Synthesizing a Virtual Height Channel from Planar Microphone Arrays

Stefan Wirler
Dpt. of Information and
Communications Engineering
Aalto University
Espoo, Finland
stefan.wirler@aalto.fi

Nils Meyer-Kahlen
Dpt. of Information and
Communications Engineering
Aalto University
Espoo, Finland
nils.meyer-kahlen@aalto.fi

Ville Pulkki
Dpt. of Information and
Communications Engineering
Aalto University
Espoo, Finland
ville.pulkki@aalto.fi

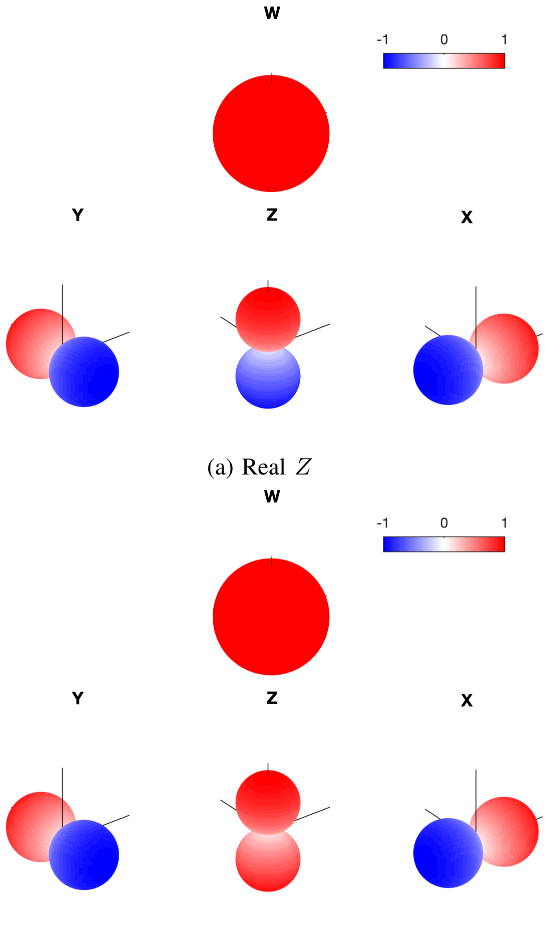
***Abstract*—In spatial audio processing, the ability to capture sound scenes in three dimensions is important for applying algorithms such as beamforming and direction-of-arrival (DoA) estimation. Conventional microphone arrays require at least four microphones that are not on the same plane to achieve three-dimensional processing. This work presents a method for synthesizing a virtual height channel from planar microphone arrays that could, for example, be placed on a flat surface such as a table, enabling full 3D Ambisonic processing without requiring physically elevated microphones. The approach relies on transforming array signals into the spherical harmonics (SH) domain, where only the W, X, and Y components are available for a planar array. The missing Z-component is synthesized using spectral subtraction, assuming that sources only originate from the upper half-space. The basic method strictly requires time-frequency sparsity; here, we also present an improved version incorporating diffuseness estimation to enhance robustness in reverberant environments. Evaluation results demonstrate the effectiveness of the synthesized height channel for DoA estimation and beamforming in simulated conditions, achieving comparable performance to recordings with a true height channel. This technique enables practical implementations of 3D sound field analysis with simpler, more flexible microphone configurations.**

***Index Terms*—Ambisonics, Beamforming**

I. Introduction

In microphone array processing, a common aim is to process a recorded sound scene concerning all three Cartesian dimensions of space. Processing might, for example, involve determining direction-of-arrival (DoA) of one or more sound sources in the recorded signal or extracting any of these sources using beamforming [1] or spatial post-filters [2]. These techniques are essential in applications such as teleconferencing and hands-free communication where good source localization and separation are required.

Other applications that rely on accurate three-dimensional processing include spatial sound reproduction techniques, such as Ambisonics [3] and parametric spatial audio methods [4], which aim to create an immersive experience by reconstructing three-dimensional sound fields. Additionally, augmented and virtual reality (AR/VR) systems depend on three-dimensional microphone array processing to enhance user experience by accurately reproducing spatial audio as well.



(a) Real Z

(b) Synthesized $\hat{Z}$

Fig. 1: First order SH patterns with real and synthesized Z-channel. For the upper half space, both patterns are identical.

To apply such algorithms in three dimensions, microphone arrays comprising at least four receivers are required. If they are omnidirectional, they may not be arranged in the same plane. However, application scenarios exist in which only three microphones are available or in which very flat array designs are desired. An ultra-flat array might, for example, be integrated into a conference table.

In this paper, we present a straightforward method for allowing three-dimensional processing based on such flat, two-dimensional arrays. For this, the array signals are first transformed to the spherical harmonics (SH) domain. Flat arrays will only allow the synthesis of the W, X and Y components of a first-order SH signal. Then, we use spectral subtraction to create a “virtual” Z-channel (see Figure 1). So far, this approach is not widely known, even though it has already been used as part of a planar variant of directional audio coding (DirAC) [5]. Here, we show that the algorithm can be used beyond the context of DirAC. We show how a perfect Z-channel can be synthesized if the sound-field is sparse in time or frequency and that all sound sources are located in one half-space. The latter is easily fulfilled in the case of a flat table-top array, but the sparsity assumption might be violated if recordings are made in real rooms that comprise reverberation. Therefore, we also describe an extended version of the method that takes the diffuseness of the recorded scene into account.

Finally, we show that the estimated Z-channel can be used for DoA estimation using the pseudo intensity vector (PIV) and for beamforming, even in simulated multispeaker scenes with reverberation, as they could be found in a meeting room. We show numerical results for DoA estimation and beamforming for the extended method, and also for the basic method, which was not done in [5] yet.

II. Method

Conventionally, when 3D microphone arrays are used, a first order Ambisonics signal $\boldsymbol{\chi}(t)$ comprises the W, X, Y and Z channels. These channels are created by linearly combining the recorded microphone signals $p(t)$. The so-called encoding matrix $\boldsymbol{E}$ required for this can be derived depending on the specific microphone configuration [6]

$$\boldsymbol{\chi}(t) = \begin{bmatrix} W(t) \\ X(t) \\ Y(t) \\ Z(t) \end{bmatrix} = \boldsymbol{E}\boldsymbol{p}(t). \tag{1}$$

For the encoding, different approaches can be employed.

For the presented approach, we assume that only a flat array is used, where the Z-channel cannot be obtained by linear combination. The energy found in the Z-channel can, however, be derived from the energies of the other channels. This is possible through the following energy consideration [5], which we write in the time-frequency domain

$$\left|\sqrt{3}W(f,t)\right|^2 = |X(f,t)|^2 + |Y(f,t)|^2 + |Z(f,t)|^2, \tag{2}$$

from which the Z-channel energy is easily obtained. The multiplication by a factor $\sqrt{3}$ is needed to scale the signals appropriately. The factor depends on the chosen normalization scheme - here N3D normalization [6], [7]. If the Ambisonics signal uses a different normalization scheme, e.g. SN3D, re-normalization needs to be applied first. To derive a signal that resembles $Z(f,t)$ which has an appropriate phase, we must assume that sound sources only arrive from one half-space. When thinking about a flat, table-top array, this is easily justifiable. Also, we must assume that there is one dominant source per time-frequency tile, to obtain the energy

$$\bar{Z}^2(f,t) = \left|\sqrt{3}W(f,t)\right|^2 - \left(|Y(f,t)|^2 + |X(f,t)|^2\right). \tag{3}$$

In [5], they called the summation of the X- and Y-channel “*Torus* signal”, describing the shape of its implied directivity pattern. Applying the phase of the omnidirectional signal, $\angle W(f,t)$, as

$$\hat{Z}(f,t) = \sqrt{\max\left(0, \bar{Z}^2\right)} e^{j\angle W(f,t)}, \tag{4}$$

allows estimating the *virtual* or *synthesized* Z-channel as shown in Figure 1. By omitting negative values, complex-valued results due to energy mismatches are avoided.

A. Diffuse Sound

Clearly, the assumption of time-frequency sparsity does not hold in many real sound-fields, for example when reverberation is present. Hence, we propose an extension to the algorithm that can take diffuse energy into account.

Similar to the underlying model of many parametric sound reproduction methods [4], we assume that the direct sound is *mixed* with a diffuse sound field, originating from reverberation, for example.

The first step for incorporating this assumption is the estimation of the diffuseness. This is done by analyzing the eigenvalues of the spatial covariance matrix $\mathbf{C}(f,t)$. In [8], Epain et al. showed that in a diffuse sound field, the eigenvalues become similar, whereas in a directional sound field the eigenvalues are the most dissimilar or exhibit the greatest spread. Here, the spatial covariance matrix is constructed with three channels as

$$\mathbf{C}(f,t) = \begin{bmatrix} S_{WW} & S_{WX} & S_{WY} \\ S_{WX}^* & S_{XX} & S_{XY} \\ S_{WY}^* & S_{XY}^* & S_{YY} \end{bmatrix}, \tag{5}$$

where $S_{WW} = \mathbb{E}[3W(f,t)W^*(f,t)]$ is the auto-power spectral density of the W-Channel, $S_{WX} = \mathbb{E}[\sqrt{3}WX^*(f,t)]$ and $S_{WY}(f,t) = \mathbb{E}[\sqrt{3}W(f,t)Y^*(f,t)]$ are the cross-power spectral densities between the W-Channel and the X- and Y-channel, respectively. Since the resulting matrix is only 3×3, its three eigenvalues can be efficiently calculated using characteristic polynomials [9] as follows:

$$\lambda_{\substack{\text{max}\\\text{min}}}(f,t) = \frac{\text{tr}(\mathbf{C}(f,t)) \pm \sqrt{\text{tr}(\mathbf{C}(f,t))^2 - 3\det(\mathbf{C}(f,t))}}{3}, \tag{6}$$

$$\lambda_{\text{mid}}(f,t) = \text{tr}(\mathbf{C}(f,t)) - \lambda_{\max}(f,t) - \lambda_{\min}(f,t). \tag{7}$$

The diffuseness of the corresponding time-frequency bin is then calculated as

$$d(f,t) = 1 - \frac{\lambda_{\max}(f,t) - \lambda_{\text{mid}}(f,t) - \lambda_{\min}(f,t)}{\lambda_{\max}(f,t) + \epsilon}, \tag{8}$$

with ϵ as a regularization parameter. Based on the spread of the eigenvalues, Eq. 8 yields values close to 0 for highly directional signals and approaches 1 for diffuse sound fields. The obtained signals are then combined as follows:

$$\tilde{Z}_e(f,t) = (1 - d(f,t))\hat{Z}(f,t) + d(f,t)\tilde{Z}^{\text{diffuse}}(f,t), \quad (9)$$

where the diffuse part is simply estimated as the magnitude of the omnidirectional channel $\tilde{Z}^{\text{diffuse}}(f,t) = |W(f,t)|$. This assumption holds in a complete diffuse field, however not in a real diffuse field which is a mix of several coherent and fully diffuse components.

Finally, the extended estimated Z-channel is synthesized with,

$$\tilde{Z} = |\tilde{Z}_e| e^{j\angle W(f,t)}. \quad (10)$$

The Z-channel in the time domain is then easily obtained by applying an inverse Fourier-transform to the corresponding time-frame. Now the newly synthesized virtual Z-channel can be used for any task usually conducted on an ordinary, 3D Ambisonics signal, such as DoA estimation and beamforming. Finally, the estimated Ambisonics signal is obtained by

$$\tilde{\chi}(t) = \begin{bmatrix} W(t) \\ X(t) \\ Y(t) \\ \tilde{Z}(t) \end{bmatrix}. \quad (11)$$

III. Evaluation

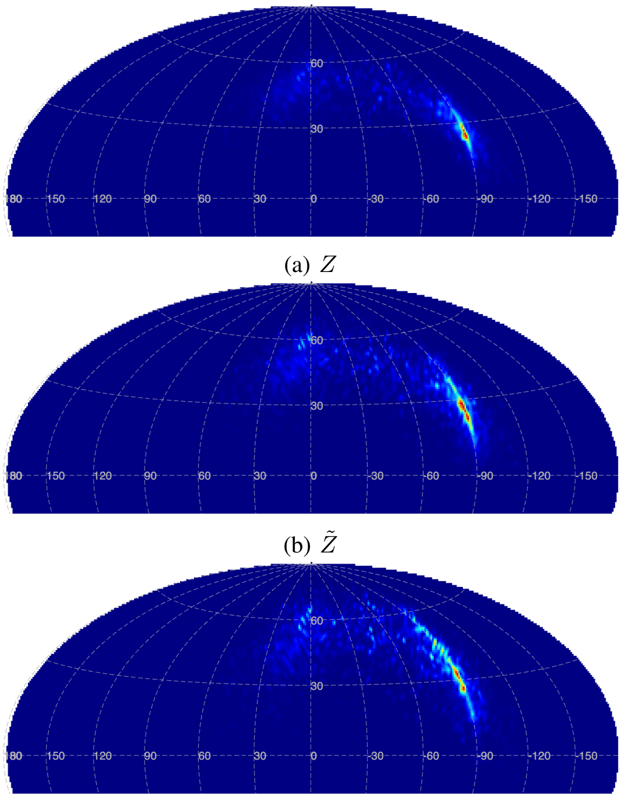
The evaluation is carried out by using simulated signals. The simulation is done by using an image-source-based shoebox room simulator tuned for spherical audio processing [10]. Rather than simulating microphone signals, the simulator directly generates the spherical harmonic signals at a single point in space. In the simulation, the assumption of a flat table-top array is included by rejecting all sound and reflections arriving from directions considered below the array for all channels. First, the capabilities of DoA estimation with the synthesized Z-channels are carried out. The comparison is performed using signals with the synthesized Z-channels and the simulated (ground-truth) Z-channel. Second, the performance of source separation is assessed by employing a first-order hypercardioid steered towards two different speech sources. The processing parameters used for the evaluation are shown in Table I

TABLE I: Simulation Parameters

Parameter	Symbol	Typical Values
Sampling Rate	f_s	48 kHz
Frame Length	N	2048 samples
Frame Overlap	O	50%
Window Function	-	$\sqrt{\text{Hann}}$
Room Dimensions	$\mathbf{D}$	(10.2, 7.1, 3.2) m
Microphone Array Position	$\mathbf{r}_m$	(4, 3.5, 0.8) m
Reverberation Time	RT_{60}	{0.2, 0.25, 0.5} s

The different methods are compared, by using the simulated W-, X-, Y-channels and the following Z-channels as:

- Z: Signal with simulated Z-channel (ground truth)
- $\tilde{Z}$: Synthesized signal with diffuse modeling (proposed)
- $\hat{Z}$: Synthesized signal without diffuse modeling [5]



(a) Z

(b) $\tilde{Z}$

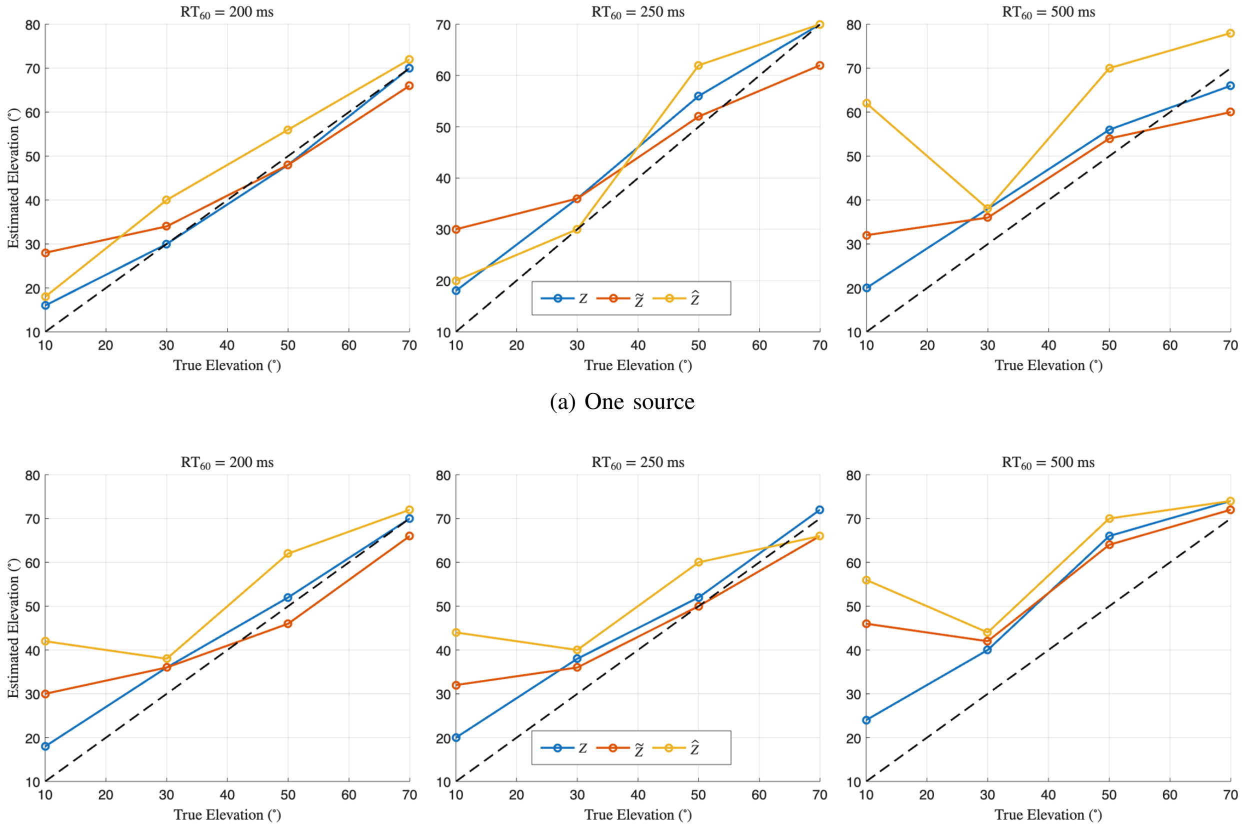
(c) $\hat{Z}$

Fig. 2: Normalized DoA estimation histogram for two sources located at $(-14, 44)°$ and $(90, 22)°$, estimated with the (a) simulated signal Z, (b) extended synthesized $\tilde{Z}$ and (c) the synthesized $\hat{Z}$ with the simple model.

A. DoA Estimation

The DoA estimation is performed using PIV-based [11] estimation, as implemented in [12]. The DoA is determined within 15 ms time frames of the signal. Figure 2 presents the estimated histogram for two sources positioned at azimuth and elevation angles of $(-14, 44)°$ and $(90, 22)°$, respectively. In this example, the reverberation time is set to $\text{RT}_{60} = 250$ ms. Comparing the direction estimates from the simulated Z-channel with the synthesized Z-channel, our approach exhibits less deviation in estimated directions. Compared to the plane-wave assumption, our method results in a more concentrated intensity direction, indicating better preservation of directivity. The extended model reduces the spread of the estimated direction, leading to a more accurate synthesis of the Z-channel. However, a slight deviation from the simulated Z-channel remains, suggesting potential areas for further refinement.

Figure 3a shows the DoA estimation for a single-source scenario of a female speaker across different elevation angles and reverberation times for the simulated, the extended and the plane-wave Z-estimation. For most elevations and reverberation time conditions, our method follows the estimated elevation of the simulated channel more closely. Only for low elevation angle, i.e., 10° and very dry rooms, the basic method



(a) One source

(b) Two sources

Fig. 3: Estimated elevation angle in the case of (a) one source and (b) two sources, with the simulated Z-channel Z, the extended method $\tilde{Z}$, and the plane-wave assumption $\hat{Z}$ for three different reverberation times.

shows a slightly lower error.

The same evaluation was conducted for a two-speaker scenario, as shown in Figure 3b. The second source is a male English speaker placed at an elevation of 34° with a separation of 75° degrees to the first source. The extended approach outperforms the basic approach in all of the cases and shows elevation estimates similar to those of the simulated Z-channel for higher reverberation times and elevations. The offset at a small elevation is still apparent but is similar for the extended Z-channel estimation compared to the one speaker scenario for lower reverberation times. Whereas the Z-channel estimated with the basic method shows a high estimation error.

B. Beamforming

The next evaluation, the capability of source separation via beamforming is investigated in a two-speaker scenario. The scene consists of two simultaneous speakers (male and female) closely spaced with different heights. The first source (Src 1) is located at (5.6 4.2 1.1) m and the second source (Src 2) at (4.7 3.8 1.7) m, resulting in an angular distance of 24° for the azimuth and 55° for the elevation. Such a scenario could be encountered due to a seated and a standing person.

For the evaluation, we use the STOI, SIR, SDR, and SAR metrics [13], [14]. To ensure consistency across all methods, the hypercardioid beamformer is always steered toward the desired speaker using the ground-truth DoA calculated from the known positions of the sources and the receiver. In addition to the elevation estimation, also a beamformed signal was used, which was only steered along the horizontal axis, i.e., no Z-channel is used (Z=0). The same processing and simulation parameters as those used in the DoA estimation are applied.

The evaluation results are presented in Table II for different reverberation times and the desired source (Src). In addition to the beamformed signals, the metrics for the input, i.e., the omnidirectional signal, are shown. All methods improve performance across all objective measures for the first source compared to the omnidirectional input signal. Omitting the SAR, which is expected to be the highest for the ”real” simulated Z-channel, our proposed method achieves the best results for all reverberation times steering towards the first source.

In the case of source 2 being the target, our method performs worse compared to the simulated and the basic method in terms of STOI, SDR, and SIR. However, this difference decreases at higher reverberation times. The results indicate a stronger focus on the dominant source for the presented method.

Further, the beamformed signal with the extended method always shows a higher SAR for all cases compared to the

TABLE II: STOI (lin), SDR (dB), SIR (dB) and SAR (dB) of the beamformed signal for different reverberation times

Src	M	$RT_{60} = 0.20s$				$RT_{60} = 0.25s$				$RT_{60} = 0.50s$			
		STOI	**SDR**	**SIR**	**SAR**	**STOI**	**SDR**	**SIR**	**SAR**	**STOI**	**SDR**	**SIR**	**SAR**
1	In	0.656	-3.315	-3.024	15.587	0.650	-3.548	-2.990	12.511	0.618	-4.888	-2.732	6.062
	Z	0.746	-0.627	0.648	19.553	0.743	-0.762	0.663	16.486	0.725	-1.516	0.698	10.044
	$\tilde{Z}$	0.764	-0.276	1.394	18.782	0.762	-0.425	1.432	15.813	0.744	-1.308	1.535	9.478
	$\hat{Z}$	0.751	-0.603	0.846	18.586	0.749	-0.736	0.919	15.686	0.734	-1.586	1.112	9.361
	$Z = 0$	0.791	1.551	2.714	18.512	0.787	1.363	2.716	15.574	0.766	0.367	2.720	9.438
2	In	0.858	2.456	2.938	15.587	0.850	1.998	2.917	12.511	0.809	-0.137	2.884	6.062
	Z	0.910	4.802	6.434	21.707	0.905	4.352	6.324	17.720	0.882	2.533	6.091	10.039
	$\tilde{Z}$	0.864	3.218	3.497	16.014	0.855	2.734	3.341	13.270	0.820	0.716	2.971	7.055
	$\hat{Z}$	0.890	3.662	5.507	14.307	0.878	2.947	5.183	11.989	0.836	0.338	4.366	6.335
	Z=0	0.777	1.446	-1.798	17.018	0.773	1.342	-1.785	14.174	0.745	0.770	-1.731	8.321

beamformed signal based on the basic method. This suggests better perceptual performance of our method.

For source 1, 2D beamforming without the Z-channel achieves better separation than all methods that use Z-channel information. Since no elevation steering is applied, source 1 (the lower source) is extracted more effectively due to its higher absolute angular distance from the interfering source, which is closer to the beamformer's minima. In contrast, source 2, which has a higher elevation, is extracted less effectively. All methods that use Z-channel information successfully separate the sources, whereas omitting Z-channel information results in worse separation than the omnidirectional input signal. By steering only in the horizontal plane, the interfering source is still located closer to the maximum of the beamformer making it unsuitable for extracting elevated sources.

IV. Conclusion

This paper describes a method for synthesizing a virtual height channel that can be applied to planar microphone arrays in the Ambisonics domain. While the basic method was mentioned in the context of DirAC before [5], where no numerical evaluation was conducted, we show that it is a generally applicable method that yields DoA estimation of elevated sources and improved source separation via beamforming for elevated sources.

Moreover, we presented an extended version that employs diffuseness estimation to recreate the missing Z-channel. The extension is able to decrease the directional spread in terms of DoA estimation encountered using only the plane-wave assumption. Nevertheless, some limitations remain, which could be addressed in future work, such as employing neural network-based estimation of the Z-channel. The accuracy of the method depends on reliable diffuseness estimation, which has not yet been validated independently. Furthermore, the diffuse part only uses the omnidirectional signal, which could be improved by a more accurate estimation of the diffuse component. Also, validation of the method with real recordings from a table-top array should be conducted.

References

[1] B. Rafaely, *Fundamentals of spherical array processing*. Springer, 2015, vol. 8.

[2] S. Wirler, N. Meyer-Kahlen, and V. Pulkki, "Enhancing spatial post-filters through non-linear combinations," in *Audio Engineering Society Convention 157*. Audio Engineering Society, 2024.

[3] M. Frank, F. Zotter, and A. Sontacchi, "Producing 3d audio in ambisonics," in *Audio Engineering Society Conference: 57th International Conference: The Future of Audio Entertainment Technology–Cinema, Television and the Internet*. Audio Engineering Society, 2015.

[4] V. Pulkki, "Spatial sound reproduction with directional audio coding," *Journal of the Audio Engineering Society*, vol. 55, no. 6, pp. 503–516, 2007.

[5] F. Kuech, M. Kallinger, R. Schultz-Amling, G. del Galdo, J. Ahonen, and V. Pulkki, "Directional Audio Coding Using Planar Microphone Arrays," in *2008 Hands-Free Speech Communication and Microphone Arrays*, May 2008, pp. 37–40.

[6] F. Zotter and M. Frank, *Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*, ser. Springer Topics in Signal Processing. Cham: Springer International Publishing, 2019, vol. 19.

[7] C. Nachbar, F. Zotter, E. Deleflie, and A. Sontacchi, "Ambix-a suggested ambisonics format," in *Ambisonics Symposium*, vol. 2011, 2011.

[8] N. Epain and C. T. Jin, "Spherical Harmonic Signal Covariance and Sound Field Diffuseness," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1796–1807, Oct. 2016.

[9] G. H. Golub and C. F. Van Loan, *Matrix computations*. JHU press, 2013.

[10] A. Politis, "Microphone array processing for parametric spatial audio techniques," Ph.D. dissertation, 2016. [Online]. Available: https://aaltodoc.aalto.fi/items/550940c9-65e0-4ee0-adfa-a2a826a41271

[11] S. Tervo, "Direction estimation based on sound intensity vectors," in *2009 17th European Signal Processing Conference*. IEEE, 2009, pp. 700–704.

[12] A. Politis. Polarch/Spherical-Array-Processing. GitHub repository. [Online]. Available: https://github.com/polarch/Spherical-Array-Processing

[13] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, Mar. 2010, pp. 4214–4217.

[14] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.