4, SH8, SH44), equalization (EQ off, EQ on), array type (SMA, EMA, XMA), source incidence direction (frontal, lateral), and interaction terms ($\times$) as population-level effects and effects for subject, trial, and slope of individual ratings as group-level (varying) effects. To account for possible variance differences between experience levels, the authors estimated variance as a function of reported experience (Exp), time spent on the task (Time), and the interaction term ($\times$).

$$p(y = k | \mu, \sigma, \{\theta_1, \dots, \theta_{K-1}\}) = (\theta_k - \mu) / \sigma - (\theta_{k-1} - \mu) / \sigma \quad (1)$$

$$\begin{aligned} \text{logit}(\mu_i) = & \alpha_{\text{Subject}_i} + \alpha_{\text{Trial}_i} + \beta_{1\text{Subject}} \text{SH2}_i + \beta_{2\text{Subject}} \text{SH4}_i + \beta_{3\text{Subject}} \text{SH8}_i + \beta_{4\text{Subject}} \text{SH44}_i \\ & + \beta_{5\text{Subject}} \text{EMA}_i + \beta_{6\text{Subject}} \text{XMA}_i + \beta_7 \text{Rating}_i + \beta_8 \text{EQon}_i + \beta_9 \text{Direction}_{\text{lateral},i} \\ & + \beta_{10} \text{Exp}_{<1\text{yr},i} + \beta_{11} \text{Exp}_{1:3\text{yr},i} + \beta_{12} \text{Exp}_{3:5\text{yr},i} + \beta_{13} \text{Exp}_{>5\text{yr},i} \\ & + \beta_{11}(\text{SH2}_i \times \text{EQon}_i) + \beta_{12}(\text{SH4}_i \times \text{EQon}_i) + \beta_{13}(\text{SH8}_i \times \text{EQon}_i) + \beta_{14}(\text{SH44}_i \times \text{EQon}_i) \\ & + \beta_{15}(\text{EQon}_i \times \text{EMA}_i) + \beta_{16}(\text{EQon}_i \times \text{XMA}_i) \\ & + \beta_{17}(\text{SH2}_i \times \text{EMA}_i) + \beta_{18}(\text{SH4}_i \times \text{EMA}_i) + \beta_{19}(\text{SH8}_i \times \text{EMA}_i) + \beta_{20}(\text{SH44}_i \times \text{EMA}_i) \\ & + \beta_{21}(\text{SH2}_i \times \text{XMA}_i) + \beta_{22}(\text{SH4}_i \times \text{XMA}_i) + \beta_{23}(\text{SH8}_i \times \text{XMA}_i) + \beta_{24}(\text{SH44}_i \times \text{XMA}_i) \\ & + \beta_{25}(\text{SH2}_i \times \text{Direction}_{\text{lateral},i}) + \beta_{26}(\text{SH4}_i \times \text{Direction}_{\text{lateral},i}) \\ & + \beta_{27}(\text{SH8}_i \times \text{Direction}_{\text{lateral},i}) + \beta_{28}(\text{SH44}_i \times \text{Direction}_{\text{lateral},i}) + \beta_{29}(\text{EQon}_i \times \text{Direction}_{\text{lateral},i}) \\ & + \beta_{30}(\text{SH2}_i \times \text{Exp}_{<1\text{yr},i}) + \beta_{31}(\text{SH2}_i \times \text{Exp}_{1:3\text{yr},i}) + \beta_{32}(\text{SH2}_i \times \text{Exp}_{3:5\text{yr},i}) + \beta_{33}(\text{SH2}_i \times \text{Exp}_{>5\text{yr},i}) \\ & + \beta_{34}(\text{SH4}_i \times \text{Exp}_{<1\text{yr},i}) + \beta_{35}(\text{SH4}_i \times \text{Exp}_{1:3\text{yr},i}) + \beta_{36}(\text{SH4}_i \times \text{Exp}_{3:5\text{yr},i}) + \beta_{37}(\text{SH4}_i \times \text{Exp}_{>5\text{yr},i}) \\ & + \beta_{38}(\text{SH8}_i \times \text{Exp}_{<1\text{yr},i}) + \beta_{39}(\text{SH8}_i \times \text{Exp}_{1:3\text{yr},i}) + \beta_{40}(\text{SH8}_i \times \text{Exp}_{3:5\text{yr},i}) + \beta_{41}(\text{SH8}_i \times \text{Exp}_{>5\text{yr},i}) \\ & + \beta_{42}(\text{SH44}_i \times \text{Exp}_{<1\text{yr},i}) + \beta_{43}(\text{SH44}_i \times \text{Exp}_{1:3\text{yr},i}) + \beta_{44}(\text{SH44}_i \times \text{Exp}_{3:5\text{yr},i}) + \beta_{45}(\text{SH44}_i \times \text{Exp}_{>5\text{yr},i}) \\ & + \beta_{46}(\text{EQon}_i \times \text{Exp}_{<1\text{yr},i}) + \beta_{47}(\text{EQon}_i \times \text{Exp}_{1:3\text{yr},i}) + \beta_{48}(\text{EQon}_i \times \text{Exp}_{3:5\text{yr},i}) + \beta_{49}(\text{EQon}_i \times \text{Exp}_{>5\text{yr},i}) \\ & + \beta_{50}(\text{Direction}_{\text{lateral},i} \times \text{Exp}_{<1\text{yr},i}) + \beta_{51}(\text{Direction}_{\text{lateral},i} \times \text{Exp}_{1:3\text{yr},i}) \\ & + \beta_{52}(\text{Direction}_{\text{lateral},i} \times \text{Exp}_{3:5\text{yr},i}) + \beta_{53}(\text{Direction}_{\text{lateral},i} \times \text{Exp}_{>5\text{yr},i}) \end{aligned} \quad (2)$$

$$\alpha_{\text{Subject}} = \sigma_{\text{Subject}} * \zeta_{\text{Subject}_i} \quad \sigma_{\text{Subject}} \sim \text{Normal}(0, 1) \quad \zeta_{\text{Subject}} \sim \text{Normal}(0, 1) \quad (3)$$

$$\alpha_{\text{Trial}} = \sigma_{\text{Trial}} * \zeta_{\text{Trial}_i} \quad \sigma_{\text{Trial}} \sim \text{Normal}(0, 1) \quad \zeta_{\text{Trial}} \sim \text{Normal}(0, 1) \quad (4)$$

$$\beta_1, \dots, \beta_{53} \sim \text{Normal}(0, 1) \quad \theta_k \sim \text{Normal}(0, 5) \quad \sigma = 1/\text{disc} \quad (5)$$

$$\begin{aligned} \log(\text{disc}) = & \gamma_1 \text{Exp}_{<1\text{yr},i} + \gamma_2 \text{Exp}_{1:3\text{yr},i} + \gamma_3 \text{Exp}_{3:5\text{yr},i} + \gamma_4 \text{Exp}_{>5\text{yr},i} + \gamma_5 \text{Time} \\ & + \gamma_6(\text{Time}_i \times \text{Exp}_{<1\text{yr},i}) + \gamma_7(\text{Time}_i \times \text{Exp}_{1:3\text{yr},i}) + \gamma_8(\text{Time}_i \times \text{Exp}_{3:5\text{yr},i}) + \gamma_9(\text{Time}_i \times \text{Exp}_{>5\text{yr},i}) \end{aligned} \quad (6)$$

$$\gamma_{1..9} \sim \text{Normal}(0, 1) \quad (7)$$

$$\begin{bmatrix} \alpha_{\text{Subject}} \\ \beta_{1\text{Subject}} \\ \beta_{2\text{Subject}} \\ \beta_{3\text{Subject}} \\ \beta_{4\text{Subject}} \\ \beta_{5\text{Subject}} \\ \beta_{6\text{Subject}} \end{bmatrix} \sim \text{MVNormal} \left(\begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \end{bmatrix}, \sum_{\text{Subject}} \right) \quad (8)$$

$$\Sigma_{\text{Subject}} = \begin{pmatrix} \sigma_{\alpha} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma_{\beta_1} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_{\beta_2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_{\beta_3} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_{\beta_4} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma_{\beta_5} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \sigma_{\beta_6} \end{pmatrix} \mathbf{R}_{\text{Subject}} \begin{pmatrix} \sigma_{\alpha} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma_{\beta_1} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_{\beta_2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_{\beta_3} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_{\beta_4} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma_{\beta_5} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \sigma_{\beta_6} \end{pmatrix} \quad (9)$$

$$\mathbf{R}_{\text{Subject}} \sim \text{LKJcorr}(1) \quad (10)$$

THE AUTHORS



Hannes Helmholz



Jeffery Crukley



Sebastià V. Amengual Garí



Zamir Ben-Hur



Jens Ahrens

Hannes Helmholz is currently a doctoral candidate for the Division of Applied Acoustics at Chalmers University of Technology, working on the spatial capture and reproduction of existing acoustic environments with microphone arrays. He received his B.Sc. degree in Applied Computer Science from the University of Applied Sciences, Berlin, Germany, in 2013 and his M.Sc. degree in Audio Communication and Technology from Technische Universität Berlin, Berlin, Germany, in 2018. He has been a Research Intern with Meta Reality Labs Research, Redmond, WA, since 2021. His research interests lie in audio signal processing, binaural reproduction, and auditory perception of spatial sound fields.

Jeff Crukley is currently an applied statistician in collaboration with scientists/researchers across academic, industrial, and government contexts after spending a decade as an industry scientist and university instructor in the field of communication sciences and disorders. He has also mentored graduate students through appointments at the University of Toronto since 2014 and McMaster University since 2017. He received his B.Sc. degree in Biology and Psychology from McMaster University, Hamilton, ON, Canada, in 2000, and he received his M.Sc. degree in Communication Sciences and Disorders, Ph.D. degree in Hearing Science, and a post-doctoral fellowship in Medical Biophysics and Neuroscience from Western University, London, ON, Canada, in 2007, 2011, and 2013, respectively.

Sebastià Amengual is currently a Research Scientist at Meta Reality Labs Research, working on room acoustics, spatial audio, and auditory perception. He received his Diploma in Telecommunications with a major in Sound and Image from the Polytechnic University of Catalonia (UPC), Barcelona, Spain, in 2014. He completed his Mas-

ter's thesis at the Norwegian University of Science and Technology (NTNU), Trondheim, Norway. He received his Doctoral degree (Dr.-Ing.) from the University of Music, Detmold, Germany, in 2017, focusing on the interaction of room acoustics and live music performance using virtual acoustic environments. His research interests lie in the intersection of audio, perception, and music.

Zamir Ben-Hur is currently a Research Scientist at Meta Reality Labs Research, working on spatial audio technologies. He received his B.Sc. (summa cum laude), M.Sc., and Ph.D. degrees in Electrical and Computer Engineering from Ben-Gurion University of the Negev, Beer-Sheva, Israel, in 2015, 2017, and 2020, respectively. He was a Research Intern with Meta Reality Labs Research, Redmond, WA, in 2016, 2017, and 2019. His research interests include spatial audio signal processing for binaural reproduction with improved spatial perception.

Jens Ahrens is currently a Professor within the Division of Applied Acoustics at Chalmers University of Technology. His primary interests are tools and methods for auralization of real and virtual acoustic environments. He received his Diplom (equivalent to M.Sc.) in electrical engineering/sound engineering jointly from Graz University of Technology and the University of Music and Dramatic Arts Graz, Graz, Austria, in 2005, and his Doctoral degree (Dr.-Ing.) from Technische Universität Berlin, Berlin, Germany, in 2010. From 2011 to 2013, Jens was a Postdoctoral Researcher with Microsoft Research, Redmond, WA, and a Visiting Scholar with the Center for Computer Research in Music and Acoustics at Stanford University, CA, from 2015 to 2016. He conducts research visits to research organizations in academia and in the industry on a regular basis.