Audio Engineering Society

Conference Paper

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

S^{3}MASH: Spatial Sound Scene Matching using Single-Channel Audio

Raimundo Gonzalez[1], Leo McCormack[1], and Archontis Politis[2]

[1] *Department of Information and Communications Engineering, Aalto University, Espoo, Finland.*
[2] *Faculty of Information Technology and Communication Sciences, Tampere University, Finland*

Correspondence should be addressed to Raimundo Gonzalez (raimundo.gonzalez@aalto.fi)

ABSTRACT

This paper describes a novel approach for recording and binaurally reproducing spatial sound scenes using the audio from a single microphone. This is realised by recording the sound scene using both a microphone array, which potentially comprises more affordable and lower quality capsules, and a monophonic microphone, possibly featuring a higher quality capsule. By adopting a perceptually motivated sound-field model and estimating the model's spatial parameters, it is possible to define target time-frequency-dependent binaural spatial covariance matrices (SCMs). The actual binaural signals can then be synthesised using an adaptive SCM matching renderer, which takes only the higher-quality monophonic audio signal as input. A perceptual study was conducted to compare this novel processing approach, using a tetrahedral array and an omnidirectional microphone, against binaural renderings achieved through traditional Ambisonic means, when using four- and 32-channel arrays. The results show that, despite utilising only a monophonic signal for the spatialisation, the proposed approach yielded binaural renderings that are perceptually in-between the two conventional Ambisonic array renderings, with regards to their perceived spatial accuracy.

1 Introduction

Flexible and high-quality sound-field capture and reproduction solutions are becoming increasingly important and sought-after within the consumer space. Binaural playback has become a dominant means of consuming Augmented/Virtual Reality (AR/VR) audio and other immersive media content, where it can be especially important that the playback system facilitates sound-field modifications (such as rotations for head-tracking), and can offer other desirable rendering features, such as importing individualised Head-Related Transfer Functions (HRTFs).

A popular spatial audio format for spatial sound scene recording and playback is Ambisonics [1]. Spherical microphone arrays (SMAs) with uniformly distributed sensors are typically used for recording sound scenes in this format, since they can capture the sound-field with equal spatial resolution for all directions. The array signals can then be conveniently converted (or *encoded*) into Ambisonic signals [2]. These Ambisonic signals can optionally be manipulated (in a capture and playback agnostic manner) based on broad-band mixing matrices, in order to incorporate rotations (for head-tracking), and other spatial audio effects (e.g., sound-field warping, and direction-dependent loudness modifications) [3]. The Ambisonic signals can then be reproduced (or *decoded*) over the target playback setup based on a linear broad-band matrixing (for loudspeaker playback [4]), or via convolution with a matrix of filters (typically for binaural playback [5]). Since this decoding operation is decoupled from the recording setup, multiple sets of decoding filters (e.g., with individualised HRTFs) can be created. Therefore, due to this inherent flexibility, Ambisonics has found wide adoption within AR/VR and immersive media contexts.

Regarding current commercially-available SMAs, the most popular array configuration is that of an open tetrahedral arrangement of cardioid microphones. In order to save costs, it is usually beneficial for the encoding filters/software (shipped alongside the SMA) to be able to generalise well across multiple samples (of the same make and model) of the array; otherwise, potentially laborious and time consuming calibration measurements of each array sample is required. This therefore mandates that the microphone capsules are all phase-matched, which generally means that the capsules either become expensive (when using high-quality condenser capsules) or become more noisy (in the case of using e.g. MEMS). Examples of the former include the Sound-field SPS200 array, and the Sennheiser Ambeo array. Whereas, one of the more affordable offerings is that of the ZOOM H3-VR, which employs the use of MEMS.

These tetrahedral microphone arrays, when using linear encoding schemes [2], are capable of producing first-order Ambisonic (FOA) signals. However, this limited spatial resolution has been shown to incur perceptual issues when decoded using a linear decoder. These issues include sound source localisation ambiguities [6], a loss of listener envelopment, and a narrowing of apparent source width [7]. These limitations have led to the proposal of two main solutions. The first solution is to simply include many more microphones in the same array, thus facilitating the acquisition of higher-order Ambisonic (HOA) signals, which has been shown to alleviate these perceptual issues when paired with a higher-order linear decoder [6, 7]. The other solution is to replace the traditional linear binaural decoders with signal-dependent parametric alternatives [8, 9, 10, 11, 12]. These parametric approaches operate by analysing the inter-channel relationships between the microphone signals, in order to gain more insight into, for example: the directions of prominent sound sources in the scene, the energy-distribution of ambient/diffuse sounds, and the balance between them. This additional information is then used to, for example, spatially sharpen directional sounds, and decorrelate ambient sounds, in a time-frequency dependent manner. Such parametric rendering has been shown to improve the perceived spatial accuracy of the rendering, with perceptual studies demonstrating similar or better performance when using FOA signals as input, compared to what is achieved when rendering third-order Ambisonics signals (or higher) with a purely linear processing chain [8, 12, 13].

However, if given more stringent requirements on lowering costs, while retaining high-quality spatial audio capture and playback, these aforementioned solutions may be insufficient. For example, the majority of HOA capable microphone arrays are generally either very expensive, feature noisy sensors, or a combination of the two [14]. Whereas, while parametric methods using FOA signals derived from tetrahedral arrays featuring high quality (low noise) capsules may lead to perceptual improvements, such high quality tetrahedral arrays tend to feature expensive phase-matched capsules. Therefore, to further lower costs, the present authors postulated that it may be possible to develop a parametric rendering method capable of estimating the necessary spatial parameters using a more affordable lower-quality tetrahedral array, (featuring potentially noisier sensors), but to then use these spatial parameters to impose the spatial characteristics of the captured sound scene onto a single high quality/low noise omnidirectional microphone signal. Such an approach could lead to substantial cost reductions, and thus lower the barrier to entry for artists and audio engineers to enter the field of spatial audio recording.

In this paper, a novel parametric method is proposed, which analyses the spatial sound scene using a more affordable (and potentially noisy) microphone array, and employs these parameters to synthesise the playback signals by adaptively mixing a signal obtained from a single-channel (high quality) microphone that is suited near to the array during the recording. This study follows on from a similar study conducted in [15], except that the focus in this paper is on the binaural format (as opposed to targeting HOA), and the spatial covariance domain rendering approach of [16] is adopted for the spatialisation task. A perceptual study is then described, whereby binaural renderings using different real-world microphone array recordings are compared against a binaural dummy head reference.

2 Method

First consider a Q-channel microphone array, which is capturing signals, $\mathbf{x}(t,f) \in \mathbb{C}^{Q\times 1}$; where t and f are the time and frequency indices, respectively. Suitable options for this time-frequency transformation include a complex-Quadrature-Mirror-Filterbank (QMF), or alias-free short-time Fourier transform (afSTFT) [17].

The array signal model is given as

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

where $\mathbf{a}(\boldsymbol{\gamma},f) \in \mathbb{C}^{Q\times 1}$ is the array transfer function (ATF) for the direction $\boldsymbol{\gamma} \in \mathbb{S}^2$, corresponding to the most dominant source signal, s, at each time-frequency point; and $\mathbf{d} \in \mathbb{C}^{Q\times 1}$ is a residual component encapsulating diffuse ambience and less prominent directional sounds. Note that it is henceforth assumed that ATFs are available for a dense V-directional grid, $\mathbf{A} \in \mathbb{C}^{Q\times V}$. These may be obtained based on free-field measurements of the array, simulations, or (in the case of e.g. SMAs) analytically [18].

The narrow-band array spatial covariance matrices (SCMs), $\mathbf{C}_\text{x} \in \mathbb{C}^{Q\times Q}$, may be obtained as

$$\mathbf{C}_\text{x}(f) = \mathscr{E}[\mathbf{x}(t,f)\mathbf{x}^\text{H}(t,f)], \tag{2}$$

where $\mathscr{E}[.]$ denotes the expectation operator.

2.1 Spatial parameter estimation

The Multiple-Signal Classification (MUSIC) [19] algorithm may be used to estimate the source direction-of-arrival (DoA) for each frequency and time window as

$$\hat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma}}{\text{argmax}} \quad \mathbf{a}^\text{H}(\boldsymbol{\gamma},f)\mathbf{V}_\text{n}(f)\mathbf{V}_\text{n}^\text{H}(f)\mathbf{a}(\boldsymbol{\gamma},f), \tag{3}$$

where $\mathbf{V}_\text{n} \in \mathbb{C}^{Q\times(Q-1)}$ are the eigenvectors corresponding to the $(Q-1)$ lowest eigenvalues, obtained by performing an eigenvalue decomposition of the array SCM.

The source power can then be estimated as

$$p_\text{s}(f) = \mathbf{w}_\text{s}(\boldsymbol{\gamma},f)\mathbf{C}_\text{x}(f)\mathbf{w}_\text{s}^\text{H}(\boldsymbol{\gamma},f), \tag{4}$$

where $\mathbf{w}_\text{s} \in \mathbb{C}^{1\times Q}$ are beamforming weights, corresponding to the DoA. In this study, the matched-filter beamformer design was selected for this task

$$\mathbf{w}_\text{s}(\boldsymbol{\gamma},f) = \left(\mathbf{a}^\text{H}(\boldsymbol{\gamma},f)\mathbf{a}(\boldsymbol{\gamma},f)\right)^{-1}\mathbf{a}^\text{H}(\boldsymbol{\gamma},f). \tag{5}$$

An estimate of the ambience power may be obtained as

$$p_\text{d}(f) = \frac{1}{Q}\mathbf{tr}[\mathbf{W}_\text{d}(\boldsymbol{\gamma},f)\mathbf{C}_\text{x}(f)\mathbf{W}_\text{d}^\text{H}(\boldsymbol{\gamma},f)], \tag{6}$$

where $\mathbf{tr}[.]$ denotes the trace operator, and $\mathbf{W}_\text{d} \in \mathbb{C}^{Q\times Q}$ is an ambience beamforming matrix, which may be viewed as an operation that spatially subtracts the matched-filter beamformer pattern from the unit sphere, and may be computed as [12]

$$\mathbf{W}_\text{d}(\boldsymbol{\gamma},f) = \mathbf{I} - \mathbf{a}(\boldsymbol{\gamma},f)\mathbf{w}_\text{s}(\boldsymbol{\gamma},f), \tag{7}$$

where $\mathbf{I} \in \mathbb{R}^{Q\times Q}$ is an identity matrix.

2.2 Formulating target binaural SCM

With the DoAs, source powers, and ambient powers now estimated, the goal is to determine the appropriate target binaural SCMs, which the binaural signals (at the end of the chain) should exhibit. Note that these target SCMs should be updated independently per frequency bin and also for every time window. However, for notational clarity, the following matrices are formulated for a single time window.

The target SCMs, given the directional sounds in the scene, is given as

$$\mathbf{C}_\text{y,s}(f) = p_\text{s}(f)\mathbf{h}(\boldsymbol{\gamma},f)\mathbf{h}^\text{H}(\boldsymbol{\gamma},f), \tag{8}$$

where $\mathbf{h} \in \mathbb{C}^{2\times 1}$ is a HRTF corresponding to the estimated source direction. Note that it is assumed that HRTFs, $\mathbf{H} \in \mathbb{C}^{2\times V}$, are available for the same measurement/simulation grid as the ATFs.

The target SCMs, given ambient sounds in the scene, may be obtained as

$$\mathbf{C}_\text{y,d}(f) = p_\text{d}(f)\mathbf{H}(f)\mathbf{G}\mathbf{H}^\text{H}(f), \tag{9}$$

where $\mathbf{G} \in \mathbb{R}^{V\times V}$ is a diagonal matrices containing grid weights (with $\mathbf{tr}[\mathbf{G}] = 1$), to account for cases where the simulation/measurement grid is not uniform.

The total target binaural SCM is therefore

$$\mathbf{C}_\text{y}(f) = \mathbf{C}_\text{y,s}(f) + \mathbf{C}_\text{y,d}(f). \tag{10}$$

2.3 Obtaining suitable prototype signals

The adopted spatial covariance matching approach proposed in [16] mandates suitable prototype signals $\mathbf{z} \in \mathbb{C}^{2\times 1}$ (typically with the same number of channels as the target playback format; i.e., in this case, two) must first be obtained. In this study, the present authors propose that a single-channel microphone, $o(t,f) \in \mathbb{C}^{1\times 1}$, which is situated nearby to the microphone array during recording, may be first replicated to obtain these two channels. Note that, ideally, this single-channel microphone should exhibit an omni-directional directivity pattern, and posses lower noise characteristics compared to the microphone array.

A signal decorrelator, $\mathscr{D}_1[.]$, is then used reduce the correlation between the two channels as

$$\mathbf{z}(t,f) = [o(t,f)\mathscr{D}_1[o(t,f)]]. \tag{11}$$

Note that this decorrelator $\mathscr{D}_1[.]$ does not necessarily need to achieve zero inter-channel coherence values for all frequencies. This is motivated by the knowledge that the target binaural SCMs of Eq. 10 will generally feature high coherence at low-frequencies, even for a perfect diffuse-field (Section 2.3: [20]). If one then considers that low-frequencies are typically the most challenging to decorrelate without introducing artefacts (such as temporal smearing of transients, or other distortions), then one may be motivated to select a high-quality decorrelator, which achieves strong decorrelation at higher-frequencies, but minimal decorrelation at low-frequencies. The velvet noise based design of [21] is one example of such a decorrelator.

2.4 SCM matching

With suitable prototype signals and the target binaural SCMs now at hand, the goal is to adaptively mix the prototype signals $\mathbf{z} \in \mathbb{C}^{2\times 1}$ (which have a SCM of $\mathbf{C}_{\mathrm{z}} = \mathbb{E}[\mathbf{z}\mathbf{z}^{\mathrm{H}}]$)) in such a manner that binaural signals are produced $\mathbf{y} \in \mathbb{C}^{2\times 1}$, which instead exhibit SCMs that are more inline with the target $\mathbf{C}_{\mathrm{y}}$.

This adaptive mixing can be realised as [16]

$$\mathbf{y}(t,f) = \mathbf{M}(t,f)\mathbf{z}(t,f) + \mathbf{M}_{\mathrm{res}}(t,f)\mathscr{D}_2[\mathbf{z}(t,f)], \tag{12}$$

where $\mathbf{M} \in \mathbb{C}^{2\times 2}$ and $\mathbf{M}_{\mathrm{res}} \in \mathbb{C}^{2\times 2}$ are the optimal mixing matrices computed using the spatial covariance matching framework described in [16] (the reader is referred to the paper for their derivation); and $\mathscr{D}_2[.]$ denotes applying a decorrelation operation to the enclosed signals. Unlike the $\mathscr{D}_1[.]$ decorrelator in Section 2.3, note that these $\mathscr{D}_2[.]$ decorrelators are required to produce strongly uncorrelated versions of the enclosed signals at all frequencies, in order to fulfil the requirements of the adopted spatial covariance matching framework. These decorrelators may therefore introduce audible artefacts. However, it is highlighted that this framework is optimised to introduce these decorrelated signals (via $\mathbf{M}_{\mathrm{res}}$) into the output audio stream only to the degree that is necessary to achieve the target SCM. Previous studies have demonstrated that this type of rendering typically leads to higher overall audio quality, when compared to using more traditional parametric synthesis strategies [22, 10]; hence motivating its adoption for the present study.

3 Evaluation

The proposed method was evaluated through a formal listening test involving real recorded sound scenes.

3.1 Implementation of the proposed method

The proposed method was first implemented using the the alias-free STFT [17], configured with a hop size of $2.\dot{6}$ ms (128 samples at 48 kHz). The spatial analysis SCMs were averaged over $5.\dot{3}$ ms (256 samples at 48 kHz). The mixing matrices (in Equation (12)) were also updated for every averaging window, but were additionally recursively averaged using a one-pole filter with a coefficient value of 0.3. A velvet noise decorrelator [21] was used for replicating the omnidirectional signal ($\mathscr{D}_1$), which served as the prototype signals. A combination of frequency-dependent delay lines and lattice allpass filters was used for decorrelating[1] the prototype signals for the residual stream rendering ($\mathscr{D}_2$).

3.2 Test scenes

For the listening test, three spatial test scenes with two different room acoustics conditions were simulated, played back over a multichannel loudspeaker array, and recorded with various microphone configurations. These recordings were then reproduced over the binaural channels using various methods.

[1] Decorrelator can be found in the Spatial_Audio_Framework: `https://github.com/leomccormack/Spatial_Audio_Framework`

The three simulated test scenes were: **1)** a homophonic musical excerpt with four brass instruments (*Horns*), **2)** two overlapping tracks of male and female speech (*Speech*), and **3)** an excerpt with seven percussive instruments (*Percussion*). These excerpts were rendered with and without reverberation. The rendered reverberation corresponded to a shoe-box room of dimensions 20 x 18 x 5 meters, simulated using the SPARTA AmbiRoomSim (v1.0.0alpha) audio plugin [23] configured with a maximum reflection order of four and with fully reflective boundaries.

The scenes were then reproduced over a 44-channel loudspeaker array located in an anechoic chamber, as shown in Figure 1, using a fourth-order All-round Ambisonic Decoder (AllRAD) [24], as implemented in SPARTA AmbiDEC (v1.7.0) audio plugin [23]. The first room acoustic condition was anechoic (*Dry*), and therefore the signals of each source object were simply quantised to a single loudspeaker located on the horizontal plane. The assigned loudspeakers were -90, -30, 30 and 90 degrees, for the first scene, -60 and 60 degrees, for the second, and -90, -60, -30, 0, 30, 60, and 90 degrees for the third. For the reverberant room acoustic condition (*Wet*), the sound objects were placed in those same directions; however, since the simulated reverberation was three-dimensional, all loudspeakers in the array emitted audio.

These simulated scenes were then recorded using two recording setups. For the first, a GRAS 45BC Head & Torso simulator (KEMAR), was placed in the centre of the loudspeaker array. Whereas, for the second, the Eigenmike32 SMA, a ZOOM H3-VR 360° Audio Recorder, an AKG C414, and a Rødе NT-1 were placed in the centre of the loudspeaker array. Note that Eigenmike32 and Zoom H3-VR microphones are capable of recording in higher-order Ambisonics (*HOA*), and first-order ambisonics (*FOA*), respectively. The latter two, C414 and NT-1, are microphones with omnidirectional properties, and are commonly utilised by audio engineers due to their specific timbral and low-noise characteristics. The two microphone arrays were positioned at the center of the room coincidentally (one on top of the other), whereas the two omnidirectional microphones were 30 cm away from each other and from the two arrays, as shown in Figure 2.

3.3 Listening test design

The perceptual study was realised as a binaural multiple-stimulus test. The known (and hidden) *reference* condition was of the KEMAR recording, and a 3.5 kHz lowpass filtered version of this KEMAR recording served as an *anchor* condition. The first test case was of the ZOOM H3-VR recording encoded into *FOA* (using the manufacturers encoder), and then decoded to binaural using the SPARTA AmbiBIN (v1.6.0) audio plugin [23]. The second test case was of the Eigenmike32 recording encoded into fourth-order *HOA* (using the manufacturers encoder), and decoded using the same plugin. The third test case was of the proposed method using the raw recorded ZOOM H3-VR signals for the spatial analysis, and the AKG C414 microphone for the spatial synthesis (*S^3MASH (C414)*). The fourth test case was the same as the third test case, except using the Rødе NT-1 (*S^3MASH (NT-1)*).

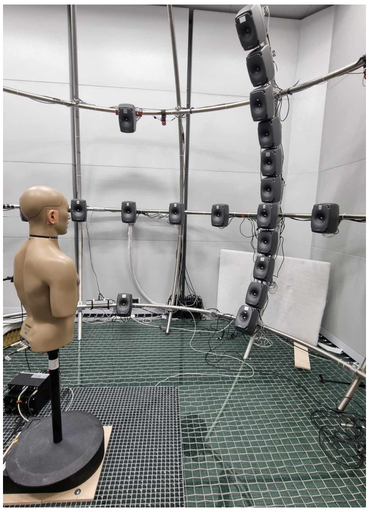

Fig. 1: Multichannel loudspeaker array used for simulating the sound scenes, with the KEMAR situated in the centre for recording the binaural reference signals.

The listening test was divided into two parts: spatial and timbral; following a similar test design and stimuli pre-processing approach described in [25]. For the **spatial** part, all test cases (except for the anchor) were equalised to the reference condition; thus, mitigating timbral differences between them, but retaining spatial differences. The listening test subjects were then instructed to rate the conditions based on their spatial similarity to the reference. For the **timbral** part, the reference condition was replicated and equalised based on the individual test cases; thus, minimising spatial differences between them, but retaining timbral colourations.

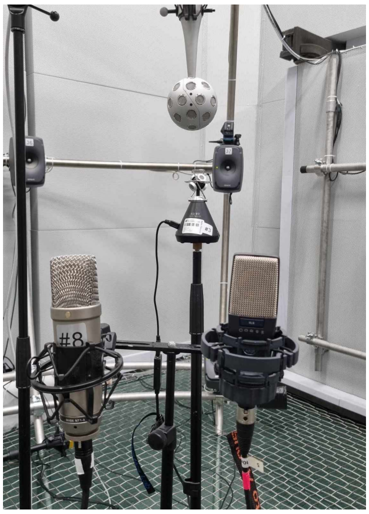

Fig. 2: Arrangement of the microphones and the microphone arrays, which were used for recording the simulated sound scenes, and passing to the different binaural rendering methods under test.

The listening test subjects were instructed to rate based on colourations with respect to the reference condition.

The test was conducted in specially-built sound-proof listening booths, and the subjects wore Sennheiser HD600 headphones. The webMUSHRA [26] framework was chosen for hosting the test material. The subjects were able to freely switch between the test cases as many times as they wished, before making their assessments. All subjects reported having normal hearing and had prior experience taking listening tests. The length of time required to complete both parts of the test was approximately 20 mins.

4 Results and discussion

The results of the perceptual study[2] given 14 participants, are presented in Figure 3 as a violin plot[3], which shows the medians as a white circle, the interquartile range as a black line, and the individual data points as coloured dots.

For the spatial portion of the test, participants rated the Eigenmike32 HOA renderings the highest for all sound scenes (other than the hidden reference), whereas the FOA H3-VR renderings generally received the lowest ratings. The proposed S^{3}MASH renderings generally received scores in-between these conventional FOA and HOA processing chains. This is more noticeable in test scenes where sources are not homophonic, such as the speech and percussion scenes. In the percussion test case, which contained seven transient sources, the proposed parametric method appears to capture and render the spatial properties of the scene more noticeably better than FoA, which is of particular interest, since previous studies have shown that parametric methods generally exhibit poorer performance when presented with transient material [10]. There then appears to be minimal perceived spatial differences between the parametric renderings of the C414 and NT-1, which indicates that the proposed method performance was largely independent of the chosen omnidirectional microphone and its placement; although, this could be explored further in a follow up study.

The timbre test did not reveal major differences between microphone arrays for the horns case. However, the H3-VR Zoom FOA and proposed method were perceived to have more noticeable timbral deviations from the binaural reference for both speech and percussion test scenes. These differences were perhaps revealed due to the broadband spectrum of the sound sources, which may more readily reveal colourations. This would be especially noticeable in the percussion case which contained a shaker instrument with high frequency signal content. However, since the precise cause of this timbral colouration remains unclear, future investigations will be carried out by the present authors.

It is finally noted that one aspect the conducted perceptual study does not reveal, is the amount of noise present in the reproduced binaural signals. It was reported by the test participants that some items were notably noisier than others, which the present authors believe are in reference to the Eigenmike32 renders in particular (as also reported in a previous study [14]). The reader is therefore encouraged to listen to the provided listening test items, in order to corroborate this particular aspect. However, this too, will be formally investigated in a future study.

5 Summary

This paper has proposed a novel parametric sound-field reproduction method, whereby the spatial properties

[2] Note that all listening test material is hosted here: `https://on.soundcloud.com/GNk7yr52JtnanK9F8`, `https://on.soundcloud.com/x7AGF5u27yfMz818A`

[3] `https://github.com/bastibe/Violinplot-Matlab`

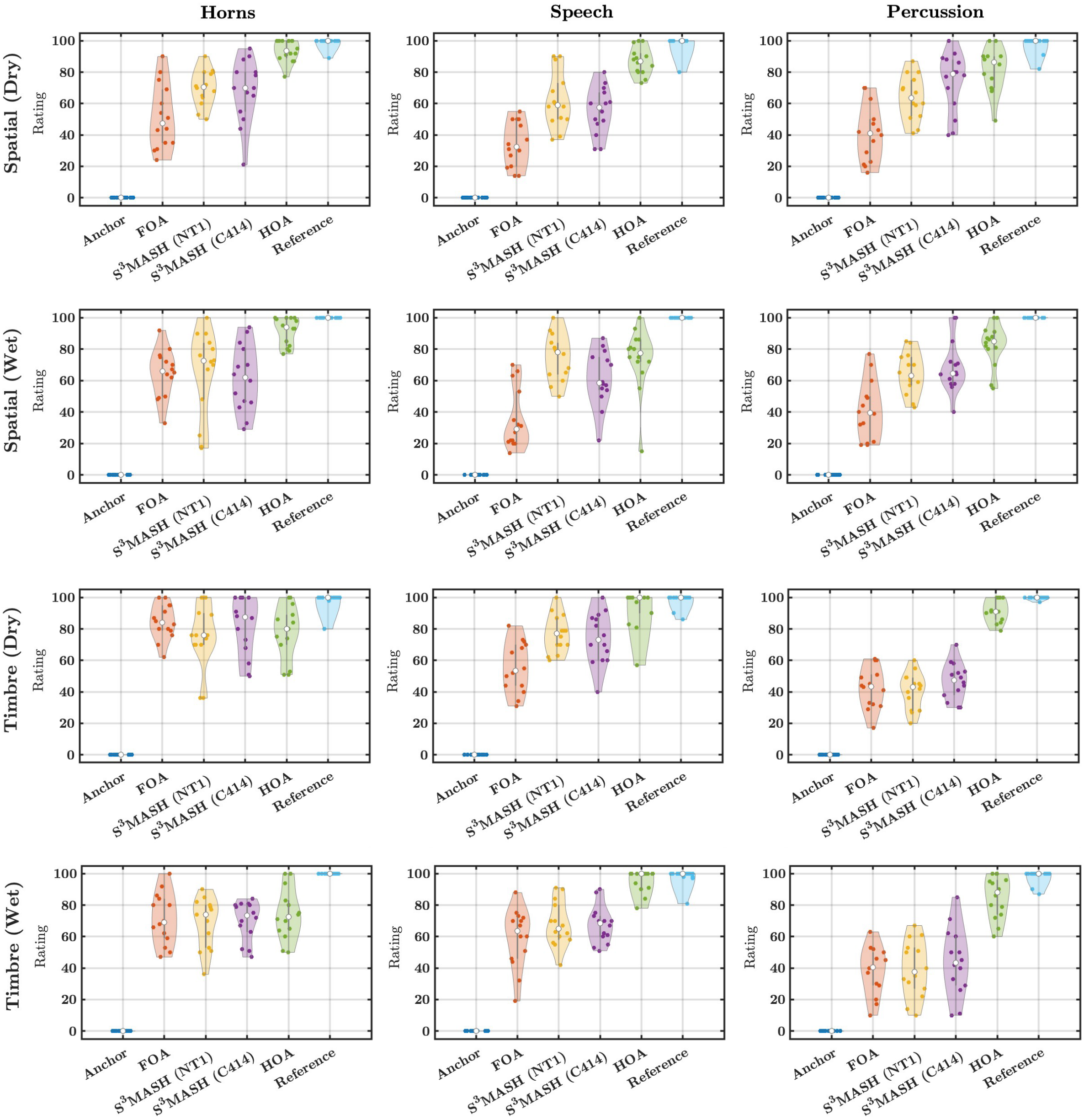

Fig. 3: Listening test results based on 14 participants.

of sound scenes are recorded and analysed using a microphone array, and then used to synthesise binaural signals using a separate monophonic microphone signal as input. Such an approach could potentially lead to cost reductions for the spatial audio recording and playback task, since the microphone array may comprise significantly cheaper (and lower quality) microphone capsules compared to the separate monophonic microphone. The binaural signals are synthesised using an established spatial covariance matching framework, which is optimised to reduce decorrelated signal energy in the output.

The results of a formal listening test indicated that the proposed method can attain binaural renderings that are spatially in-between renders achieved through conventional Ambisonic rendering means, when using four- and 32-channel microphone arrays, and compared against a binaural reference recording. However, the results also indicated that the proposed approach incurs noticeable timbral colourations, and ascertaining the root cause of this colouration was identified as a topic of future work.

References

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

[2] 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.

[3] Kronlachner, M. and Zotter, F., "Spatial transformations for the alteration of ambisonic recordings," *M. Thesis, University of Music and Performing Arts, Graz, Institute of Electronic Music and Acoustics*, 7, 2014.

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

[5] 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.

[6] 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.

[7] 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.

[8] Pulkki, V., "Spatial sound reproduction with directional audio coding," *Journal of the Audio Engineering Society*, 55(6), pp. 503–516, 2007.

[9] Berge, S. and Barrett, N., "High angular resolution planewave expansion," in *Proc. of the 2nd International Symposium on Ambisonics and Spherical Acoustics May*, pp. 6–7, 2010.

[10] Politis, A., McCormack, L., and Pulkki, V., "Enhancement of ambisonic binaural reproduction using directional audio coding with optimal adaptive mixing," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 379–383, IEEE, 2017.

[11] Schörkhuber, C. and Höldrich, R., "Linearly and quadratically constrained least-squares decoder for signal-dependent binaural rendering of ambisonic signals," in *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*, Audio Engineering Society, 2019.

[12] Politis, A., Tervo, S., and Pulkki, V., "COMPASS: Coding and multidirectional parameterization of ambisonic sound scenes," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6802–6806, IEEE, 2018.

[13] 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.

[14] Bates, E., Gorzel, M., Ferguson, L., O'Dwyer, H., and Boland, F. M., "Comparing Ambisonic

Microphones–Part 1," in *Audio Engineering Society Conference: 2016 AES International Conference on Sound Field Control*, Audio Engineering Society, 2016.

[15] McCormack, L., Gonzalez, R., Fernandez, J., Hold, C., and Politis, A., "Parametric Ambisonic Encoding using a Microphone Array with a One-plus-Three Configuration," in *Audio Engineering Society Conference: AES 2022 International Audio for Virtual and Augmented Reality Conference*, Audio Engineering Society, 2022.

[16] Vilkamo, J., Bäckström, T., and Kuntz, A., "Optimized covariance domain framework for time–frequency processing of spatial audio," *Journal of the Audio Engineering Society*, 61(6), pp. 403–411, 2013.

[17] Vilkamo, J. and Bäckström, T., "Time–frequency processing: Methods and tools," *Parametric Time-Frequency Domain Spatial Audio*, pp. 1–24, 2017.

[18] Williams, E. G., *Fourier acoustics: sound radiation and nearfield acoustical holography*, Academic press, 1999.

[19] Schmidt, R., "Multiple emitter location and signal parameter estimation," *IEEE transactions on antennas and propagation*, 34(3), pp. 276–280, 1986.

[20] Menzer, F., "Binaural audio signal processing using interaural coherence matching," Technical report, EPFL, 2010.

[21] Schlecht, S. J., Alary, B., Välimäki, V., Habets, E. A., et al., "Optimized velvet-noise decorrelator," in *Proc. Int. Conf. Digital Audio Effects (DAFx-18), Aveiro, Portugal*, pp. 87–94, 2018.

[22] Vilkamo, J. and Pulkki, V., "Minimization of decorrelator artifacts in directional audio coding by covariance domain rendering," *Journal of the Audio Engineering Society*, 61(9), pp. 637–646, 2013.

[23] McCormack, L. and Politis, A., "SPARTA & COMPASS: Real-time implementations of linear and parametric spatial audio reproduction and processing methods," in *AES International Conference on Immersive and Interactive Audio*, pp. 1–12, Audio Engineering Society, 2019.

[24] Zotter, F. and Frank, M., "All-round ambisonic panning and decoding," *Journal of the audio engineering society*, 60(10), pp. 807–820, 2012.

[25] 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.

[26] Schoeffler, M., Bartoschek, S., Stöter, F.-R., Roess, M., Westphal, S., Edler, B., and Herre, J., "webMUSHRA—A comprehensive framework for web-based listening tests," *Journal of Open Research Software*, 6(1), p. 8, 2018.