

# Spatial Audio Signal Enhancement: A Multi-output MVDR Method in The Spherical Harmonic-domain

Huawei Zhang<sup>1</sup>, Jihui (Aimee) Zhang<sup>2,1,3</sup>, Huiyuan (June) Sun<sup>1</sup>, Prasanga Samarasinghe<sup>1</sup>

<sup>1</sup>Audio & Acoustic Signal Processing Group, The Australian National University, Canberra, Australia

<sup>2</sup>School of Electrical Engineering and Computer Science, The University of Queensland, Brisbane, Australia

<sup>3</sup>Institute of Sound and Vibration Research, University of Southampton, Southampton, U.K.

u7178294@anu.edu.au, jihuiaimee.zhang@uq.edu.au, u5870643@anu.edu.au, prasanga.samarasinghe@anu.edu.au

**Abstract**—Spatial audio signal enhancement aims to reduce interfering source contributions while preserving the desired sound field with its spatial cues. Existing methods generally rely on impractical assumptions (e.g. accurate estimations of impractical information) or have limited applicability. This paper presents a spherical harmonic (SH)-domain minimum variance distortionless response (MVDR)-based spatial signal enhancer using Relative Harmonic Coefficients (ReHCs) to extract clean SH coefficients from noisy recordings in reverberant environments. A simulation study shows the proposed method achieves lower estimation error, higher speech-distortion-ratio (SDR), and comparable noise reduction (NR) within the sweet area in a reverberant environment, compared to a beamforming-and-projection method as the baseline.

## I. INTRODUCTION

With the rapid increase of spatial audio applications in the last decade, there has been a growing demand for spatial audio signal enhancement (also known as ambisonic-to-ambisonic separation) [2]–[9], which separates a desired sound field from interference source contributions while preserving the desired spatial cues. This technology acts as a fundamental step in many audio signal processing applications including sound field decomposition [9], sound field recording [10], sound field reproduction [11]–[13], and spatial active noise control [14]–[16].

We categorize existing methods for spatial audio enhancement into three kinds: (i) Beamforming-and-projection methods [2]–[5], (ii) Multi-channel Wiener Filter methods [5], [6], and (iii) Learning-based methods [7]–[9]. Beamforming-and-project methods often include at least two stages: a beamforming stage to capture the desired source signal or the desired sound pressure at a point, and a projection stage to reconstruct the desired sound field using the point-to-region acoustic transfer functions (ATFs). While these solutions have been achieved over space via multi-point processing [3] as well as spherical harmonic (SH) domain processing [2], [4], [5], they make one or more impractical assumptions such as the accurate estimation of ATFs. Multi-channel Wiener Filter methods achieve enhancement by preserving the estimated power spectral density (PSD) matrix of the desired SH coefficients and reducing the estimated PSD matrix of interference

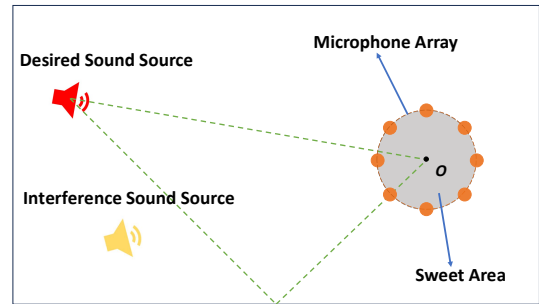


Fig. 1. A desired sound source (red speaker), an interference sound source (yellow speaker), and a spherical microphone array (orange solid dots) at the origin  $O$  in a room environment. The grey area is the spherical sweet area. The green dash lines are examples of the acoustic paths from the desired sound source to the array. The acoustic paths from the interference sound source are omitted.

SH coefficients. However, the accurate estimation of the former PSD matrix is difficult. Learning-based methods, up to today, have required a suitable dataset for pre-training.

Relative Harmonic Coefficients (ReHCs) [17], also named as SH-domain relative transfer functions [18], are denoted as the ratio between SH coefficients and the specific SH coefficient at the 0-th order and the 0-th mode. These ReHCs have been used in beamformer design [18] and spatial audio signal enhancement [2], [5]. These ReHCs can be estimated based on corresponding SH-domain PSD matrices obtained from the received microphone signals [17]. However, the potential of ReHCs in audio enhancement has not been fully studied.

In this paper, we develop an SH-domain minimum variance distortionless response (MVDR) spatial audio signal enhancement method with multiple beamformers based on ReHCs. Motivated by a binaural MVDR beamforming method [19] and a multi-output MVDR beamforming method [3], we extend the conventional SH-domain single-output MVDR beamforming method [18], [20] to be a multi-output variant. This proposed method extracts the SH coefficients due to the desired sound source from the mixed SH domain recording. Different from the conventional single-output MVDR beamforming method (losing the spatial information) and the beamforming-and-projection methods (requiring a projection stage to reconstruct spatial information), this proposed method can directly output a set of SH coefficients representing the extracted desired

This work is sponsored by the Australian Research Council (ARC) Discovery Projects funding scheme with project number DE230101567. We briefly discussed the preliminary idea of this work in [1].

sound field with its spatial cues. Moreover, compared with relying on estimated ATFs in [3], the proposed beamforming method is developed with a more practical foreknowledge of estimated ReHCs, which is used to construct a set of spatial constraints to preserve desired SH coefficients.

## II. PROBLEM FORMULATION

As shown in Fig. 1, consider a far-field scenario with a desired sound source, an interference sound source, with a spherical sweet area with the radius  $r_s$  and a  $Q$ -microphone spherical array of the radius  $r_a$  placed concentrically at the origin  $O$  in a reverberated room. The two sound sources are placed at fixed positions outside the spherical area, causing a desired sound field and an interference sound field within the sweet area, respectively. We assume the desired sound source is uncorrelated with the interference sound source, and the interference sound source is stable.

Each sound field pressure within the sweet area is denoted as  $x(t, k, r, \theta, \phi)$  in the time-frequency (TF) domain, where  $t$  is the time frame,  $k = 2\pi f/c$  is the wavenumber,  $f$  is the frequency,  $c$  is the speed of sound,  $r$ ,  $\theta$ , and  $\phi$  are the radius, the elevation angle, and the azimuth angle, respectively. The sound field pressure  $x(t, k, r, \theta, \phi)$  can be decomposed as

$$x(t, k, r, \theta, \phi) = d(t, k, r, \theta, \phi) + v(t, k, r, \theta, \phi), \quad (1)$$

where  $d(t, k, r, \theta, \phi)$  and  $v(t, k, r, \theta, \phi)$  are the sound pressure of the desired field and the interference field, respectively. The received microphone signal at the  $q$ -th microphone is denoted as

$$\begin{aligned} x_q(t, k) &= x(t, k, r_q, \theta_q, \phi_q) + u(t, k) \\ &= d_q(t, k) + v_q(t, k) + u(t, k), \end{aligned} \quad (2)$$

where  $d_q(t, k)$ ,  $v_q(t, k)$ ,  $u(t, k)$  are the desired microphone signal, the interference microphone signal, and the random sensor noise, respectively,  $r_q$ ,  $\theta_q$ , and  $\phi_q$  are the radius, the elevation angle, and the azimuth angle, respectively, at the  $q$ -th microphone. The microphone signal  $x_q(t, k)$  can be decomposed into SH coefficients to represent the mixed field as [20], [21]

$$x_q(t, k) = \sum_{n=0}^{N_k} \sum_{m=-n}^n \tilde{x}_{nm}(t, k) j_n(kr_q) Y_{nm}(\theta_q, \phi_q), \quad (3)$$

where  $\tilde{x}_{nm}(t, k)$  is the mixed SH coefficient representing the mixed field at the  $n$ -th order and the  $m$ -th mode,  $j_n(\cdot)$  is the Spherical Bessel function of the first kind,  $Y_{nm}(\cdot)$  is the SH function, and  $N_k$  is the corresponding maximum order for the wavenumber  $k$  defined as [20], [21]

$$N_k = \text{ceil}(kr_a) = \text{ceil}\left(\frac{2\pi f}{c} r_a\right), \quad (4)$$

where  $\text{ceil}(\cdot)$  is the ceiling function. For brevity, the time frame  $t$  and the wave number  $k$  will be omitted in the rest of this paper. Obtained  $L = (N + 1)^2$  mixed SH coefficients  $\tilde{\mathbf{x}} = [\tilde{x}_{00}, \tilde{x}_{1-1}, \dots, \tilde{x}_{nm}, \dots, \tilde{x}_{NN}]^T$  can be divided as

$$\tilde{\mathbf{x}} = \text{SHT}(\mathbf{x}) = \text{SHT}(\mathbf{d} + \mathbf{v} + \mathbf{u}) = \tilde{\mathbf{d}} + \tilde{\mathbf{v}} + \tilde{\mathbf{u}}, \quad (5)$$

where the SHT is the Spherical Harmonic Transform based on Eq. (3) [20], [21],  $\mathbf{x} = [x_1, \dots, x_q, \dots, x_Q]^T$  are received microphone signals at the array,  $\mathbf{d}$ ,  $\mathbf{v}$  and  $\mathbf{u}$  are received signals from the desired field, the interference field, and sensor noises, respectively,  $\tilde{\mathbf{d}}$ ,  $\tilde{\mathbf{v}}$  and  $\tilde{\mathbf{u}}$  are SH coefficients of the desired field, the interference field, and sensor noises, respectively.

The task of this work is to extract desired SH coefficients  $\tilde{\mathbf{d}}$  from mixed SH coefficients  $\tilde{\mathbf{x}}$  obtained from the mixture recording  $\mathbf{x}$ .

## III. PROPOSED MULTI-OUTPUT MVDR METHOD

In this section, we use ReHCs to develop an SH-domain multi-output MVDR beamforming method to estimate desired SH coefficients.

The proposed method is a multi-output extension of a SH-domain MVDR beamforming method, consisting of  $L$  beamformers  $\tilde{\mathbf{w}} = [\tilde{\mathbf{w}}_{00}^T, \tilde{\mathbf{w}}_{1-1}^T, \dots, \tilde{\mathbf{w}}_{nm}^T, \dots, \tilde{\mathbf{w}}_{NN}^T]^T$  as a  $L^2 \times 1$  vector [3], [19], [20]. The estimated desired SH coefficient  $\hat{d}_{nm}$  at the  $n$ -th order and the  $m$ -th mode is obtained with the corresponding beamformer  $\tilde{\mathbf{w}}_{nm}$  as

$$\hat{d}_{nm} = \tilde{\mathbf{w}}_{nm}^H \tilde{\mathbf{x}}. \quad (6)$$

The beamformer  $\tilde{\mathbf{w}}_{nm}$  is required to preserve the desired SH coefficient and reduce SH coefficients of interference and sensor noises. As it is difficult to obtain accurate ATFs, similar to [18], [22], we use ReHCs to develop a cost function with a spatial constraint as

$$\begin{aligned} &\underset{\tilde{\mathbf{w}}_{nm}}{\text{minimize}} && \tilde{\mathbf{w}}_{nm}^H \tilde{\mathbf{R}}_{\mathbf{v}+\mathbf{u}} \tilde{\mathbf{w}}_{nm} \\ &\text{subject to} && \tilde{\mathbf{w}}_{nm}^H \tilde{\mathbf{h}} = \tilde{h}_{nm}, \end{aligned} \quad (7)$$

where  $\tilde{\mathbf{R}}_{\mathbf{v}+\mathbf{u}} = \tilde{\mathbf{R}}_{\mathbf{v}} + \tilde{\mathbf{R}}_{\mathbf{u}}$  is the summation of the SH-domain PSD matrices of interference  $\tilde{\mathbf{R}}_{\mathbf{v}} = \mathbb{E}[\tilde{\mathbf{v}}\tilde{\mathbf{v}}^H]$  and that of sensor noises  $\tilde{\mathbf{R}}_{\mathbf{u}} = \mathbb{E}[\tilde{\mathbf{u}}\tilde{\mathbf{u}}^H]$ ,  $\mathbb{E}[\cdot]$  is math expectation,  $\tilde{\mathbf{h}} = [\tilde{h}_{00}, \tilde{h}_{1-1}, \dots, \tilde{h}_{nm}, \dots, \tilde{h}_{NN}]^T$  are ReHCs for the desired field, and  $\tilde{h}_{nm}$  is the corresponding ReHC at the  $n$ -th order and the  $m$ -th mode. These ReHCs  $\tilde{\mathbf{h}}$  can be obtained as [17], [23]

$$\tilde{\mathbf{h}} = \frac{\tilde{\mathbf{d}}}{\tilde{d}_{00}} = \frac{\tilde{\mathbf{R}}_{\mathbf{d}} \tilde{\mathbf{e}}_1}{\mathbf{e}_1^H \tilde{\mathbf{R}}_{\mathbf{d}} \tilde{\mathbf{e}}_1} \approx \frac{\hat{\mathbf{R}}_{\mathbf{d}+\mathbf{u}} \tilde{\mathbf{e}}_1}{\tilde{\mathbf{e}}_1^H \hat{\mathbf{R}}_{\mathbf{d}+\mathbf{u}} \tilde{\mathbf{e}}_1}, \quad (8)$$

where  $\tilde{d}_{00}$  is the desired SH coefficient at the 0-th order and the 0-th mode,  $\tilde{\mathbf{R}}_{\mathbf{d}} = \mathbb{E}[\tilde{\mathbf{d}}\tilde{\mathbf{d}}^H]$  is the SH-domain PSD matrix of desired SH coefficients,  $\tilde{\mathbf{e}}_1 = [1 \quad \mathbf{0}_{1 \times (L-1)}]^T$ ,  $\hat{\mathbf{R}}_{\mathbf{d}+\mathbf{u}}$  is the estimated SH-domain PSD matrix obtained from received microphone signals when only the desired source is active.

Combining cost functions as Eq. (7) in all orders and all modes, a new cost function is derived [19] to design  $\tilde{\mathbf{w}}$  as

$$\begin{aligned} &\underset{\tilde{\mathbf{w}}}{\text{minimize}} && \tilde{\mathbf{w}}^H \tilde{\mathbf{R}}_{\mathbf{v}+\mathbf{u}} \tilde{\mathbf{w}} \\ &\text{subject to} && \tilde{\mathbf{C}}^H \tilde{\mathbf{w}} = \tilde{\mathbf{b}}, \end{aligned} \quad (9)$$

where  $\tilde{\mathbf{b}} = \text{conj}(\tilde{\mathbf{h}})$  is a  $L \times 1$  column vector,  $\text{conj}(\cdot)$  is complex conjugate operation,  $\tilde{\mathbf{R}}_{\mathbf{v}+\mathbf{u}}$  and  $\tilde{\mathbf{C}}$  are two large

matrices constructed by  $\tilde{\mathbf{R}}_{\mathbf{v}+\mathbf{u}}$  and  $\tilde{\mathbf{h}}$ , respectively, as

$$\tilde{\mathbf{R}}_{\mathbf{v}+\mathbf{u}} = \begin{bmatrix} \tilde{\mathbf{R}}_{\mathbf{v}+\mathbf{u}} & & & \\ & \tilde{\mathbf{R}}_{\mathbf{v}+\mathbf{u}} & & \\ & & \dots & \\ & & & \tilde{\mathbf{R}}_{\mathbf{v}+\mathbf{u}} \end{bmatrix}_{L^2 \times L^2}, \quad (10)$$

$$\tilde{\mathbf{C}} = \begin{bmatrix} \tilde{\mathbf{h}} & & & \\ & \tilde{\mathbf{h}} & & \\ & & \dots & \\ & & & \tilde{\mathbf{h}} \end{bmatrix}_{L^2 \times L}. \quad (11)$$

The optimal solution of Eq. (9) can be obtained as [24]

$$\tilde{\mathbf{w}} = \tilde{\mathbf{R}}_{\mathbf{v}+\mathbf{u}}^{-1} \tilde{\mathbf{C}} [\tilde{\mathbf{C}}^H \tilde{\mathbf{R}}_{\mathbf{v}+\mathbf{u}}^{-1} \tilde{\mathbf{C}}]^{-1} \tilde{\mathbf{b}}. \quad (12)$$

Each beamformer  $\tilde{\mathbf{w}}_{nm}$  can be obtained from  $\tilde{\mathbf{w}}$  directly. After that, we can obtain an estimation  $\hat{\tilde{\mathbf{d}}}$  of desired SH coefficients for all orders and all modes with Eq. (6).

Estimated sound field pressures with spatial cues over the sweet area can be further obtained based on the Inverse Spherical Harmonic Transform (ISHT) [20], [21] from the estimated desired SH coefficients  $\hat{\tilde{\mathbf{d}}}$ . However, spatial cues of the residual interference field may not be preserved, similar to [3].

Note the proposed method requires to obtain (1) a time interval of microphone recordings when only the desired source is active, and (2) the summation of the SH-domain PSD matrix of interference and that of sensor noises in advance. The requirement of such a time interval can be waived by estimating the relative transfer matrix [25] instead of ReHCs, but it is out of our scope in this paper.

#### IV. SIMULATION AND RESULTS

In this section, we evaluate the performance of the proposed method compared with a leading-edge beamforming-and-projection method [3] as the baseline in a reverberant room.

The baseline is implemented by three stages: First, we use TF-domain single-output MVDR beamforming [19] to estimate a single-channel desired signal. The beamforming requires prior knowledge of the accurate direction of arrival (DoA) and the summation of the TF-domain PSD matrix of interference and that of sensor noises. Second, we estimate a set of ATFs based on microphone signals and our estimated desired signal, where frequency smooth operation [3] has been performed with 9 frequency bins around each target frequency bin. Third, we implement the projection to obtain the estimated desired microphone signals.

As the output of the baseline is the estimated desired microphone signals, for a fair comparison with the proposed method, the SHT is applied to the output of the baseline to estimate desired SH coefficients. We can then reconstruct the corresponding estimated desired field by applying the ISHT from estimated desired SH coefficients of both the proposed method and the baseline method.

#### A. Setup

A speech signal [26] is used as the desired source at (4.60, 4.05, 1.70) m. A recorded washer-dryer noise signal [27] is used as the interference source at (1.60, 1.05, 1.20) m. We set a 32-mic open spherical array at (1.60, 4.05, 1.70) m, with the radius  $r_a = 0.042$  m. The array has the same size and microphone positions as the em32 Eigenmike spherical microphone array [28]. White Gaussian noises are added to the microphone array as sensor noises in terms of 35 dB signal-sensor noise-ratio (SSNR), where the 'signal' in SSNR refers to the desired source signal. The room size is  $5 \times 6 \times 4$  m and the reverberation time is  $T_{60} = 0.2$  s. The sampling frequency is 16000 Hz. The radius of the sweet area is  $r_s = r_a = 0.042$  m. The speed of sound is 343 m/s. All room impulse responses (RIRs) are simulated with a toolbox using the image source method [29], [30].

We first apply the Short-Time Fourier Transform with the frame size as 16384 and 75% overlap to the time-domain received microphone signals. Then we apply the SHT with the corresponding maximal order  $N_k$  to each frequency bin of obtained TF-domain microphone signals.

For a fair comparison, we simulate accurate PSD matrices of interference signals and sensor noises in the TF domain and the SH domain, facilitating the baseline method and the proposed method, respectively. The PSD matrix of interference signals is calculated based on the simulated interference source signal power and corresponding ATFs from simulated RIRs. The PSD matrix of sensor noises is calculated by the averaged estimated PSD matrix of simulated sensor noises.

We evaluate the performance of both the proposed method and the baseline method among the common telephone bandwidth (from 300 Hz to 3400 Hz) [31], [32] due to: (1) The key information of speech signals exist within the bandwidth; (2) The Spherical Bessel function of the first kind  $j_n(kr)$  will achieve values close to 0 among some higher frequency bins, resulting in the Bessel Zero problem [20]. Therefore, at most 3-order SH coefficients are required based on Eq. (4) in the following simulations. The Bessel Zero problem can be relieved by using the spherical-shell microphone array [33].

#### B. Performance Analysis Based on Sound Field Estimation Error

In Fig. 2, we compare the estimated desired sound fields and evaluate the corresponding estimation errors within the sweet area at one time frame at 1500 Hz on the x-y plane. The signal-noise-ratio (SNR) is 0 dB, where the 'signal' in SNR refers to the desired source signal and the 'noise' in SNR refers to the interference source signal. Here, we denote the normalized square error  $\epsilon(t, k)$  as the estimation error at each point represented by

$$\epsilon(t, k) = 10 \log_{10} \frac{|d_{\text{point}}(t, k) - \hat{d}_{\text{point}}(t, k)|^2}{|d_{\text{point}}(t, k)|^2}, \quad (13)$$

where  $d_{\text{point}}(t, k)$  and  $\hat{d}_{\text{point}}(t, k)$  are the true desired sound field pressure and the estimated desired sound field pressure

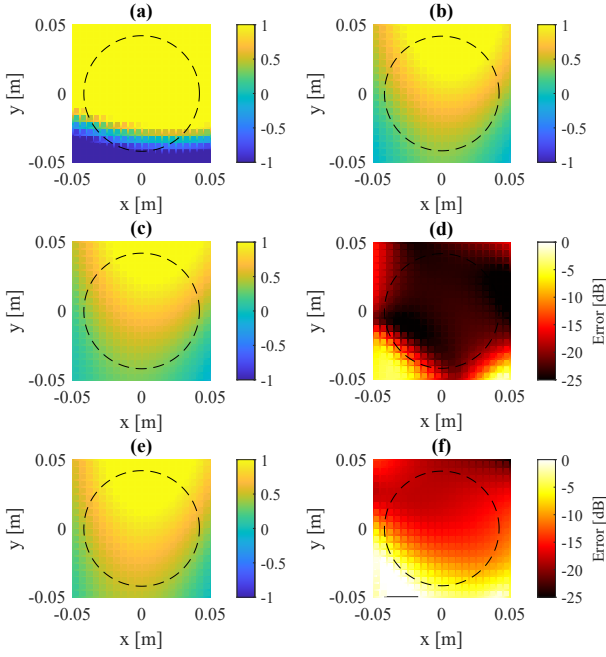


Fig. 2. (a) True mixed field (desired field + interference field); (b) True desired field; (c) Estimated desired field by the proposed method; (d) Estimation error for the proposed method; (e) Estimated desired field by the baseline method; (f) Estimation error for the baseline method at 1500 Hz on the x-y plane. The black dashed circle is the border of the sweet area and the array.

at the point, respectively. Each sound field with corresponding estimation errors defined in Eq. (13) is presented for 441 observation points evenly distributed over the x-y plane. By comparing Fig. 2(b), Fig. 2(c), and Fig. 2(e), we observe that both the proposed method and the baseline method reconstructed a desired field similar to the true desired field within the sweet area. Fig. 2(d) and Fig. 2(f) show that the proposed method achieved a lower estimation error of less than about  $-15$  dB, compared with the baseline method.

### C. Performance Analysis over Frequency

In Fig. 3, we further evaluate the average of the sound field estimation error, the speech-distortion-ratio (SDR), and the noise reduction (NR) within the spherical sweet area across 15 time frames against different signal frequencies. The evaluation is based on corresponding sound pressures at 107 observation points evenly distributed over the spherical sweet area. Here, estimation error, SDR, and NR at these 107 observation points are defined as [5]

$$\text{Error}^{\text{ob}}(t, k) = 10 \log_{10} \frac{\|\mathbf{d}_s(t, k) - \hat{\mathbf{d}}_s(t, k)\|^2}{\|\mathbf{d}_s(t, k)\|^2}, \quad (14)$$

$$\text{SDR}^{\text{ob}}(t, k) = 10 \log_{10} \frac{\|\mathbf{d}_s(t, k)\|^2}{\|\mathbf{d}_s^{\text{res}}(t, k) - \mathbf{d}_s(t, k)\|^2}, \quad (15)$$

$$\text{NR}^{\text{ob}}(t, k) = 10 \log_{10} \frac{\|\mathbf{v}_s(t, k)\|^2}{\|\mathbf{v}_s^{\text{res}}(t, k) + \mathbf{u}_s^{\text{res}}(t, k)\|^2}, \quad (16)$$

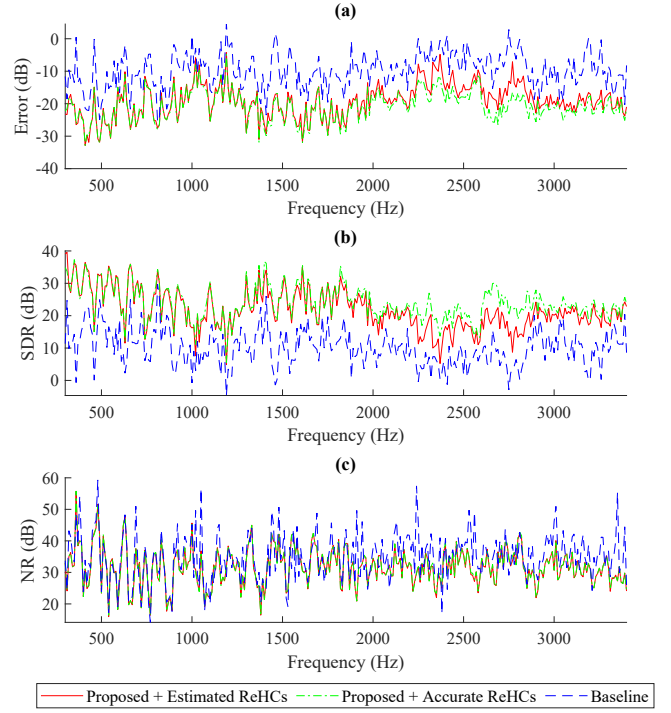


Fig. 3. The average of estimation error (a), SDR (b), and NR (c) of the proposed with estimated ReHCs (red), the proposed with accurate ReHCs (green), and the baseline (blue) inside the spherical sweet area among frequency bins from 300 Hz to 3400 Hz.

where  $\mathbf{d}_s(t, k)$ ,  $\hat{\mathbf{d}}_s(t, k)$ , and  $\mathbf{v}_s(t, k)$  are true desired field pressures, estimated desired field pressures, and true interference field pressures, respectively,  $\mathbf{d}_s^{\text{res}}(t, k)$  and  $\mathbf{v}_s^{\text{res}}(t, k)$  are residual desired sound field pressures and residual interference sound field pressures after processing, respectively,  $\mathbf{u}_s^{\text{res}}(t, k)$  is representing the influence of residual sensor noises, at these observation points within the spherical sweet area, and  $\|\cdot\|$  denotes the 2-norm operation. As shown in Fig. 3, the proposed method with estimated ReHCs achieved lower estimation error, higher SDR, and comparable NR (more than 25 dB NR is effective enough) than the baseline method ranging from 300 Hz to 3400 Hz. In detail, the proposed with estimated ReHCs achieved a lower than  $-15$  dB estimation error and a higher than 15 dB SDR at the majority of chosen frequency bins, which outperforms the baseline.

Fig. 3 also shows the influence of the accuracy of ReHCs. As a comparison, we play a segment of 0 dB white Gaussian noise instead of the original desired source signal at the same source position, and then use the same method in Sec. IV-A to obtain another set of ReHCs as accurate ReHCs. These ReHCs are more accurate as white Gaussian noise is more stable than speech signals at different frequency bins. In Fig. 3, we can find the proposed method with accurate ReHCs outperformed that with estimated ReHCs in terms of estimation error and SDR, and achieved comparable NR performance, which implies the performance of the proposed method is influenced by the accuracy of ReHCs estimation.

TABLE I

THE AVERAGED ESTIMATION ERROR, SDR, AND NR INSIDE THE SPHERICAL SWEET AREA AMONG FREQUENCY BINS FROM 300 HZ TO 3400 HZ ACROSS 15 TIME FRAMES AGAINST DIFFERENT REVERBERATION TIMES  $T_{60}$ : (A) THE PROPOSED METHOD; (B) THE BASELINE METHOD.

	$T_{60} = 0$ s			$T_{60} = 0.2$ s			$T_{60} = 0.4$ s		
	Error (dB)	SDR (dB)	NR (dB)	Error (dB)	SDR (dB)	NR (dB)	Error (dB)	SDR (dB)	NR (dB)
(A) The proposed	-20.7	24.0	32.3	-19.3	21.8	31.7	-16.3	19.3	28.3
(B) The baseline	-14.7	14.7	73.2	-9.8	10.2	34.9	-6.3	7.0	27.9

TABLE II

THE AVERAGED ESTIMATION ERROR, SDR, AND NR INSIDE THE SPHERICAL SWEET AREA AMONG FREQUENCY BINS FROM 300 HZ TO 3400 HZ ACROSS 15 TIME FRAMES AGAINST DIFFERENT SNR LEVELS: (A) THE PROPOSED METHOD; (B) THE BASELINE METHOD.

	SNR = 5 dB			SNR = 0 dB			SNR = -5 dB		
	Error (dB)	SDR (dB)	NR (dB)	Error (dB)	SDR (dB)	NR (dB)	Error (dB)	SDR (dB)	NR (dB)
(A) The proposed	-19.9	21.8	28.5	-19.3	21.8	31.7	-18.2	21.8	34.0
(B) The baseline	-11.8	12.0	35.2	-9.8	10.2	34.9	-7.1	7.6	34.4

#### D. Performance Analysis over Varying Reverberation Times and SNR Levels

Here, we evaluate the performance of both the proposed method with estimated ReHCs and the baseline method for varying reverberation times  $T_{60}$  and varying SNRs, averaging within the spherical sweet area at the bandwidth for 15 time frames. As Sec. IV-C, three indexes are evaluated at these 107 observation points.

Firstly, we evaluate the performance of these two methods over varying reverberation times  $T_{60}$  as shown in TABLE I. Here, we vary the  $T_{60}$  from 0 s to 0.4 s, while the SNR is set to be 0 dB. With the reverberation time  $T_{60}$  increased, both methods achieved higher estimation error, lower SDR, and lower NR. In addition, the proposed method always achieved lower estimation error, higher SDR, and comparable NR than the baseline method for varying  $T_{60}$  (more than 25 dB NR is effective enough).

Secondly, we evaluate the performance of these two methods over varying SNR levels as shown in TABLE II. Here, we vary the SNR level from 5 dB to -5 dB, and remain the  $T_{60}$  as 0.2 s. TABLE II shows that the proposed method maintained stable performance for varying SNR levels in terms of estimation error (about -19 dB) and SDR (about 22 dB), always outperforming the baseline method. By contrast, with the SNR level decreased, both estimation error and SDR achieved by the baseline method declined. In addition, both the proposed method and the baseline method achieved comparable and stable NR (more than about 28 dB) for varying SNR levels.

#### V. CONCLUSION

In this paper, we propose a spherical harmonic (SH)-domain minimum variance distortionless response (MVDR) method to estimate the desired sound field from the mixture recording at a spherical microphone array. We use a cost function with a set of spatial constraints to extract desired SH coefficients and suppress SH coefficients of interference and sensor noises. The field due to the desired sound source within the sweet area can be hence reconstructed. Simulation results show that the proposed method can outperform the baseline methods within the sweet area in a reverberant room. In the future, we plan to

theoretically analyze the influence of ReHC accuracy on the performance of our proposed method.

#### REFERENCES

- [1] H. Zhang, J. Zhang, F. Ma, H. Sun, and P. N. Samarasinghe, "A directional spatial active noise control system with a sound field separation algorithm," *J. Acoust. Soc. Am.*, vol. 154, no. 4\_supplement, A162–A162, 2023.
- [2] Y. Hu, P. N. Samarasinghe, and T. D. Abhayapala, "Acoustic signal enhancement using relative harmonic coefficients: Spherical harmonics domain approach," en, in *Interspeech 2020*, ISCA, 2020, pp. 5076–5080.
- [3] M. Lugasi, A. Menon, V. Tourbabin, and B. Rafaely, "Spatial audio signal enhancement by a two-stage source—system estimation with frequency smoothing for improved perception," en, *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 2584–2596, 2022, ISSN: 2329-9290, 2329-9304.
- [4] C. Borrelli, A. Canclini, F. Antonacci, A. Sarti, and S. Tubaro, "A denoising methodology for higher order ambisonics recordings," in *2018 Int. Workshop Acoust. Signal Enhancement (IWAENC)*, IEEE, 2018, pp. 451–455.
- [5] A. Herzog and E. A. P. Habets, "Direction and reverberation preserving noise reduction of ambisonics signals," en, *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 2461–2475, 2020, ISSN: 2329-9290, 2329-9304.
- [6] M. Hafsati, N. Epain, R. Gribonval, and N. Bertin, "Sound source separation in the higher order ambisonics domain," in *2019 Int. Conf. Digit. Audio Eff. (DAFx)*, 2019, pp. 1–7.
- [7] A. Herzog, S. R. Chetupalli, and E. A. P. Habets, "Ambisep: Ambisonic-to-ambisonic reverberant speech separation using transformer networks," en, in *2022 Int. Workshop Acoust. Signal Enhancement (IWAENC)*, IEEE, 2022, pp. 1–5, ISBN: 978-1-66546-867-1.
- [8] A. Herzog, S. R. Chetupalli, and E. A. P. Habets, "Ambisep: Joint ambisonic-to-ambisonic speech separation and noise reduction," en, *IEEE/ACM Trans. Audio*

- Speech Lang. Process.*, pp. 1–13, 2023, ISSN: 2329-9290, 2329-9304.
- [9] R. Matsuda and M. Otani, “Sound field decomposition based on two-stage neural networks,” *arXiv preprint arXiv:2309.06661*, 2023.
- [10] W. Zhang, P. Samarasinghe, H. Chen, and T. Abhayapala, “Surround by sound: A review of spatial audio recording and reproduction,” en, *Appl. Sci.*, vol. 7, no. 5, p. 532, 2017, ISSN: 2076-3417.
- [11] P.-A. Gauthier and A. Berry, “Adaptive wave field synthesis for sound field reproduction: Theory, experiments, and future perspectives,” in *Audio Eng. Soc. Conv. 123*, Audio Engineering Society, 2007.
- [12] S. Spors, H. Wierstorf, A. Raake, F. Melchior, M. Frank, and F. Zotter, “Spatial sound with loudspeakers and its perception: A review of the current state,” en, *Proc. IEEE*, vol. 101, no. 9, pp. 1920–1938, 2013, ISSN: 0018-9219, 1558-2256.
- [13] J. Yang, M. Wu, and L. Han, “A review of sound field control,” *Appl. Sci.*, vol. 12, no. 14, p. 7319, 2022.
- [14] J. Zhang, T. D. Abhayapala, W. Zhang, P. N. S., and S. Jiang, “Active noise control over space: A wave domain approach,” en, *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 4, pp. 774–786, 2018, ISSN: 2329-9290, 2329-9304.
- [15] F. Ma, W. Zhang, and T. D. Abhayapala, “Active control of outgoing broadband noise fields in rooms,” en, *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 529–539, 2020, ISSN: 2329-9290, 2329-9304.
- [16] D. Shi, B. Lam, W.-S. Gan, J. Cheer, and S. J. Elliott, “Active noise control in the new century: The role and prospect of signal processing,” in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, Institute of Noise Control Engineering, vol. 268 issue 3, 2023, pp. 5141–5151.
- [17] Y. Hu, P. N. Samarasinghe, S. Gannot, and T. D. Abhayapala, “Semi-supervised multiple source localization using relative harmonic coefficients under noisy and reverberant environments,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 3108–3123, 2020.
- [18] Y. Peled and B. Rafaely, “Linearly-constrained minimum-variance method for spherical microphone arrays based on plane-wave decomposition of the sound field,” en, *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 12, pp. 2532–2540, 2013, ISSN: 1558-7916, 1558-7924.
- [19] E. Hadad, D. Marquardt, S. Doclo, and S. Gannot, “Theoretical analysis of binaural transfer function mvdr beamformers with interference cue preservation constraints,” en, *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 12, pp. 2449–2464, 2015, ISSN: 2329-9290, 2329-9304.
- [20] B. Rafaely, *Fundamentals of Spherical Array Processing* (Springer Topics in Signal Processing), en. Springer Berlin Heidelberg, 2015, vol. 8. DOI: 10.1007/978-3-662-45664-4.
- [21] E. G. Williams, *Fourier acoustics: sound radiation and nearfield acoustical holography*. Academic press, 1999.
- [22] E. Hadad, S. Doclo, and S. Gannot, “The binaural LCMV beamformer and its performance analysis,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 3, pp. 543–558, 2016.
- [23] S. Markovich-Golan and S. Gannot, “Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method,” en, in *2015 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, IEEE, 2015, pp. 544–548, ISBN: 978-1-4673-6997-8.
- [24] B. D. Van Veen and K. M. Buckley, “Beamforming: A versatile approach to spatial filtering,” en, *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4–24, 1988, ISSN: 0740-7467.
- [25] T. D. Abhayapala, L. Birnie, M. Kumar, D. Grixti-Cheng, and P. N. Samarasinghe, “Generalizing the relative transfer function to a matrix for multiple sources and multichannel microphones,” en, in *2023 31st Eur. Signal Process. Conf. (EUSIPCO)*, IEEE, 2023, pp. 336–340, ISBN: 978-94-6459-360-0.
- [26] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An asr corpus based on public domain audio books,” in *2015 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, IEEE, 2015, pp. 5206–5210.
- [27] C. K. A. Reddy, E. Beyrami, J. Pool, R. Cutler, S. Srinivasan, and J. Gehrke, “A scalable noisy speech dataset and online subjective test framework,” en, in *Interspeech 2019*, ISCA, 2019, pp. 1816–1820.
- [28] M. H. Acoustics, “Em32 eigenmike microphone array release notes (v17. 0),” *25 Summit Ave, Summit, NJ 07901, USA*, 2013.
- [29] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-room acoustics,” en, *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.
- [30] D. P. Jarrett, E. A. P. Habets, M. R. P. Thomas, and P. A. Naylor, “Rigid sphere room impulse response simulation: Algorithm and applications,” *J. Acoust. Soc. Am.*, vol. 132, no. 3, pp. 1462–1472, 2012.
- [31] W. Hioki, *Telecommunications*. Prentice Hall, 1998.
- [32] H. Pulakka, “Development and evaluation of artificial bandwidth extension methods for narrowband telephone speech,” Ph.D. dissertation, Aalto University, 2013.
- [33] B. Rafaely, “The spherical-shell microphone array,” *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, no. 4, pp. 740–747, 2008.