ICA 2022 24th INTERNATIONAL CONGRESS ON ACOUSTICS October 24th-28th, 2022 Gyeongju, Korea

PROCEEDINGS of the
24th International Congress on Acoustics

October 24 to 28, 2022 in Gyeongju, Korea

ABS-0745

A spatial enhancement approach for binaural rendering of head-worn microphone arrays

Janani FERNANDEZ[1]; Leo MCCORMACK[1]; Petteri HYVÄRINEN[1];
Archontis POLITIS[2]; Ville PULKKI[1]

[1] Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland

[2] Department of Information Technology and Communication Sciences, Tampere University, Finland

ABSTRACT

This paper builds upon a recently proposed spatial enhancement approach, which has demonstrated improvements in the perceived spatial accuracy of binaurally rendered signals using head-worn microphone arrays. The foundation of the approach is a parametric sound-field model, which assumes the existence of a single source and an isotropic diffuse component for each time-frequency index. The enhancement approach involves the post-processing of an initial estimate of the binaural signals, in order to obtain a refined estimate of binaural signals which more closely represent the inter-aural cues corresponding to the sound-field model. In this contribution, the enhancement approach has been implemented as an open-source framework, written in both the MATLAB and C programming languages, and as a real-time audio plug-in. The framework was also extended to offer direction-dependent gain control of sound sources relative to the listener, and a frequency-dependent control of the direct-to-diffuse balance, which are modifications that may find application within future augmented reality headsets and assistive hearing devices.

Keywords: microphone array processing, spatial audio, augmented reality, binaural hearing aids

1. INTRODUCTION

Binaural rendering of head-worn microphone arrays is receiving increased attention, largely due to the emergence of commercially-available augmented reality headsets and binaural hearing aid devices. In such applications, it is desirable for the captured surrounding sound scene to be reproduced over headphones with high perceived spatial accuracy and transparency, and for the reproduction method to facilitate some degree of hearing augmentation. There have been a number of signal-independent methods proposed recently [1, 2], which are specifically intended for processing head-worn microphone array input, and operate based on a linear mapping of the array signals to the binaural channels. However, while such methods have low computational requirements and achieve high signal fidelity, their maximum attainable spatial resolution is inherently limited by the number of microphones and their placement on the array geometry.

There are binaural rendering approaches which have been proposed by the hearing-aid research community [3, 4, 5], and proposals involving spherical microphone arrays (SMAs) within the spatial audio research community [6, 7], which also extract additional spatial information through observations of the inter-channel relationships between the array signals themselves. Typically, such methods (often referred to as parametric methods within the spatial audio field) impose assumptions regarding the composition of the input sound scene, and estimate meaningful spatial parameters over time and frequency. These elements are then used to dictate the rendering; typically, by informing the look-direction of spatial filters (beamformers) with their signals convolved with the respective head-related transfer functions (HRTFs). Provided that the spatial analysis techniques are robust, such signal-dependent methods have been shown to outperform their signal-independent counterparts in formal perceptual studies [7]. However, due to the nature of time-frequency domain processing, audible artefacts can occur.

[1] janani.fernandez@aalto.fi

Recently, a general spatial enhancement approach, based on spatial covariance matching, was proposed in [8]. In principle, the enhancement approach may be applied to the output of any existing binaural rendering approach; but the focus of the study however, revolved around head-worn microphone arrays for augmented reality and hearing aid applications.

Building on the recent work of [8], this paper details a MATLAB toolbox and a real-time VST audio plugin implementation of the algorithms detailed therein[1].

2. BINAURAL RENDERING FRAMEWORK

This section provides an overview of the binaural rendering framework.

2.1 Sound-field model

The employed signal model assumes that the M input microphone array signals $\mathbf{x} \in \mathbb{C}^{M\times 1}$ describe a single source signal s, a diffuse field $\mathbf{d} \in \mathbb{C}^{M\times 1}$, or a combination of the two at each time-frequency index

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

where $\mathbf{a}(\gamma,f) \in \mathbb{C}^{M\times 1}$ is the array transfer function for incident direction γ. Note that it is henceforth assumed that these array transfer functions, $\mathbf{A} \in \mathbb{C}^{M\times V}$ have been either measured or simulated for a dense grid of V incident directions $\mathbf{\Gamma} = [\gamma_1, ..., \gamma_V]$.

The array spatial covariance matrix (SCM) is given as

$$\mathbf{C_x}(f) = \mathbb{E}[\mathbf{x}(t,f)\mathbf{x}^{\mathrm{H}}(t,f)] = \mathbf{a}(\gamma,f)\mathbf{a}^{\mathrm{H}}(\gamma,f)\mathbb{E}[|s(t,f)|^2] + \mathbb{E}[\mathbf{d}(t,f)\mathbf{d}^{\mathrm{H}}(t,f)], \tag{2}$$

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

2.2 Spatial parameter estimation

Many source number detectors and diffuseness estimation approaches operate based upon the analysis of the eigenvalues of the time-averaged array SCMs. Typically, such approaches rely on the SCMs becoming diagonal (with all eigenvalues being equal) when there are no sources active or when high levels of diffuse sound is present. However, especially at lower frequencies, the array SCMs can deviate from this diagonal structure when capturing diffuse-fields. Therefore, a spatial whitening operation is applied as

$$\mathbf{C_x}^{(w)}(f) = \mathbf{T}(f)\mathbf{C_x}(f)\mathbf{T}^{\mathrm{H}}(f), \tag{3}$$

where $\mathbf{C_x}^{(w)} \in \mathbb{C}^{M\times M}$ is the spatially whitened array SCM, and $\mathbf{T} \in \mathbb{C}^{M\times M}$ is the whitening matrix as described in further detail in [9].

The spatially whitened array SCM is then decomposed using an eigenvalue decomposition as

$$\mathbf{C_x}^{(w)}(f) = \mathbf{V}\mathbf{\Sigma}\mathbf{V}^{\mathrm{H}} = \sum_{m=1}^{M} \sigma_m \mathbf{v}_m \mathbf{v}_m^{\mathrm{H}}, \tag{4}$$

where σ are the eigenvalues sorted in descending order and $\mathbf{v}$ are their respective eigenvectors.

A diffuseness estimate is then obtained, based on the method described in [10], as

$$\psi(f) = 1 - \frac{\beta}{\beta_0} \tag{5}$$

where $\beta_0 = 2(M-1)$, $\beta = \frac{1}{\langle\sigma\rangle}\sum_{m=1}^{M} |\sigma_q - \langle\sigma\rangle|$, and $\langle\sigma\rangle = \frac{1}{M}\sum_{m=1}^{M}\sigma_m$.

For estimating the direction-of-arrival of the most prominent sound source in the scene, the MUtiple-Signal Classification (MUSIC) approach [11] is employed as

$$P_{\mathrm{MUSIC}}(\gamma,f) = \frac{1}{||\mathbf{V}_{\mathrm{n}}^{\mathrm{H}}\mathbf{T}(f)\mathbf{a}(\gamma,f)||^2} \quad \text{for } \gamma \in \mathbf{\Gamma}, \tag{6}$$

[1]The open-source framework and audio plugin may be found here: `https://github.com/jananifernandez/HADES`

with the frequency-dependent DoA estimates subsequently extracted from the resulting pseudo-spectrum, P_{MUSIC}, by identifying the direction at which the function is minimised.

2.3 Baseline binaural rendering approach

The spatial enhancement approach described in the following subsection is a post-processing operation, which is applied onto binaural signals which are produced by an existing baseline method. The baseline signals are therefore first obtained as

$$\mathbf{y}_{\text{bl}}(t,f) = \mathbf{Q}(t,f)\mathbf{x}(t,f), \tag{7}$$

where $\mathbf{Q} \in \mathbb{C}^{2\times M}$ is the baseline binaural rendering mixing matrix.

For example, in [8], three different baseline approaches, which are found in hearing aid research literature, were explored. One of the approaches involved the selection of two reference signals, the nearest microphones to each ear canal, and then to simply route them to the respective ear canals. The other two approaches were based on establishing a balance between binaural beamformers [3, 12, 13] and the reference signals, which has previously been formulated in [14, 15] using a user-controllable parameter. In [8], however, the balance was dictated by the time-frequency-dependent diffuseness term.

Note that in the case of simply routing two reference signals to the binaural channels, this baseline mixing matrix is both frequency- and time-independent. Whereas, for other time-invariant methods, e.g. [1, 2], the baseline mixing matrix may be frequency-dependent. While other baseline methods may employ signal-dependent beamformers, e.g. [3], and thus the mixing matrix may be both frequency- and time-dependent.

2.4 Spatial covariance matching based enhancement

The spatial enhancement approach is based on adaptively mixing the baseline binaural signals, in order to obtain refined estimates of binaural signals, which should more closely match the employed sound-field model. The enhancement approach is based on first defining target narrow-band binaural SCMs as

$$\begin{aligned}\mathbf{C}_{\text{y}}(f) = \big(1-\psi(f)\big)P_{\text{total}}(f)\mathbf{h}(\gamma,f)\mathbf{h}(\gamma,f)^{\text{H}} + \\ \psi(f)P_{\text{total}}(f)\mathbf{D}_{\text{bin}}(f),\end{aligned} \tag{8}$$

where $P_{\text{total}} = \mathbf{tr}[\mathbf{C}_{\mathbf{x}}]$ is an estimate of the total signal energy, $\mathbf{h} \in \mathbf{C}^{2\times 1}$ are HRTFs corresponding to the analysed DoA, and $\mathbf{D}_{\text{bin}} = \mathbf{H}\mathbf{H}^{\text{H}} \in \mathbb{C}^{2\times 2}$ is a binaural coherence matrix derived using a dense grid of HRTF measurements $\mathbf{H} = [\mathbf{h}_1, ...\mathbf{h}_V] \in \mathbb{C}^{2\times V}$.

The spatial enhancement operation is then applied via the mixing matrix $\mathbf{M} \in \mathbb{C}^{2\times 2}$ as

$$\mathbf{y}_{\text{enh}}(t,f) = \mathbf{M}(f)\mathbf{y}_{\text{bl}}(t,f) = \mathbf{M}(f)\mathbf{Q}(t,f)\mathbf{x}(t,f), \tag{9}$$

which has been derived, through an optimisation process, as the matrix required to best ensure the following holds true:

$$\mathbb{E}[\mathbf{y}_{\text{enh}}(t,f)\mathbf{y}_{\text{enh}}^{\text{H}}(t,f)] = \mathbf{M}(f)\mathbf{Q}(t,f)\mathbf{C}_{\mathbf{x}}(f)\mathbf{Q}^{\text{H}}(t,f)\mathbf{M}^{\text{H}}(f) \approx \mathbf{C}_{\text{y}}(f). \tag{10}$$

Note that the solution to this problem is detailed in [8].

3. IMPLEMENTATION

The binaural rendering framework and the spatial enhancement approach described in [8] was implemented in both the MATLAB and C languages. Both implementations are divided into two main stages: analysis and synthesis. In the analysis stage, the DoA and diffuseness parameters are estimated over time and frequency, based on the input microphone array signals. The time-frequency signals are stored in a *signal container*, and the results of the analysis are placed into a *parameter container*. These containers may then be optionally modified, before being passed onto the synthesis stage, which renders the binaural signals.

The C code implementation was also integrated into a real-time VST audio plugin; the graphical user interface for which is depicted in Fig. 1. The plugin supports the loading of arbitrary sets of array transfer functions and HRTFs via the SOFA standard. All three baseline binaural rendering methods described

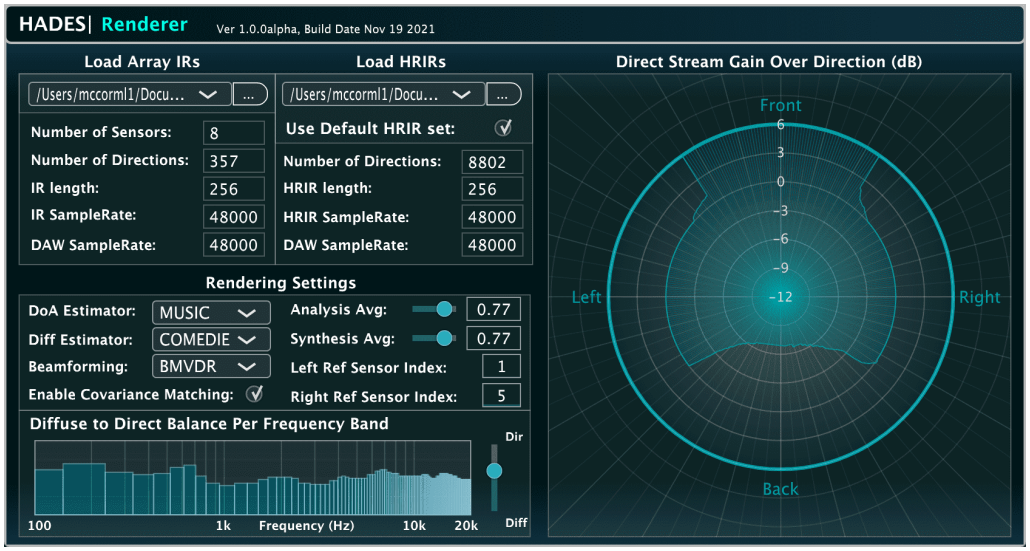

Figure 1: The graphical user interface of the developed VST audio-plugin.

in [8] are integrated. Due to the decoupling of the analysis and synthesis stages of the framework, it is possible to manipulate the signal and/or parameters containers prior to synthesising the scene. This may be carried out to augment the rendering, in order to realise certain spatial audio effects and other sound-field modifications. Note that an overview of such parametric effects may be found in [16] using an alternative rendering framework; but with many of these effects still applicable to the present framework. In the developed VST audio plugin, the frequency-dependent direct-to-diffuse balance control is implemented, as described in [16], this allows diffuse sounds to be attenuated (i.e. de-reverberation), or exaggerated. Furthermore, a direction-dependent gain control was incorporated within the plugin, which enables an addition gain factor to be applied only to the direct stream components.

4. CONCLUSIONS

This paper has presented an open-source MATLAB, C, and audio-plugin implementation of a spatial enhancement approach, and the associated binaural rendering algorithms. The implemented spatial enhancement approach, which is based on signal-dependent spatial covariance matching operations, is a post processing algorithm intended to be applied to the output of an existing binaural rendering method. The implementation of the present binaural rendering framework also decouples the spatial analysis and synthesis stages. This allows spatial audio effects and sound-field modifications to be applied through simple manipulations of the estimated parameters. The developed audio plugin supports the application of direction dependent gain factors, which only affect sound components which are characterised as having a clear directionality, and also allows frequency-dependent balancing modification of the direct-to-diffuse balance.

REFERENCES

1. Madmoni L, Donley J, Tourbabin V, 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. .
2. Ahrens J, Helmholz H, Alon DL, Garí SVA. A head-mounted microphone array for binaural rendering. In: 2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA). IEEE; 2021. p. 1-7.
3. Hadad E, Gannot S, Doclo S. Binaural linearly constrained minimum variance beamformer for hearing aid applications. In: IWAENC 2012; International Workshop on Acoustic Signal Enhancement. VDE; 2012. p. 1-4.
4. As' ad H, Bouchard M, 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. 2019;27(10):1549-63.
5. Marquardt D, Hadad E, Gannot S, Doclo S. Theoretical analysis of linearly constrained multi-channel Wiener filtering algorithms for combined noise reduction and binaural cue preservation in binaural hearing aids. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2015;23(12):2384-97.
6. Laitinen MV, Pulkki V. Binaural reproduction for directional audio coding. In: 2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. IEEE; 2009. p. 337-40.
7. Politis A, Tervo S, Pulkki V. COMPASS: Coding and multidirectional parameterization of ambisonic sound scenes. In: IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP); 2018. p. 6802-6.
8. Fernandez J, McCormack L, Hyvärinen P, Politis A, 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. 2022;151(4):2624-35.
9. McCormack L, Politis A, Gonzalez R, Lokki T, Pulkki V. Parametric Ambisonic Encoding of Arbitrary Microphone Arrays. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2022;30:2062-75.
10. Epain N, Jin CT. Spherical harmonic signal covariance and sound field diffuseness. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2016;24(10):1796-807.
11. Schmidt R. Multiple emitter location and signal parameter estimation. IEEE transactions on Antennas and Propagation. 1986;34(3):276-80.
12. Hadad E, Marquardt D, Doclo S, Gannot S. Theoretical analysis of binaural transfer function MVDR beamformers with interference cue preservation constraints. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2015;23(12):2449-64.
13. Doclo S, Gannot S, Moonen M, Spriet A, Haykin S, Liu KR. Acoustic Beamforming for Hearing Aid Applications. Handbook on Array Processing and Sensor Networks. 2010:269-302.
14. Van den Bogaert T, Doclo S, Wouters J, Moonen M. Speech enhancement with multichannel Wiener filter techniques in multimicrophone binaural hearing aids. The Journal of the Acoustical Society of America. 2009;125(1):360-71.
15. Corey RM. Microphone array processing for augmented listening. University of Illinois at Urbana-Champaign; 2019.
16. McCormack L, Politis A, Pulkki V. Parametric Spatial Audio Effects Based on the Multi-Directional Decomposition of Ambisonic Sound Scenes. In: Proceedings of the 24th International Conference on Digital Audio Effects (DAFx20in21); 2021. p. 214-21.