Audio Engineering Society

Conference Paper 7

Presented at the AES 5th International Conference on Audio for Virtual and Augmented Reality
2024 August 19–21, Redmond, WA, USA

This paper was peer-reviewed as a complete manuscript for presentation at this conference. This paper is available in the AES E-Library (http://www.aes.org/e-lib), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Binaural reproduction of head-worn microphone array recordings with adjustable field-of-view control

Janani Fernandez[1], David Lou Alon[1], Zamir Ben-hur[1], and Vladimir Tourbabin[1]

[1]*Reality Labs Research, Meta, Redmond, WA, USA*

Correspondence should be addressed to Janani Fernandez (jananifernandez@gmail.com)

ABSTRACT

This paper investigates an approach for reproducing head-worn microphone array recordings over headphones, such that the listener is also able to augment the rendering to emphasise sounds arriving within a particular field-of-view (FoV), while also attempting to preserve the spatial properties of the captured scene. This type of processing may find application within future augmented reality contexts. The directional emphasis is realised by applying an additional direction-dependent weighting term, when conducting the magnitude least-squares fitting of the array directivities to the binaural directivities. The proposed approach is presented alongside perceptual metric analysis and evaluated via a perceptual study involving 20 listeners. The results suggest that achieving a gain within a defined FoV is attainable, but there exists a trade-off between increasing gain and negatively impacting the spatial aspects of the reproduced sound scene.

1 Introduction

Due to the growing adoption of head-worn microphone arrays within the consumer space, increased attention has been given to the task of sound-field capture and binaural reproduction. Here, a user may spatially record sound scenes from their perspective, using a microphone array integrated into a device, such as a pair of smartglasses, and share or relive this auditory experience binaurally over headphones. Additionally, it may be desirable for the user to be able to manipulate the rendering. For example, one may wish to warp, distort, and/or rotate the spatial properties of the sound scene, in order to offer support for listener movements, such as rotations and translations during playback (i.e., 3DoF/6DoF), or to realise other creative spatial audio effects [1, 2, 3]. One may also attempt to preserve the original spatial characteristics of the scene, while manipulating the relative direction-dependent energy-distribution within the scene, resulting in sound sources located at specific directions being emphasised or attenuated [4, 5, 6, 7]. The focus of this paper is on this latter task, within the context of using smartglasses to audio-visually record the scene. Specifically, it concerns the emphasis of sounds arriving at the array within a specified field-of-view (FoV) (i.e., within a *cone* matching the camera zoom), while attempting to leave the spatial aspects of the scene largely unchanged.

Traditionally, flexible capture and reproduction of spatial sound scenes has involved the use of uniform spherical microphone arrays (SMAs), which can capture sound-fields with equal spatial resolution over the sphere. A popular rendering framework, espe-

cially within the context of augmented/virtual reality (AR/VR) and immersive media, is Ambisonics [8, 9], which involves the transformation of the array signals into the spherical harmonic domain, and then mapping these spherical harmonic (SH) signals to the target playback setup. Sound-field modifications may be realised by manipulating the SH signals in this domain in-between these two stages, and may include applying: sound-field warping functions [10, 2], directional-loudness modifications [2], rotations [11] and translations [12, 13, 3].

Limitations of the Ambisonics format, however, become strongly evident within the context of head-worn and other irregularly shaped arrays, such as those integrated into mobile phones and 360 degree cameras (where spatial audio capture is not their only or primary function). This is because, due to their irregular geometry and/or non-uniform sensor placements, SH signals derived from such arrays are typically contaminated by spatial aliasing for a sizeable portion of the perceivable frequency range [14, 15], which also varies depending on the source direction. Additionally, the maximum order achievable through a conventional linear encoder, using such arrays, is usually far below what is required in order to mitigate perceptual errors [16, 17, 18]; owing to the limited number of microphones. In order to circumvent this problematic conversion, and to better leverage the full spatial resolution of this limited capture format; some researchers have recently turned their attention to the directly rendering the head-worn microphone array signals into binaural signals.

Options for direct binaural reproduction of head-worn array capture include plane-wave decomposition (PWD) followed by binauralisation (also known as beamformer-based binaural reproduction (BFBR) [19]); conducting a least-squares (LS) fitting (also known as binaural signal matching (BSM) [20]); or employing perceptually-motivated optimisations of the LS fitting, such as magnitude-least-squares (MagLS). These approaches have also recently been extended to support listener translations and/or rotations [21, 22]. However, to the best of the authors' knowledge, there does not currently exist a report of a study where these smartglasses-to-binaural rendering approaches are modified to accommodate FoV enhancement/emphasis. There is one potential exceoption to this [5], which may be considered, but this study utilised the parametric rendering framework of [23], and may therefore not be suitable for low-power devices, which include many of the head-worn devices available today. Additionally, while similar processing has been explored within hearing aid contexts, the focus there has almost entirely been on improving speech intelligibility without attempting to preserve the spatial properties of the scene, with only few exceptions [6, 7].

Therefore, in the work described within this paper, the MagLS approach was selected and modified to achieve FoV enhancement, which is realised by introducing a direction dependent weighting matrix into its formulation. Objective analysis and a subjective evaluation are then performed, based on a 5-sensor microphone array integrated into a pair of smartglasses (and worn by a manikin during recording), in order to investigate the feasibility of the proposed approach.

2 Method

Consider Q microphones being used to record array signals $\mathbf{x}(t,f) \in \mathbb{C}^{Q\times 1}$, which are represented in the short-time Fourier transform (STFT)-domain, indexed with time-frequency indices t, f. We assume that the sound-field can be described using a large number of V plane-wave signals, $\mathbf{s}(t,f) = [s(\Omega_1,t,f),...,s(\Omega_V,t,f)]^{\mathrm{T}} \in \mathbb{C}^{V\times 1}$, which are impinging on the array from different directions Ω_i (with $i = [1,2,...,V]$), as

$$\mathbf{x}(t,f) = \mathbf{A}(f)\mathbf{s}(t,f), \tag{1}$$

where $\mathbf{A}(f) = [\mathbf{a}(\Omega_1,f),...,\mathbf{a}(\Omega_V,f)]^{\mathrm{T}} \in \mathbb{C}^{Q\times V}$ are the array transfer functions (ATFs), corresponding to those same V directions, which may be obtained through free-field measurements or simulations of the array.

These recorded microphone array signals may be linearly mapped to the binaural channels, $\mathbf{y}(t,f) \in \mathbb{C}^{2\times 1}$, and relayed to the listener reliving the auditory experience via a pair of headphones, with the following mixing matrix $\mathbf{M}(f) \in \mathbb{C}^{2\times Q}$

$$\mathbf{y}(t,f) = \mathbf{M}(f)\mathbf{x}(t,f). \tag{2}$$

There are then a number of different methods for computing this linear mixing matrix, including: constructing it by first beamforming in multiple directions, followed by convolving the resulting signals with HRTFs for those same directions [24, 19]; a combination of an Ambisonic encoder [25] and a subsequent binaural decoder [26]; or a direct least-squares (LS) fitting [20]

of the array directivities (described by the ATFs), to the binaural directivities (described by HRTFs). In this work, however, the sensor-domain magnitude-least-squares (MagLS) approach [27, 28], was selected which may be formulated as [21]

$$\mathbf{M}(f) = \mathbf{H}^{(\text{mod})}(f)\mathbf{W}\mathbf{A}^{\text{H}}(f)\big(\mathbf{D}(f) + \lambda\mathbf{I}\big)^{-1}, \quad (3)$$

where $\mathbf{W} \in \mathbb{R}^{V\times V}$ is a diagonal matrix of integration weights (with $\mathbf{tr}[\mathbf{W}] = 1$) to account for cases where the measurement grid is not uniform, $\lambda = 0.01$ is a regularisation term, $\mathbf{D}(f) = \mathbf{A}(f)\mathbf{W}\mathbf{A}^{\text{H}}(f)$ is the diffuse coherence matrix [29], and $\mathbf{H}^{(\text{mod})}$ are phase modified HRTFs. Note that this phase modification is a perceptually-motivated optimisation [26], inspired by the Duplex Theory, whereby inter-aural level differences are deemed to be perceptually more important than phase differences at higher frequencies; hence motivating prioritising a better fit to HRTF magnitudes, rather than HRTF phases, at these higher frequencies. In this work, we adopt the algorithm proposed in [9], which defines $\mathbf{H}^{(\text{mod})}$ as

$$\mathbf{H}^{(\text{mod})}(f) = \begin{cases} \mathbf{H}(f) & f \le f_c \\ |\mathbf{H}(f)|e^{i\phi(f)} & f > f_c \end{cases} \quad (4)$$

where the cut-off frequency $f_c = 1500$ Hz dictates the point where the method attempts to transition from the standard LS solution to a MagLS solution, and $\phi = \arg[\mathbf{M}(f-1)\mathbf{A}(f-1)]$ denotes the phase response of the reconstructed HRTFs for the previous frequency.

Note that these frequency-dependent mixing matrices may be applied as in Eq. (2) in the STFT domain, or they can be transformed into a matrix of filters (via an inverse Fourier transform), and applied via a matrix convolution in the time-domain. The end result will be the same in either case, but depending on the length of the filters (or frequency resolution adopted), their computational requirements can differ. However, due to the linear nature of the mapping, either approach should be less computationally demanding compared to the application direction-dependent gains within the parametric rendering framework described in [5].

2.1 FoV control

The baseline MagLS solution of Eq. (3) places equal gain for all directions on the unit sphere. In this paper, we propose to alter this MagLS solution, in order to allow the method to instead emphasise directions that are within predefined FoV limits. This alteration is motivated by a desire to ensure that the audio scene reproduction appropriately aligns with the intended focus of the recorded scene, particularly in situations wherein the listener chooses to apply a camera zoom during the video playback. The proposed FoV control is formulated as

$$\mathbf{M}_{\text{FoV}}(f) = \mathbf{H}^{(\text{mod})}(f)(\mathbf{W} + \gamma\mathbf{W}_{\mathbf{C}})\mathbf{A}^{\text{H}}(f)\big(\mathbf{D}(f) + \lambda\mathbf{I}\big)^{-1}, \quad (5)$$

where $\gamma \ge 0$ is a hyper-parameter whose value influences the degree to which a biasing matrix, $\mathbf{W}_C \in \mathbb{R}^{V\times V}$, is introduced into the solution. This biasing matrix is a diagonal matrix with elements of value $1/V$ if a measurement grid direction falls within the FoV cone, and is 0 otherwise. In this study, the FoV is defined as +/- 30 degrees around the front-facing direction on the horizontal plane.

To normalise the rendering matrix, to ensure that the overall output level remains consistent for different values of γ, the following normalisation was applied

$$\hat{\mathbf{M}}_{\text{FoV}} = \left(\frac{\mathbf{tr}[\mathbf{M}(f)\mathbf{D}(f)\mathbf{M}^{\text{H}}(f)]}{\mathbf{tr}[\mathbf{M}_{\text{FoV}}(f)\mathbf{D}(f)\mathbf{M}^{\text{H}}_{\text{FoV}}(f)]}\right)^{1/2}\mathbf{M}_{\text{FoV}}, \quad (6)$$

where $\mathbf{tr}[.]$ denotes the trace operator.

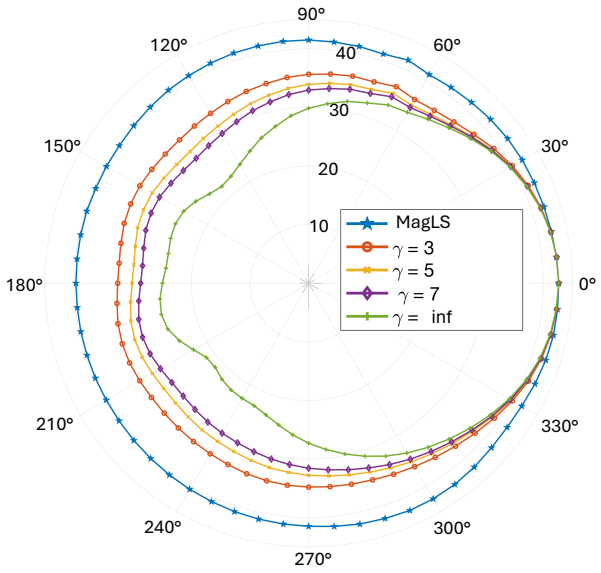
3 Objective analysis

The proposed FoV enhancement approach was first investigated by inspecting a number of objective metrics. The head-worn array featured 5 microphones embedded into the frames of a pair of smartglasses, whose ATFs were measured in an anechoic chamber for every one degree on the horizontal plane. The microphone array consisted of microphones located above each ear, on the mid point of each of the temples, and a fifth microphone located near to the nose bridge. These 360 measured ATFs were used when computing $\mathbf{M}$ and $\mathbf{M}_{\text{FoV}}$.

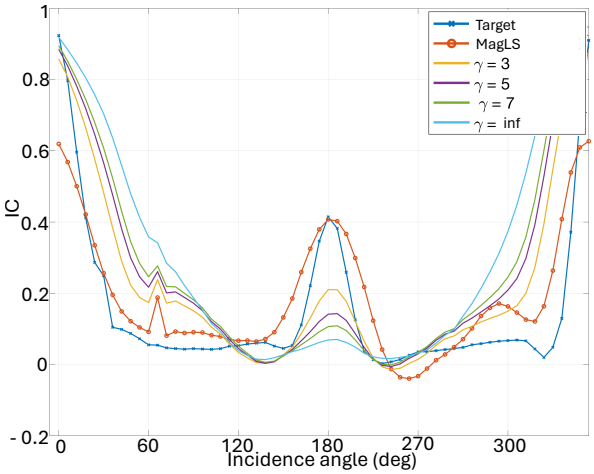
The objective metrics consisted of a combination of binaural and energy based values, which were calculated from the binaural spatial covariance matrices (SCMs) of the output binaural signals

$$\mathbf{C}_{\text{y}}(f) = \begin{pmatrix} c_{1,1}(f) & c_{1,2}(f) \\ c_{2,1}(f) & c_{2,2}(f) \end{pmatrix} = \mathbb{E}[\mathbf{y}(t,f)\,\mathbf{y}^{\text{H}}(t,f)], \quad (7)$$

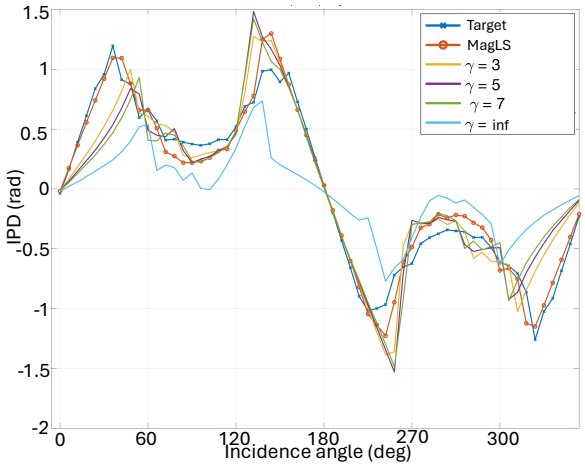
where $\mathbb{E}[.]$ is the expectation operator.



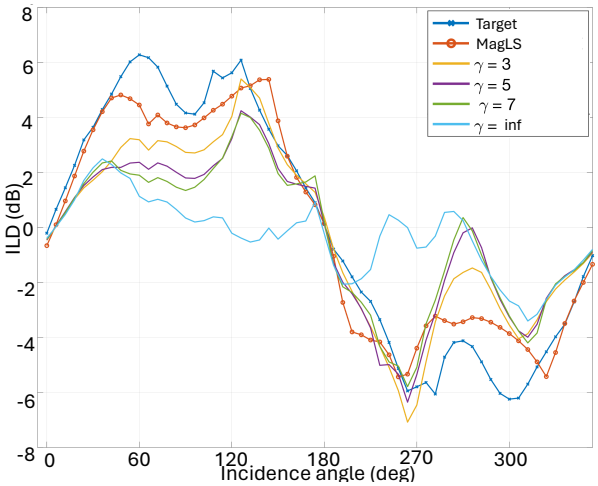
(a) Direction-dependent gain averaged between [0,3] kHz.



(b) IC averaged between [0,3] kHz.



(c) IPD averaged between [0,1.5] kHz.



(d) ILD averaged between [0,3] kHz.

Fig. 1: Computed objective metrics of the proposed algorithm.

Since the goal of this study was to apply a direction-dependent gain, the first objective metric was to compute the overall output gain as a function of source direction

$$G(\Omega, f) = 10log_{10}(\frac{1}{2}\mathbf{tr}[\tilde{\mathbf{C}}_{\mathrm{y}}(\Omega, f)]), \qquad (8)$$

where $\tilde{\mathbf{C}}_{\mathrm{y}}$ is the binaural SCM computed from the result of the algorithm rendering a single plane-wave, i.e., computed based on

$$\tilde{\mathbf{y}}(\Omega, t, f) = \mathbf{M}_{\mathrm{FoV}}(f)\mathbf{a}(\Omega, f)s(\Omega, t, f). \qquad (9)$$

In Figure 1(a), $G(\Omega, f)$ was averaged over frequencies between 0 Hz and 3 kHz (which was the approximate spatial aliasing frequency of the microphone array for the frontal direction), given $\gamma = 0$ (normal MagLS) and $\gamma = [3, 5, 7, \mathrm{Inf}]$ (increasing FoV effect), and presented as a polar plot with 5 degree resolution in the horizontal plane. Here, it is clearly evident that for directions

outside of the FoV, the proposed approach leads to more attenuation with increasing values of γ.

However, an additional goal of this study was to investigate how the perceived spatial properties of the scene are affected by this FoV control. Therefore, the remaining objective metrics of interest are the binaural spatial cues; namely the interaural level difference (ILD), the interaural coherence (IC), and the interaural phase difference (IPD). These were computed over frequency and incident direction with

$$ILD(\Omega, f) = 10log_{10}\frac{|\tilde{c}_{1,1}(\Omega, f)|}{|\tilde{c}_{2,2}(\Omega, f)|}, \quad (10)$$

$$IC(\Omega, f) = 10log_{10}\frac{real(\tilde{c}_{1,2}(\Omega, f))}{\sqrt{\tilde{c}_{1,1}(\Omega, f)\tilde{c}_{2,2}(\Omega, f)|}}, \quad (11)$$

$$IPD(\Omega, f) = arg(\tilde{c}_{1,2}(\Omega, f)). \quad (12)$$

These remaining objective metrics are plotted in Figures 1(b,c,d). Note that these metrics were also averaged between 0 and 3 kHz, with the exception of IPD, which was instead averaged between 0 and 1.5 kHz, since this is a more perceptually relevant range. These metrics were also computed based on $\gamma = [0, 3, 5, 7, \text{Inf}]$. In addition, a target/reference case was computed based on artificially introducing a 6 dB gain for directions that fall within the field of view. This was done by increasing the amplitude of these output signals by a factor of 2 before computing the metrics for them. This was deemed by the present authors to represent a realistic target to aim for.

Based on the metrics shown in Figures 1(b,c,d), there is some indication that the perceptual cues are also deviating from the original ones, with increasing values of gamma. The magnitude of the deviation was also dependent on the incident angle of the plane wave, as incidence angles outside the FoV seemed to have larger deviations. In Figure 1 (d) in particular, it can be seen that for incident angles within the FoV limits, the difference in the ILD values between the target and the proposed algorithm are near or below 1 dB, which is the average just noticeable difference for ILD values in persons with normal-hearing [30]. However, the perceptual implications of this can only be truly ascertained by conducting a subjective listening test, which is described in the following section.

4 Perceptual evaluation

In order to evaluate the perceived performance of the proposed FoV enhancement approach, a subjective multiple stimulus listening test was conducted.

The intention of this listening test was to determine how the spatial and timbral properties of the reproduced scene are impacted by the FoV enhancement. The authors postulated that increasing values of gamma, (which increasingly leads to the FoV enhancement), would also lead to a degradation in the perceptual aspects, and thus the test was conducted to see where such trade-offs may lay.

4.1 Assessors

20 expert listeners with normal hearing were recruited to be assessors through Force Technologies (Denmark). All assessor received monetary compensation for their participation in the perceptual evaluation.

4.2 Test scenes

Three sound scenes were recorded using the same pair of smartglasses described in Section 3. The first scene, termed “Pool Trick Shot", consisted of a person explaining their next shot in a game of pool. This scene had one sound source located within the FoV limits set (+/- 30 degrees around the front-facing direction), which moved within these limits during the recording. There were also other people and background noise present in the scene, outside the FoV. The scene was approximately 12 seconds in duration and was recorded in a cafeteria room. The second recorded sound scene had two sources and was labelled as “Music Interferer". The first source was a loudspeaker playing music, which was located directly in-front of the smartglasses wearer, while the second source was a person speaking, who was located 90 degrees to the left of the wearer. Note that the smartglasses wearer in this recording was a manikin mounted onto a programmable turntable. During the recording, the smartglasses wearer first faced the loudspeaker, but then turned their head 90 degrees to face the second source for approximately 12 seconds, before returning to face the loudspeaker again. This scene was a total of approximately 20 seconds in duration. The third sound scene was the same as the second scene, with the exception of the frontal loudspeaker playing a recording of someone speaking, as opposed to playing music. It was therefore

named as the “Speech Interferer" scene. It was also approximately 20 seconds in duration. These second and third scenes were both recorded in a living room environment.

4.3 Test cases

Each sound scene was rendered using the proposed algorithm with $\gamma = [1, 3, 7, \text{inf}]$. The proposed algorithm was applied up to the approximate spatial aliasing frequency of the headset for the frontal direction (3 kHz), after which regular MagLS fitting was applied (i.e., without FoV enhancement). Baseline, i.e., MagLS (the proposed algorithm with $\gamma = 0$), and target/reference renderings of the sound scenes were also made. The reference (Ref) was created by applying an artificial 6 dB gain to sources located within the set FoV limits in the recorded sound scene itself, similar to the target simulated for the objective analysis. This was possible because each sound source in each scene was recorded separately, which allowed for artificial manipulation of the scene in post processing/mixing. An anchor (Anchor) test case was also included, which varied depending upon the particular part of the perceptual evaluation being performed. These parts are described in the following subsection.

4.4 Test methodology

The listening test was divided into three parts: spatial, timbral, and overall. The order of the spatial and timbral tests was randomised. Whereas, the overall part was always presented last. Each trial had the 7 test cases present on the page, (with one page per sound scene), and assessors were tasked with rating the test cases based on how similar they were to the given reference/target rendering, with 100 being perceptually identical and 0 being perceptually very different. A training phase was included before each section in the listening test.

In the spatial part, the assessors were asked to focus on the spatial differences between the test cases and the reference. Each of the test cases were equalised to have the same timbral characteristics as the reference, following a similar test methodology to that used in previous studies [23, 15]. This equalisation was performed by computing the average binaural energy (over both channels and time) for each time-frequency tile of the reference, $E_{ref}(f)$, performing the same calculation for each test case, $E_{test}(f)$, then computing an equalisation curve, $eq(f) = \sqrt{\frac{E_{ref}(f)}{E_{test}(f)}}$, and the applying it to each test case, independently. The anchor used in this section was a monaural MagLS rendering of the respective sound scene, created by averaging the two binaural signals and routing that same signal to both output channels.

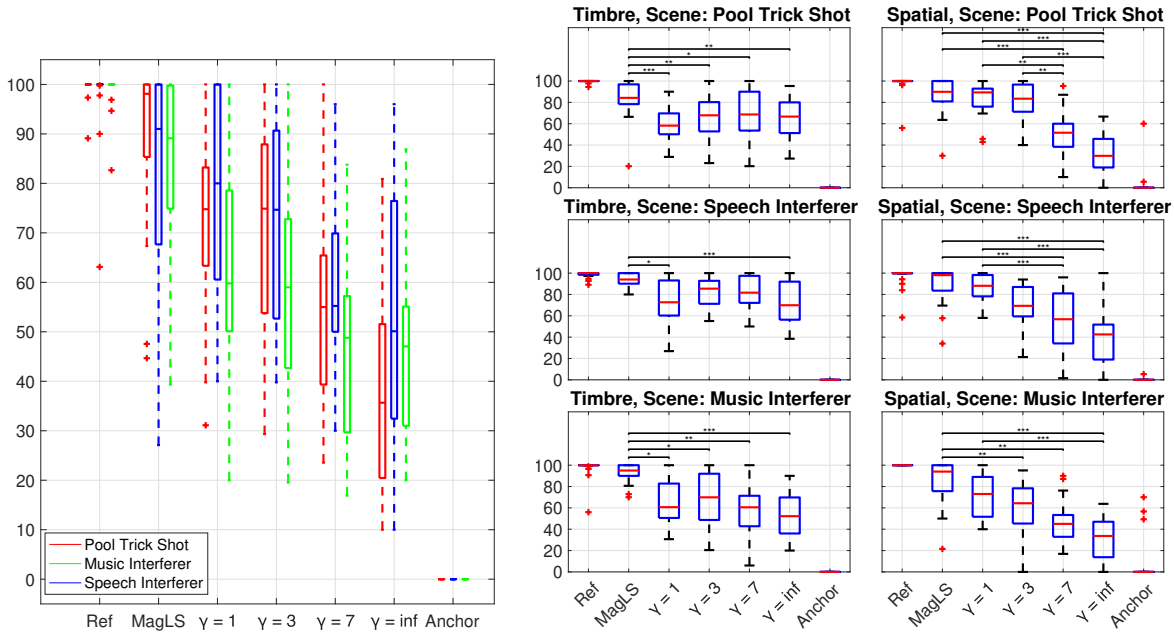
In the timbral part of the evaluation the assessors were asked to focus on the timbral difference between the reference and the test cases. Here, the test cases were created by duplicating the reference case 5 times, but then imposing the average magnitude response (over both channels and time) of the test cases onto these reference rendering duplicates, i.e., the equalisation gains to create the stimuli were $eq(f) = \sqrt{\frac{E_{test}(f)}{E_{ref}(f)}}$. The anchor in this section was a low-pass filtered rendering of the reference, with the cut-off frequency of the filter set at 500 Hz.

In the final overall part, the assessors were tasked with ranking the stimuli to score the test cases according to their overall similarity to the reference. The anchor in this instance was a low-passed (500 Hz), monaural rendering of the sound scene; i.e., a combination of the previously described anchors. No equalisation was conducted for this test part. Instead, the test cases were loudness normalised to the reference based on averaged broad-band root-mean-squared values.

5 Results and discussion

The “overall" results of the listening tests are displayed in Figure 2(a). There is a clear trend in the scores, with increasing values of γ resulting in a lower overall score. This indicates that increasing the value of γ causes more perceptual deviations from the simulated target rendering. However, this “overall" section of the results does not offer insight into whether the perceptual differences were caused by the spatial or timbral differences, or a combination of both.

The “spatial" and “timbral" test results were statistically analysed to provide insight. The results of the analyses are plotted in Figure 2(b). The analyses were conducted between the scores for the test cases for each sound scene, with the exception of the scores for the reference and anchor test cases which were excluded. An initial Friedman analysis found that there were statistically significant differences between the scores in each sound scene in both sections, i.e., for all six analyses



(a) Box plot of the overall section results. (b) Post-hoc analysis of the spatial and timbral section results.

Fig. 2: Perceptual evaluation results.

conducted. Subsequent post-hoc analysis was then conducted utilising Matlab's multcompare function with the Tukey Honest Significant Difference criterion applied. The post-hoc analyses found several statistically significant differences, indicated in Figure 2(b) by horizontal lines joining the box plots for the renderings between which these differences were found. The significance of the result is indicated by the number of asterisks above the line, with '*' indicating $p < 0.05$, '**' indicating $p < 0.01$, and '***' indicating $p < 0.001$.

The post-hoc analyses performed on the "timbral" test results reveal that while there are timbral effects resulting from the proposed algorithm, these effects do not seem to be proportionate to the value of γ. This is implied by the lack of significant differences found between test cases that have different values of γ as the only significant differences found in this section were between the regular MagLS test cases and the other test cases. Furthermore, the timbral deviations seem to be lower for speech interferers, i.e., speech sources located outside the FoV limits, compared to other more broadband interferers. This is indicated by slightly higher scores and fewer significant differences found in the speech interferer scene recordings versus the other two scenes.

The "spatial" test results, on the other hand, show that increasing the value of γ does cause increasing spatial deviations from the target. This is evident from the trend in the plotted results and is supported by the the post-hoc analyses for this section, which (unlike in the timbral section post-hoc analyses) found significant differences between renderings with increasing values of γ. There was, however, no significant difference found between the MagLS test scores and the scores for $\gamma = 1$ renderings, for all three scenes. Additionally, no significant difference was found between the former test scores and $\gamma = 3$ test scores. This indicates that, by utilising lower values of γ, a certain level of gain may be achieved for sources within the FOV limitswithout perceivable changes in spatial of the scene. It may also be concluded from these results that the deviations seen during the objective analysis are, indeed, perceptible depending on the chosen value for the γ parameter, as indicated by the objective analysis. In this part of the listening test there does not seem to be a significant difference arising from the type of the interferers present in the sound scene, unlike in the timbral section.

6 Summary

This paper presents a modification to the sensor-domain Magnitude-Least-Squares (MagLS) method, which conducts a fitting of the microphone array directivities (described by array transfer functions (ATFs)) to the binaural directivities (described by head related transfer functions (HRTFs)). The modification in question allows the application of different gains to sound sources that are emitting sounds from a selected range of directions, such as within a "field of view" (FoV) in-front of the listener. The proposed algorithm realises this by introducing direction-dependent gains into the formulation, controlled by a user parameter, $\gamma \geq 0$, which dictates the amount of emphasis placed on sounds within the FoV (or, conversely, the de-emphasis of sounds outside of the FoV).

An objective analysis of this proposed method was performed using measured ATFs of a pair of smartglasses, which featured an array of five microphones. The objective analysis found that direction dependent gain can be achieved. However, it also demonstrated that the binaural spatial cues are altered in a manner proportional to the magnitude of γ. Therefore, to study the implications of this, a multiple stimulus perceptual listening test was performed using sound scenes recorded by the same pair of smartglasses. The results of listening test indicate that the spatial properties of the sound scene are altered with increasing FoV emphasis, causing increasing perceptual deviations from a reference rendering. The listening test also found that there were perceivable timbral alterations, but these differences were not found to be statistically significant between different values of γ. There is, therefore, a compromise between increasing γ to achieve the desired emphasis within the FoV, and preserving the spatial characteristics of the recorded sound scene. Determining the exact threshold at which the perceived spatial attributes of the scene remain unchanged, while still achieving some degree of FoV enhancement, is a topic of future work. However, based on the results of the present study, it is thought to lay within the range $\gamma = [1,3]$, for the microphone array integrated into this particular pair of smartglasses.

Other avenues for future work include determining how altering the FoV limits may affect the outcome, since this study only tested the effect of changing values of γ with a fixed FoV range of +/- 30 degrees. Additionally, the emphasis proposed in this study may be viewed as a direct-dependent binary mask, i.e., with equal emphasis (1) given to all directions within the defined FoV and equal de-emphasis (0) given to all directions outside those limits. The effect of a more gradual change in emphasis when transitioning from within the FoV limits to outside of them could be investigated.

References

[1] Schultz-Amling, R., Kuech, F., Thiergart, O., and Kallinger, M., "Acoustical zooming based on a parametric sound field representation," in *Audio Engineering Society Convention 128*, Audio Engineering Society, 2010.

[2] Kronlachner, M. and Zotter, F., "Spatial transformations for the enhancement of Ambisonic recordings," in *Proceedings of the 2nd International Conference on Spatial Audio, Erlangen*, 2014.

[3] McCormack, L., Politis, A., and Pulkki, V., "Parametric spatial audio effects based on the multi-directional decomposition of ambisonic sound scenes," in *2021 24th International Conference on Digital Audio Effects (DAFx)*, pp. 214–221, IEEE, 2021.

[4] Favrot, A. and Faller, C., "Wiener-based spatial B-format equalization," *Journal of the Audio Engineering Society*, 68(7/8), pp. 488–494, 2020.

[5] Fernandez, J., McCormack, L., Hyvärinen, P., Politis, A., and Pulkki, V., "A spatial enhancement approach for binaural rendering of head-worn microphone arrays," in *24th International Congress on Acoustics (ICA2022)*, International Commission for Acoustics, 2022.

[6] As' ad, H., Bouchard, M., and Kamkar-Parsi, H., "A robust target linearly constrained minimum variance beamformer with spatial cues preservation for binaural hearing aids," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(10), pp. 1549–1563, 2019.

[7] Nikunen, J., Diment, A., Virtanen, T., and Vilermo, M., "Binaural rendering of microphone array captures based on source separation," *Speech Communication*, 76, pp. 157–169, 2016.

[8] Gerzon, M. A., "Periphony: With-height sound reproduction," *Journal of the audio engineering society*, 21(1), pp. 2–10, 1973.

[9] Zotter, F. and Frank, M., *Ambisonics: A practical 3D audio theory for recording, studio production, sound reinforcement, and virtual reality*, Springer Nature, 2019.

[10] Pomberger, H. and Zotter, F., "Warping of 3D ambisonic recordings," in *3rd International Symposium on Ambisonics and Spherical Acoustics*, 2011.

[11] Ivanic, J. and Ruedenberg, K., "Rotation matrices for real spherical harmonics. Direct determination by recursion," *The Journal of Physical Chemistry A*, 102(45), pp. 9099–9100, 1998.

[12] Schultz, F. and Spors, S., "Data-based binaural synthesis including rotational and translatory head-movements," in *Audio Engineering Society Conference: 52nd International Conference: Sound Field Control-Engineering and Perception*, Audio Engineering Society, 2013.

[13] Tylka, J. G. and Choueiri, E. Y., "Performance of linear extrapolation methods for virtual sound field navigation," *Aes: Journal of the Audio Engineering Society*, 2020.

[14] Moreau, S., Daniel, J., and Bertet, S., "3d sound field recording with higher order ambisonics–objective measurements and validation of a 4th order spherical microphone," in *120th Convention of the AES*, pp. 20–23, 2006.

[15] McCormack, L., Politis, A., Gonzalez, R., Lokki, T., and Pulkki, V., "Parametric ambisonic encoding of arbitrary microphone arrays," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30, pp. 2062–2075, 2022.

[16] Avni, A., Ahrens, J., Geier, M., Spors, S., Wierstorf, H., and Rafaely, B., "Spatial perception of sound fields recorded by spherical microphone arrays with varying spatial resolution," *The Journal of the Acoustical Society of America*, 133(5), pp. 2711–2721, 2013.

[17] Bertet, S., Daniel, J., Parizet, E., and Warusfel, O., "Investigation on localisation accuracy for first and higher order ambisonics reproduced sound sources," *Acta Acustica united with Acustica*, 99(4), pp. 642–657, 2013.

[18] Fernandez, J., McCormack, L., Hyvärinen, P., and Kressner, A. A., "Investigating sound-field reproduction methods as perceived by bilateral hearing aid users and normal-hearing listeners," *The Journal of the Acoustical Society of America*, 155(2), pp. 1492–1502, 2024.

[19] Ifergan, I. and Rafaely, B., *Theoretical framework for beamformer distribution in Beamforming based Binaural Reproduction*, Ph.D. thesis, Ben-Gurion University of the Negev, 2020.

[20] Madmoni, L., Donley, J., Tourbabin, V., and Rafaely, B., "Beamforming-based binaural reproduction by matching of binaural signals," in *Audio Engineering Society Conference: 2020 AES International Conference on Audio for Virtual and Augmented Reality*, Audio Engineering Society, 2020.

[21] McCormack, L., Meyer-Kahlen, N., Alon, D. L., Ben-Hur, Z., Gari, S. V. A., and Robinson, P., "Six-Degrees-of-Freedom Binaural Reproduction of Head-Worn Microphone Array Capture," *Journal of the Audio Engineering Society*, 71(10), pp. 638–649, 2023.

[22] Madmoni, L., Donley, J., Tourbabin, V., and Rafaely, B., "Binaural reproduction from microphone array signals incorporating head-tracking," in *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, pp. 1–5, IEEE, 2021.

[23] Fernandez, J., McCormack, L., Hyvärinen, P., Politis, A., and Pulkki, V., "Enhancing binaural rendering of head-worn microphone arrays through the use of adaptive spatial covariance matching," *The Journal of the Acoustical Society of America*, 151(4), pp. 2624–2635, 2022.

[24] Salvador, C. D., Sakamoto, S., Trevino, J., and Suzuki, Y., "Design theory for binaural synthesis: Combining microphone array recordings and head-related transfer function datasets," *Acoustical Science and Technology*, 38(2), pp. 51–62, 2017.

[25] Ahrens, J., Helmholz, H., Alon, D. L., and Garí, S. V. A., "Spherical harmonic decomposition of

a sound field using microphones on a circumferential contour around a non-spherical baffle," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30, pp. 3110–3119, 2022.

[26] Schörkhuber, C., Zaunschirm, M., and Höldrich, R., "Binaural rendering of ambisonic signals via magnitude least squares," in *Proceedings of the DAGA*, volume 44, pp. 339–342, 2018.

[27] Deppisch, T., Helmholz, H., and Ahrens, J., "End-to-end magnitude least squares binaural rendering of spherical microphone array signals," in *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, pp. 1–7, IEEE, 2021.

[28] Lübeck, T., Amengual Garí, S. V., Calamia, P., Alon, D. L., Crukley, J., and Ben-Hur, Z., "Perceptual evaluation of approaches for binaural reproduction of non-spherical microphone array signals," *Frontiers in Signal Processing*, 2, p. 883696, 2022.

[29] Politis, A., "Diffuse-field coherence of sensors with arbitrary directional responses," *arXiv preprint arXiv:1608.07713*, 2016.

[30] Yost, W. A. and Dye Jr, R. H., "Discrimination of interaural differences of level as a function of frequency," *The Journal of the Acoustical Society of America*, 83(5), pp. 1846–1851, 1988.