Estimating and Reproducing Ambience in Ambisonic Recordings

Leo McCormack[1] and Archontis Politis[2]
[1]*Department of Signal Processing and Acoustics, Aalto University*, Espoo, Finland
[2]*Department of Information Technology and Communication Sciences, Tampere University*, Finland
leo.mccormack@aalto.fi

***Abstract*—Spatial audio coding and reproduction methods are often based on the estimation of primary directional and secondary ambience components. This paper details a study into the estimation and subsequent reproduction of the ambient components found in ambisonic sound scenes. More specifically, two different ambience estimation approaches are investigated. The first estimates the ambient Ambisonic signals through a source-separation and spatial subtraction approach, and therefore requires an estimate of both the number of sources and their directions. The second instead requires only the number of sources to be known, and employs a multi-channel Wiener filter (MWF) to obtain the estimated ambient signals. One approach for reproducing estimated ambient signals is through a signal processing chain of: a plane-wave decomposition, signal decorrelation, and subsequent spatialisation for the target playback setup. However, this reproduction approach may be sensitive to spatial and signal fidelity degradations incurred during the beamforming and decorrelation operations. Therefore, an optimal mixing alternative is proposed for this reproduction task, which achieves spatially incoherent rendering of ambience directly for the target playback setup; bypassing intermediate plane-wave decomposition and excessive decorrelation. Listening tests indicate improved perceived quality when using the proposed reproduction method in conjunction with both tested ambience estimation approaches.**

***Index Terms*—ambisonics, spatial audio, microphone array processing**

I. Introduction

Separating multi-channel audio content into primary and ambient components, along with the estimation of related parameters, has been an integral part of a number of spatial upmixing [1]–[4] and parametric spatial audio reproduction methods [5]–[7]. Primary components, expressing spatially localisable sounds, are re-spatialised based on their estimated directions, whereas the ambience is reproduced in a diffuse manner over the target setup; often through the use of decorrelators. For ambisonic [8] recordings, in the simplest case, the separation model may comprise a single plane-wave source and an isotropic diffuse field for each time-frequency bin, as employed, for example, by the Directional Audio Coding (DirAC) method [5]. In the first formulation of DirAC, which was intended mainly for low-bitrate applications, an omnidirectional signal is separated into a primary directional stream and an ambience stream, based on a single-channel time-frequency mask. This mask is directly dictated by estimates of a diffuseness parameter. The isotropic ambient stream is then rendered based on distributing decorrelated copies of the monophonic signal to all channels of the playback setup. Subsequent DirAC formulations instead featured anisotropic ambient stream rendering by using the full ambisonic input [9]. These formulations apply the same single-channel diffuseness mask to all ambisonic channels, followed by spatially decoding them to a *virtual loudspeaker* arrangement. These virtual loudspeaker channels are then decorrelated and spatialised for the target playback setup.

An alternative model for ambience estimation was explored by the Coding and Multi-Parameterisation of Ambisonic Sound Scenes (COMPASS) method [6], which relies on first estimating a variable number of directional source signals; followed by re-encoding them and spatially subtracting them from the original ambisonic signals. This approach resembles the blocking matrix of linearly-constrained minimum variance spatial filtering, and can achieve improved primary-ambience separation compared to the single-channel diffuseness mask; provided that the estimation of the source signals is robust. This particular ambience estimation approach therefore relies on both source number estimation, applied per time-frequency tile, and subsequent direction-of-arrival (DoA) estimation of the detected sources. Source number estimation is typically performed through the analysis of the eigenvalues of the spatial covariance matrix of the array or ambisonic signals [10]. The identification of dominant eigenvalues and the respective eigenvectors then permits the segregation of the *signal* and *noise* subspaces, which are a common input for DoA estimation [11], [12] and spatial filtering methods [13], [14].

This subspace processing paradigm also gives rise to an alternative primary-ambience decomposition approach, which does not require DoA estimation. In the simplest case, the noise subspace eigenvalues and eigenvectors can be made to re-assemble a spatial covariance matrix corresponding to the scene ambience [14]. The ambient signals may then be estimated using a MWF. Alternatively, the method proposed in [15] avoids hard thresholding between the signal and noise subspace by instead devising a smooth eigenvalue weighting scheme, which aims to suppress strong directional components present in the input ambisonic signals. The approach, therefore, forgoes the need to estimate the number of sources. The method was, however, presented in the context of acoustic spatial visualisation, rather than for high-quality reproduction purposes. Similar subspace modeling and separation of ambience has also been used recently for virtual translation in a

higher-order ambisonic recording [16].

Regardless of how ambient signals are estimated, the reproduction of them is often conducted through: a plane-wave decomposition, decorrelation, and subsequent spatialisation over the intended playback setup; as employed, for example, by the DirAC formulation detailed in [9], and by the COMPASS method [6]. However, it is highlighted that high-quality decorrelation is challenging in practice, and any artefacts incurred during such operations will be aggregated at the output. This is then exacerbated by the need to conduct the plane-wave decomposition over many directions to preserve potential anisotropy of the ambience. In acknowledgement of these issues, an optimised covariance domain framework was proposed in [17], which aims to both appropriately reproduce spatially segregated audio streams, while also minimising the amount of decorrelated signal energy in the output. The framework has been previously employed by the DirAC formulations detailed in [18], [19], and also for the closed-form solution of a linearly and quadratically constrained decoder described in [7]. These previous formulations all employ a soft time-frequency mask to segregate the primary-ambience components, with the covariance domain framework used to solve the whole rendering problem; i.e. to render both the directional source and diffuse ambient components. However, for methods which estimate source or primary signals differently, with the aim of modifying them separately or remixing them, as with COMPASS (or other source separation approaches [20]), it may be beneficial to use the framework for the task of spatially enhancing, and reducing decorrelated signal energy, for only the ambience present in Ambisonic recordings.

Therefore, in this paper, the covariance domain framework described in [17] is investigated for the task of reproducing the ambience of sound scenes; using two different ambience estimation approaches. A multiple-stimulus test was conducted whereby reference binaural simulations were compared to first-order magnitude-least squares (MagLS) [21] and first-order COMPASS; with the latter substituting the ambient rendering with the four combinations of the two ambience estimation approaches and two reproduction methods[1]. The results indicate that the proposed reproduction approach improves the perceived quality in the majority of cases.

II. Preliminaries

It is assumed that the Ambisonic receiver signals of spherical harmonic order N have been first transformed into the time-frequency domain, $\mathbf{x}(t,f) \in \mathbb{C}^{(N+1)^2 \times 1}$, where t and f denote the down-sampled time and frequency indices, respectively. Note that the Ambisonic channel numbering (ACN) and ortho-normalised (N3D) Ambisonics conventions are employed for this study. The second order statistics are then given by the spatial covariance matrices (SCM) as

$$\mathbf{C}_\mathbf{x}(t,f) = \mathcal{E}[\mathbf{x}(t,f)\mathbf{x}^\mathrm{H}(t,f)], \tag{1}$$

[1]Note that all tested ambience estimation and reproduction approaches have been integrated into the COMPASS Binaural VST plugin, which can be downloaded from here: https://leomccormack.github.io/sparta-site

where $\mathcal{E}[.]$ denotes the expectation operator, which, in practice, involves applying temporal averaging in the range of tens of milliseconds.

The spatial analysis is based upon the subspace decomposition of the receiver SCMs as

$$\mathbf{C}_\mathbf{x} = \mathbf{V}\mathbf{\Sigma}\mathbf{V}^\mathrm{H} = \sum_{k=1}^{K} \sigma_k \mathbf{v}_k \mathbf{v}_k^\mathrm{H} + \sum_{k=K+1}^{(N+1)^2} \sigma_k \mathbf{v}_k \mathbf{v}_k^\mathrm{H}, \tag{2}$$

where $\sigma_1 > \sigma_2 > ... > \sigma_{(N+1)^2}$ are the eigenvalues in descending order, and $\mathbf{v}$ are their respective eigenvectors. Note that the signal subspace comprises the eigenvectors $\mathbf{V}_s \in \mathbb{C}^{(N+1)^2 \times K}$ corresponding to the largest K eigenvalues, whereas the noise subspace $\mathbf{V}_n \in \mathbb{C}^{(N+1)^2 \times [(N+1)^2 - K]}$ instead contains the eigenvectors corresponding to the lowest $(N+1)^2 - K$ eigenvalues. It is henceforth assumed that a source number estimate K, and corresponding DoA estimates $\mathbf{\Gamma_s}$, have been determined across time and frequency, through the application of the SORTE algorithm [10] on the eigenvalues, and MUSIC [11] using the noise subspace $\mathbf{V}_n$; as suggested and conducted by the COMPASS method [6].

A. Direct components rendering approach

The direct components of the sound scene may then be reproduced over the target V-channel playback system as [6]

$$\mathbf{y}_\mathrm{dir}(t,f) = \mathbf{G_s}\mathbf{D_s}\mathbf{x}(t,f), \tag{3}$$

where $\mathbf{D_s} \in \mathbb{C}^{K \times (N+1)^2}$ is a matrix of beamformers and $\mathbf{G_s} \in \mathbb{C}^{V \times K}$ are spatialisation weights (e.g. VBAP gains [22], binaural filters, or spherical harmonic vectors), which correspond to the estimated DoAs. This rendering approach for the direct components remains the same across all methods under test in this study.

III. Ambience estimation approaches under test

In this paper, two different ambience estimation approaches are investigated for their subsequent reproduction using both the proposed and baseline [9] ambience reproduction approaches. The first, is based on using both the source number and DoA estimates, in order to obtain an Ambisonics representation of the residual components in the scene, after the source signals have been subtracted from the input [6]

$$\mathbf{x}_\mathrm{d}^{(\mathrm{COMPASS})}(t,f) = (\mathbf{I} - \mathbf{Y}\mathbf{D_s})\mathbf{x}(t,f), \tag{4}$$

where $\mathbf{Y} \in \mathbb{R}^{(N+1)^2 \times K}$ and $\mathbf{D_s} \in \mathbb{R}^{K \times (N+1)^2}$ are SH vectors and beamforming weights for the estimated DoAs, and $\mathbf{I} \in \mathbb{R}^{(N+1)^2 \times (N+1)^2}$ is an identity matrix. This therefore produces ambient signals with the following SCM

$$\mathbf{C}_{\mathbf{x},\mathrm{d}}^{(\mathrm{COMPASS})} = (\mathbf{I} - \mathbf{Y}\mathbf{D_s})\mathbf{C}_\mathbf{x}(\mathbf{I} - \mathbf{Y}\mathbf{D_s})^\mathrm{H}, \tag{5}$$

The second approach considered is based on using a MWF and only the source number estimates to estimate the ambient Ambisonic signals as

$$\mathbf{x}_\mathrm{d}^{(\mathrm{KT-MWF})}(t,f) = \mathbf{C}_{\mathbf{x},\mathrm{d}}^{(\mathrm{KT})}\mathbf{C}_\mathbf{x}^{-1}\mathbf{x}(t,f), \tag{6}$$

where $\mathbf{C}_{\mathbf{x},\mathrm{d}} \in \mathbb{C}^{(N+1)^2 \times (N+1)^2}$ is an estimate of the ambient SCM using

$$\mathbf{C}_{\mathbf{x},\mathrm{d}}^{(\mathrm{KT})} = \mathbf{C}_{\mathbf{x}} - \mathbf{V}_s \mathrm{diag}[\sigma_1 - \sigma_{K+1}, ..., \sigma_K - \sigma_{K+1}]\mathbf{V}_s \quad (7)$$

where $\mathrm{diag}[.]$ denotes constructing a diagonal matrix from the enclosed vector.

IV. Baseline Ambience reproduction approach

Once the ambient Ambisonic signals have been estimated, they may then be reproduced over the target playback setup, based on spatialising a plane-wave decomposition of them, as conducted by more recent DirAC formulations [9] and by the COMPASS method [6]

$$\mathbf{y}_{\mathrm{diff}} = \psi \mathcal{D}[\mathbf{G}_{\mathrm{d}} \mathbf{Y}_{\mathrm{d}} \mathbf{x}_{\mathrm{d}}] + (1 - \psi)\mathbf{G}_{\mathrm{d}} \mathbf{Y}_{\mathrm{d}} \mathbf{x}_{\mathrm{d}}, \quad (8)$$

where $\mathbf{Y}_{\mathrm{d}} \in \mathbb{R}^{T \times (N+1)^2}$ are spherical harmonic weights for T directions uniformly distributed over the sphere, $\mathbf{G}_{\mathrm{d}} \in \mathbb{C}^{V \times T}$ are spatialisation gains for mapping the plane-wave decomposed signals to the playback setup, $\mathcal{D}[.]$ denotes a decorrelation operation on the enclosed signals, and $\psi \in [0, 1]$ is a diffusion level parameter dictating how much decorrelated signal energy is introduced into the output.

The final output signals are then obtained as

$$\mathbf{y} = \mathbf{y}_{\mathrm{dir}} + \mathbf{y}_{\mathrm{diff}}. \quad (9)$$

V. Proposed covariance domain based Ambience reproduction

In this paper, an alternative rendering approach is proposed. Narrow-band target covariance matrices $\mathbf{C}_{\mathbf{y},\mathrm{d}} \in \mathbb{C}^{V \times V}$, representing the inter-channel dependencies and channel energies which the ambient signals should exhibit, are first obtained as

$$\mathbf{C}_{\mathbf{y},\mathrm{d}} = \mathbf{G}_{\mathrm{d}}((\mathbf{Y}_{\mathrm{d}} \mathbf{C}_{\mathbf{x},\mathrm{d}} \mathbf{Y}_{\mathrm{d}}^{\mathrm{H}}) \odot \mathbf{F})\mathbf{G}_{\mathrm{d}}^{\mathrm{H}}, \quad (10)$$

where $\odot$ denotes the Hadamard product, and

$$\mathbf{F} = \begin{bmatrix} 1 & (1-\psi) & \dots & (1-\psi) \\ (1-\psi) & 1 & \dots & (1-\psi) \\ \vdots & \vdots & \ddots & \vdots \\ (1-\psi) & (1-\psi) & \dots & 1 \end{bmatrix}, \quad (11)$$

is a $T \times T$ matrix allowing for the same user-controllable diffusion parameter used in the baseline ambience reproduction approach, by scaling all non-diagonal elements of the plane-wave decomposed ambient SCM by $1 - \psi$. It is noted that this matrix may also be modified to permit direction-dependent diffusion control; for example, one may apply decorrelation to only the rear hemisphere of the recording. Although, investigating the perceptual ramifications of exercising this freedom was beyond the scope of the present study.

Prototype signals, which represent the starting point for the proposed reproduction approach, are then obtained as

$$\mathbf{y}_{\mathrm{proto}} = \mathbf{G}_{\mathrm{d}} \mathbf{Y}_{\mathrm{d}} \mathbf{x}, \quad (12)$$

with the optimal mixing problem outlined as [17]

$$\mathbf{y}_{\mathrm{diff}} = \mathbf{M}\mathbf{y}_{\mathrm{proto}} + \mathbf{M}_{\mathrm{r}} \mathcal{D}[\mathbf{y}_{\mathrm{proto}}], \quad (13)$$

where $\mathbf{M} \in \mathbb{C}^{V \times V}$ and $\mathbf{M}_{\mathrm{r}} \in \mathbb{C}^{V \times V}$ are the primary and residual mixing matrices, respectively. The solution to this problem is

$$\arg\min_{\mathbf{M},\mathbf{M}_{\mathrm{r}}} \mathbb{E}[||\mathbf{y}_{\mathrm{diff}} - \mathbf{A}\mathbf{y}_{\mathrm{proto}}||^2], \quad \text{subject to}$$
$$\mathbf{M}\mathbf{C}_{\mathrm{proto}}\mathbf{M}^{\mathrm{H}} + \mathbf{M}_{\mathrm{r}}(\mathrm{Diag}[\mathbf{C}_{\mathrm{proto}}])\mathbf{M}_{\mathrm{r}}^{\mathrm{H}} = \mathbf{C}_{\mathbf{y},\mathrm{d}}, \quad (14)$$

where $\mathbf{C}_{\mathrm{proto}}$ is the SCM of the prototype signals; $\mathrm{Diag}[.]$ denotes the construction of a diagonal matrix, which comprises the diagonal entries of the enclosed matrix; and

$$\mathbf{A} = \left(\mathrm{Diag}[\mathbf{C}_{\mathbf{y},\mathrm{d}}]\mathrm{Diag}[\mathbf{C}_{\mathrm{proto}}]^{-1}\right)^{-\frac{1}{2}}, \quad (15)$$

is an equalisation term to bring the prototype channel energies to be inline with the target energies.

The primary mixing matrix may be computed as

$$\mathbf{M} = \mathbf{K}_{\mathrm{diff}} \mathbf{V} \mathbf{U}^{\mathrm{H}} \mathbf{K}_{\mathrm{proto}}^{-1}, \quad (16)$$

based on the eigenvalue decompositions, or Cholesky factorisations, of $\mathbf{C}_{\mathbf{y},\mathrm{d}} = \mathbf{K}_{\mathrm{diff}}\mathbf{K}_{\mathrm{diff}}^{\mathrm{H}}$ and $\mathbf{C}_{\mathrm{proto}} = \mathbf{K}_{\mathrm{proto}}\mathbf{K}_{\mathrm{proto}}^{\mathrm{H}}$. The matrices $\mathbf{U}, \mathbf{V}$ are obtained based on the singular value decomposition of $\mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{H}} = \mathbf{K}_{\mathrm{proto}}^{\mathrm{H}} \mathbf{A} \mathbf{K}_{\mathrm{diff}}$.

The residual mixing may then be obtained with

$$\mathbf{M}_{\mathrm{r}} = \mathbf{K}_{\mathrm{r}} \mathbf{V}_{\mathrm{r}} \mathbf{U}_{\mathrm{r}}^{\mathrm{H}} \mathbf{K}_{\mathrm{proto}}^{-1}, \quad (17)$$

instead using the decompositions $\mathbf{C}_{\mathbf{y},\mathrm{d}} - \mathbf{M}\mathbf{C}_{\mathrm{proto}}\mathbf{M}^{\mathrm{H}} = \mathbf{K}_{\mathrm{r}}\mathbf{K}_{\mathrm{r}}^{\mathrm{H}}$, and $\mathbf{U}_{\mathrm{r}}\mathbf{S}\mathbf{V}_{\mathrm{r}}^{\mathrm{H}} = \mathbf{K}_{\mathrm{proto}}^{\mathrm{H}}\mathbf{A}\mathbf{K}_{\mathrm{r}}$.

VI. Evaluation

A multiple-stimulus listening test was conducted to investigate the perceived differences between the two ambience estimation methods, when combined with either the baseline reproduction approach or the proposed covariance domain alternative approach. Binaural reference scenarios were first simulated using the image-source method for two different shoebox room configurations. The first was a *medium* sized $10 \times 7 \times 4$ m (Width $\times$ Depth $\times$ Height) room, with RT60 times defined in octave band as [0.5 0.55 0.5 0.35 0.2 0.15] s (from 125 Hz to 4 kHz); while the second was a *large* $13 \times 8 \times 4$ m room, with RT60 times [0.8 0.7 0.6 0.4 0.25 0.2] s. The same configurations were then used to obtain first-order ambisonic signals, which were rendered to binaural using the four possible combinations of the ambience estimation approaches and ambience reproduction approaches under test, which were named: *compass_o1, compass_cm_o1, kt_mwf_o1, kt_cm_o1*, where *cm* denotes the use of the proposed covariance matching approach. All of which used the same direct stream rendering as described by Eq. 3.

The methods under test were all implemented using the 90% overlap alias-free short-time Fourier transform (STFT) described in [23], with a hop size of 128 samples and with the additional hybrid filtering of the lower-bands to obtain 133 frequency bands in total. Signal decorrelation was achieved through cascaded lattice all-pass filters, with longer filter structures for lower frequencies, as described in [24]. Both reproduction methods used $\psi = 1$ for all frequencies. As

TABLE I
OVERVIEW OF THE TEST CASES.

Name	Ambience Estimation	Ambience Reproduction
hidden_ref	N/A	Direction binauralisation
compass_cm_o1	As given by Eq. 4	Proposed approach (Eq. 13)
kt_cm_o1	As given by Eq. 6	Proposed approach (Eq. 13)
compass_o1	As given by Eq. 4	Baseline approach (Eq. 8)
kt_mwf_o1	As given by Eq. 6	Baseline approach (Eq. 8)
magls_o1	N/A	First-order MagLS decoder

TABLE II
LISTENING TEST SCENES. THE STIMULI ARE LISTED IN THE SAME ORDER AS THEY WERE POSITIONED FROM LEFT TO RIGHT.

Name	Room	Source stimuli
medium_speech	Medium	two male and two female speakers
medium_mix	Medium	clapping, water fountain, piano, speech
medium_band	Medium	shaker, drums, bass guitar, strings
large_speech	Large	two male and two female speakers
large_mix	Large	clapping, water fountain, piano, speech
large_band	Large	shaker, drums, bass guitar, strings

a further control, the MagLS [21] approach was included (*magls_o1*), to represent a state-of-the-art linear decoder. Furthermore, to have more parity with this control condition, the general $\mathbf{G}_\mathrm{d}\mathbf{Y}_\mathrm{d}$ operations in Eqs. 8 and 12, (which would represent virtual-loudspeaker based decoding in the present case), were replaced with this same MagLS decoder.

The simulations were conducted using three contrasting sets of four simultaneously played input stimuli: 1) multiple simultaneous speakers (*speech*), a mix of contrasting sound sources (*mix*), and a modern funk band ensemble (*band*). The source positions were placed on the horizontal plane at [90,30,-30,-90] degrees azimuth, one metre from the receiver located in the centre of the room. Note that the test cases have been summarised in Table I and the stimuli corresponding to the four source positions are listed in Table II (in order of appearance from left to right).

The listening test was divided into three parts and conducted similarly as described in [25]. In the first part, all of the test cases were equalised to match the reference test case, by passing the reference signals through the same STFT and averaging the magnitude responses over the entire simulated recording; prior to obtaining and then applying the appropriate equalisation curves. This operation therefore suppressed the timbral differences between all of the test cases. The listening test participants were asked to assess the test cases based on their **spatial** similarity with the reference, and ignore any timbral differences that they may perceive. The second part of the test involved instead duplicating the reference case and equalising them based on the other test cases. Therefore, the test cases all had spatial equivalency, with the subjects instead asked to rate them in terms of their **timbral** differences, and to ignore any spatial differences that they may perceive. For the final part of the test, the test cases were simply normalised to the reference based on their broad-band root-mean-square levels, and the listeners were asked to rate them based on their personal subjective weighting of the two previously isolated

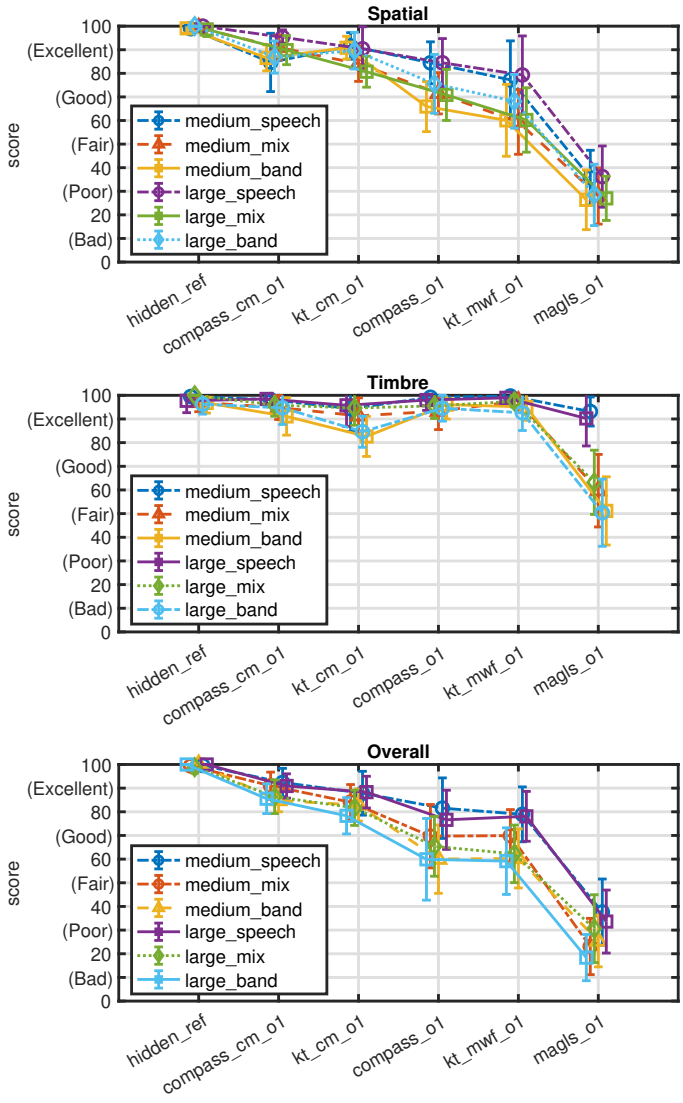

Fig. 1. Listening test results displaying means and 95% confidence intervals.

attributes for an **overall** score. For all three parts, the test interface had a scale between 0 and 100 and displayed the verbal anchors: “Bad”, “Poor”, “Fair”, “Good”, and “Excellent”, in steps of 20. The approximate duration for the three-part test was 40 minutes.

VII. RESULTS AND DISCUSSION

The results, represented as means and 95% confidence intervals based on 13 participants, are presented in Fig. 1 for all three listening test parts. For the **spatial** test, *magls_o1* was consistently rated the lowest for all tested scenes. Test cases using the proposed reproduction approach, *compass_cm_o1* and *kt_cm_o1*, in conjunction with the *mix* and *band* test scenes were rated notably higher than their baseline reproduction approach counterparts, *compass_o1* and *kt_mwf_o1*. However, this trend is less noticeable for the *speech* test scenes, with all four parametric test cases rated similarly, and within the range denoted by the “Excellent” verbal anchor.

For the **timbre** case, all combinations of the test cases and test scenes were largely rated as being transparent, or near-transparent, with respect to the reference, with the exception of *magls_o1* for the *mix* and *band* test scenes. Since lower-order ambisonics is known to result in low-pass-like behaviour [26], this may explain the high test scores for the speech test scenes where high-frequency content was limited. The *kt_cm_o1* test case was rated slightly lower for the two *band* test scenes, but still largely within the range denoted as "Excellent".

Finally, the **overall** test scores are more inline with the spatial test scores, as apposed to the timbre attribute test. It can therefore be inferred that the main differences between the methods under test were with respect to their spatial characteristics, where the proposed ambience rendering approach is found to produce output signals that are perceptually closer to the reference test case.

VIII. Conclusion

In this paper, the perceptual performance of two different ambience estimation approaches, when coupled with one of two different ambience reproduction approaches, was investigated. The Coding and Multi-Parameterisation of Ambisonic Sound Scenes (COMPASS) method served as a baseline, which estimates the ambient signals through a source signal estimation and subsequent subtraction approach; followed by reproducing the ambient Ambisonic signals through a plane-wave decomposition, decorrelation, and spatialisation for the target playback setup. The method therefore relies on both source number detection and DoA estimation. The second ambience estimation approach under test utilised only the information regarding the number of sources to first estimate the ambience spatial covariance matrix, subsequently using a MWF in order to estimate the ambient signals. These ambient signals may then be reproduced in the same manner as conducted by the COMPASS method. This paper then proposes an alternative ambience reproduction approach based on spatial covariance matching, which aims to better reproduce the intended spatial characteristics dictated by the sound-field model, while also reducing the use of decorrelation. This proposed reproduction approach is general, and may be used with any ambient signal estimation approach.

A binaural listening test was then conducted, where it is demonstrated that the proposed ambience reproduction approach produces output signals that are perceptually more similar to binaural reference scenarios, when compared to those produced by the plane-wave decomposition based alternative.

References

[1] C. Avendano and J.-M. Jot, "A frequency-domain approach to multichannel upmix," *J. Audio Eng. Soc.*, vol. 52, no. 7/8, pp. 740–749, 2004.

[2] C. Faller, "Upmixing and beamforming in professional audio," in *Parametric Time-Frequency Domain Spatial Audio*, V. Pulkki, S. Delikaris-Manias, and A. Politis, Eds. Wiley, 2017, p. 329.

[3] M. M. Goodwin and J.-M. Jot, "Primary-ambient signal decomposition and vector-based localization for spatial audio coding and enhancement," in *IEEE Int. Conf. Acoustics, Speech, and Sig. Proc. (ICASSP)*, 2007.

[4] J. He, E.-L. Tan, and W.-S. Gan, "Linear estimation based primary-ambient extraction for stereo audio signals," *IEEE/ACM Trans. Audio, Speech, and Language Proc.*, vol. 22, no. 2, pp. 505–517, 2014.

[5] V. Pulkki, A. Politis, M.-V. Laitinen, J. Vilkamo, and J. Ahonen, "First-order directional audio coding (DirAC)," in *Parametric Time-Frequency Domain Spatial Audio*, V. Pulkki, S. Delikaris-Manias, and A. Politis, Eds. Wiley, 2017, pp. 89–138.

[6] A. Politis, S. Tervo, and V. Pulkki, "COMPASS: Coding and multi-directional parameterization of ambisonic sound scenes," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6802–6806.

[7] C. Schörkhuber and R. Höldrich, "Linearly and quadratically constrained least-squares decoder for signal-dependent binaural rendering of ambisonic signals," in *2019 AES Int. Conf. on Immersive and Interactive Audio*, 2019.

[8] M. A. Gerzon, "Periphony: With-height sound reproduction," *J. Audio Engineering Society*, vol. 21, no. 1, pp. 2–10, 1973.

[9] M.-V. Laitinen and V. Pulkki, "Binaural reproduction for directional audio coding," in *IEEE Work. on Applications of Sig. Proc. to Audio and Acoustics (WASPAA)*, 2009, pp. 337–340.

[10] K. Han and A. Nehorai, "Improved source number detection and direction estimation with nested arrays and ULAs using jackknifing," *IEEE Trans. Signal Processing*, vol. 61, no. 23, pp. 6118–6128, 2013.

[11] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[12] B. Jo, F. Zotter, and J.-W. Choi, "Extended vector-based EB-ESPRIT method," *IEEE/ACM Trans. on Audio, Speech, and Language Proc.*, vol. 28, pp. 1692–1705, 2020.

[13] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. Sig. Proc.*, vol. 50, no. 9, pp. 2230–2244, 2002.

[14] F. Asano, S. Hayamizu, T. Yamada, and S. Nakamura, "Speech enhancement based on the subspace method," *IEEE Transactions Speech and Audio Proc.*, vol. 8, no. 5, pp. 497–507, 2000.

[15] N. Epain and C. T. Jin, "Super-resolution sound field imaging with sub-space pre-processing," in *IEEE Int. Conf. on Acoustics, Speech and Sig. Proc.*, 2013, pp. 350–354.

[16] M. Kentgens and P. Jax, "Ambient-aware sound field translation using optimal spatial filtering," in *IEEE Work. Applications of Sig. Proc. to Audio and Acoustics (WASPAA)*, 2021, pp. 236–240.

[17] J. Vilkamo, T. Bäckström, and A. Kuntz, "Optimized covariance domain framework for time–frequency processing of spatial audio," *J. Audio Eng. Soc.*, vol. 61, no. 6, pp. 403–411, 2013.

[18] A. Politis, J. Vilkamo, and V. Pulkki, "Sector-based parametric sound field reproduction in the spherical harmonic domain," *IEEE J. Selected Topics in Sig. Proc.*, vol. 9, no. 5, pp. 852–866, 2015.

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

[20] J. Nikunen and A. Politis, "Multichannel NMF for source separation with ambisonic signals," in *IEEE Int. Work. Acoustic Sig. Enhancement (IWAENC)*, 2018, pp. 251–255.

[21] C. Schörkhuber, M. Zaunschirm, and R. Höldrich, "Binaural rendering of Ambisonic signals via magnitude least squares," in *Proc. DAGA*, vol. 44, 2018, pp. 339–342.

[22] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *J. Audio Eng. Soc.*, vol. 45, no. 6, pp. 456–466, 1997.

[23] J. Vilkamo and T. Bäckstrom, "Time-frequency processing: Methods and tools," in *Parametric Time-Frequency Domain Spatial Audio*, V. Pulkki, S. Delikaris-Manias, and A. Politis, Eds. Wiley, 2017, pp. 1–24.

[24] J. Herre, H. Purnhagen, J. Breebaart, C. Faller, S. Disch, K. Kjörling, E. Schuijers, J. Hilpert, and F. Myburg, "The reference model architecture for MPEG spatial audio coding," Tech. Rep., 2005.

[25] J. Fernandez, L. McCormack, P. Hyvärinen, A. Politis, and V. Pulkki, "Enhancing binaural rendering of head-worn microphone arrays through the use of adaptive spatial covariance matching," *The Journal of the Acoustical Society of America*, vol. 151, no. 4, pp. 2624–2635, 2022.

[26] Z. Ben-Hur, F. Brinkmann, J. Sheaffer, S. Weinzierl, and B. Rafaely, "Spectral equalization in binaural signals represented by order-truncated spherical harmonics," *J. Acoustical Soc. of America*, vol. 141, no. 6, pp. 4087–4096, 2017.