Interspeech 2024
1-5 September 2024, Kos, Greece

Spatial Acoustic Enhancement Using Unbiased Relative Harmonic Coefficients

Liang Tao[1], *Maoshen Jia*[1,*], *Yonggang Hu*[2], *Changchun Bao*[1]

[1]Speech and Audio Signal Processing Laboratory, Beijing University of Technology, Beijing, China
[2]Shanghai branch of the Southwest Institute of Electronics and Telecommunication Technology of China, Shanghai, China

liangtaobjut@163.com, jiamaoshen@bjut.edu.cn,
yongganghu@outlook.com, baochch@bjut.edu.cn

Abstract

This paper targets at enhancing the noisy soundfield over the entire recording area and all the individual channels, while preserving the spatial clues of the original soundfield. For the goal, we utilize a recently proposed spherical harmonics (SH) domain feature denoted *relative harmonic coefficients* (RHC) as it compactly contains the source's spatial information. Specifically, we (i) propose an unbiased estimator of RHC in noisy environments; (ii) estimate the source signal in noisy environments using a SH domain beamformer; (iii) enhance the SH coefficients by multiplying the estimated RHC and source signal; and (iv) reconstruct the entire soundfield based on the enhanced SH coefficients. Finally, we evaluate and validate the performance of the enhancement algorithm using extensive simulations.

Index Terms: spatial acoustic enhancement, spherical harmonics domain, relative harmonic coefficients

1. Introduction

In practical application scenarios such as speech communication and human-computer interaction, noise or interfering sources noise would disrupt the target signals, thereby degrading speech quality and intelligibility. Fortunately, this negative effect can be effectively mitigated with the multi-channel speech enhancement [1, 2], which has been an active topic in acoustic signal processing, aiming at recovering the clean speech signal from the noisy recordings [3, 4].

Over the past few decades, we have witnessed notable advancements in enhancing the noisy signal using planar microphone arrays (e.g., linear and circular arrays) [5, 6, 7, 8]. For these methods, the multi-channel signals are directly processed by steering a spatial filter (beamformer), then the source signal on the microphones is enhanced. In contrast, spherical microphone array, capturing the three dimension (3-D) soundfiled, is more suitable as it decomposes the soundfield into the spherical harmonics (SH) domain [9, 10, 11]. The spherical arrays encode the sound pressure into the SH domain first, then filter and sum the SH coefficients [12, 13]. Early methods such as [14, 15] focus on only the enhancement of the received signal at the origin of the array, losing the spatial clues of the desired soundfield. A solution to this problem is the directional audio coding [16, 17], extracting the source signal and spatial parameters independently during the recording phase, then these clues are used to synthesize desired acoustic signal. Another solution in [18] is to use a least mean squares filter to denoise the SH coefficients concerning the entire soundfield, while assuming the source DOA is known as prior knowledge. To overcome the drawback relying on an additional localization technique, Hu et al. proposed to enhance the whole soundfiel using relative harmonic coefficients (RHC) [19], as this feature compactly contains the source spatial information [9, 20, 21]. However, the RHC is estimated biasedly as it ignores the impact of the noise power spectral density (PSD). Additionally, the source signal is obtained by utilizing a fixed SH-domain beamformer limited to the far-field scenarios [22, 23], thus makes it is suboptimal.

In contrast, this paper presents an improved solution concerning the RHC based enhancement algorithm. According to the definition of the SH coefficient, we can decompose it into the product of RHC and source signal. Consequently, it is essential to estimate these components as accurate as possible. Compared to the RHC estimation method developed in [19], the proposed scheme is more robust, as the noise statistics is taken into consideration, leading to an approximate unbiased RHC estimator. Besides, an accurate RHC estimates further enhances estimating the source signal. It is worth mentioning that the method not only recovers the individual channels, but enhances the entire 3-D soundfield, as it estimate the SH coefficients of the whole soundfield order rather than the source signal at the array origin. In the sequel, we first introduce the system model. Then, we present the improved spatial acoustic enhancement method in Section 3. In section 4, the enhancement performance is evaluated and verified. Finally, section 5 concludes this work.

2. Preliminaries

2.1. Signal model

Consider a spherical microphone array consisting of J omnidirectional sensors with the spherical coordinates of each sensor $(r, \varphi_j, \vartheta_j)$, where r is the radius of array, φ_j and ϑ_j are the respective azimuth and elevation in the range of $(0, 2\pi]$ and $[0, \pi]$ (see Figure 1). The signal received by the j-th channel in the short-time Fourier transform (STFT) domain is modeled as,

$$X_j(l,k) = S(l,k)H_j(l,k) + V_j(l,k), \tag{1}$$

where $l \in \{1, 2, \ldots, L\}$ and $k \in \{1, 2, \ldots, K\}$ are the time frame and frequency indices, respectively; $X_j(l,k)$, $S(l,k)$, $H_j(l,k)$ and $V_j(l,k)$ denote the STFT coefficients of the measured sound pressure, the acoustic impulse response from the sound source to the j-th channel, and the additive noise at the j-th channel. Note that the source signal is typically assumed to be uncorrelated with the noise.

2.2. Spherical harmonics decomposition of sound pressure

The sound pressure in (1) can be decomposed into the SH domain [9],

$$X_j(l,k) = \sum_{n=0}^{N} \sum_{m=-n}^{n} \alpha_{nm}(l,k)\beta_n(kr)Y_{nm}(\varphi_j, \vartheta_j), \tag{2}$$

10.21437/Interspeech.2024-141

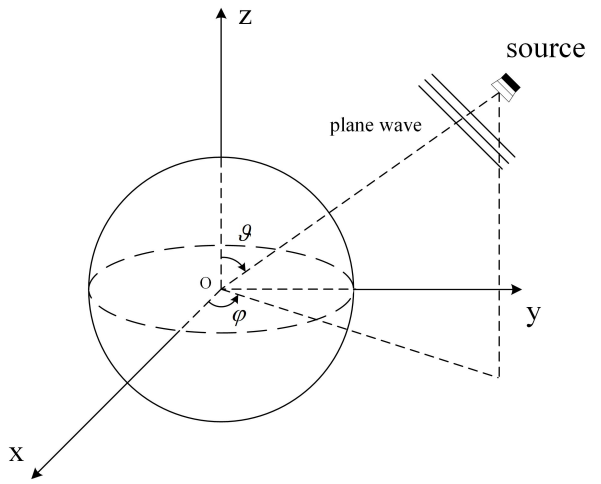

Figure 1: *Illustration of a spherical microphone array for sound pressure measurement.*

where $N = \lceil kr \rceil$ is the truncation order of soundfield, $\lceil \cdot \rceil$ denotes the ceiling operator, $k = 2\pi f/c$ is the wave number, f is the frequency, c is the velocity of sound, and $\beta_n(\cdot)$ is a function dependent on the array configuration [24],

$$\beta_n(\varepsilon) = \begin{cases} j_n(\varepsilon), & \text{for open array} \\ j_n(\varepsilon) - \frac{j_n^{'}(\varepsilon)}{h_n^{'}(\varepsilon)} h_n(\varepsilon), & \text{for rigid array} \end{cases} \tag{3}$$

where $j_n^{'}(\cdot)$ and $h_n^{'}(\cdot)$ denote the first derivative of spherical Bessel and Hankel functions, respectively. The term $Y_{nm}(\varphi_j, \vartheta_j)$ in (2) is the SH function with order n and mode m, defined as,

$$Y_{nm}(\varphi, \vartheta) = \sqrt{\frac{(2n+1)}{4\pi} \frac{(n-m)!}{(n+m)!}} P_{nm}(\cos\vartheta) e^{im\varphi}, \tag{4}$$

where $P_{nm}(\cdot)$ is the associated Legendre polynomial, and,

$$\alpha_{nm}(l,k) = \frac{1}{\beta_n(kr)} \sum_{j=1}^{J} w_j X_j(l,k) Y_{nm}^*(\varphi_j, \vartheta_j), \tag{5}$$

where $\alpha_{nm}(l,k)$ denotes the SH coefficients containing the soundfield information withing the recording area, w_j is the sampling weights designed to maximize the validity of (2), and $(\cdot)^*$ denotes the conjugate operator.

2.3. Problem formulation

Decomposing the noisy multi-channel pressure into the SH domain, the SH coefficients also denote a combination of source signal and noise,

$$\begin{aligned} \alpha_{nm}(l,k) &= \tilde{\alpha}_{nm}(l,k) + v_{nm}(l,k) \\ &= \lambda_{nm}(l,k)\tilde{\alpha}_{00}(l,k) + v_{nm}(l,k) \end{aligned} \tag{6}$$

where $\tilde{\alpha}_{nm}(l,k)$ and $v_{nm}(l,k)$ represent the SH coefficients of desired signal and noise. The term $\lambda_{nm}(l,k)$ denotes the RHC,

$$\lambda_{nm}(l,k) = \tilde{\alpha}_{nm}(l,k)/\tilde{\alpha}_{00}(l,k). \tag{7}$$

The aim by this paper is to estimate $\tilde{\alpha}_{nm}(l,k)$ from the $(N+1)^2$ noisy SH coefficients, $\alpha_{nm}(l,k)$. From (6) we know that the coefficient of desired signal, $\tilde{\alpha}_{nm}(l,k)$, can be written as the product of $\lambda_{nm}(l,k)$ and $\tilde{\alpha}_{00}(l,k)$. Hence, we divide the signal denoising task into three steps, (i) estimating RHC, $\lambda_{nm}(l,k)$; (ii) estimating the desired source signal, $\tilde{\alpha}_{00}(l,k)$; (iii) Synthesizing the desired signal of any channel by substituting the estimated $\tilde{\alpha}_{nm}(l,k)$ to $\alpha_{nm}(l,k)$ in (2).

3. Proposed Algorithm

The proposed method comprises of three steps elaborated in detail below.

3.1. Estimation of RHC

Assume the source signal is nonstationary over a short time segment, while the acoustic transfer function and the statistics of the noise signal is stationary [25, 26]. And, such properties also hold in the SH domain, thus we rewrite the second row in (6),

$$\alpha_{nm}(l) = \lambda_{nm}(l)\alpha_{00}(l) + \gamma_{nm}(l), \tag{8}$$

where

$$\gamma_{nm}(l) = v_{nm}(l) - \lambda_{nm}(l,k)v_{00}(l). \tag{9}$$

The cross-PSD at the l-th frame is given as,

$$\Phi_{\alpha_{nm}\alpha_{00}}(l) = \lambda_{nm}(l)\Phi_{\alpha_{00}\alpha_{00}}(l) + \Phi_{\gamma_{nm}\alpha_{00}}(l), \tag{10}$$

where $\Phi_{\alpha_{nm}\alpha_{00}}(l) = \mathbb{E}[\alpha_{nm}(l)\alpha_{00}^*(l)]$, and $\mathbb{E}[\cdot]$ denotes the mathematical expectation operator. Note that the frequency index k is omitted in this subsection for notational convenience. We estimate the cross-PSD at the l-th frame using a short-time average over a time segment, where a time segment is composed of R successive frames,

$$\widehat{\Phi}_{\alpha_{nm}\alpha_{00}}(l) = \lambda_{nm}(l)\widehat{\Phi}_{\alpha_{00}\alpha_{00}}(l) + \Phi_{\gamma_{nm}\alpha_{00}}(l) + \epsilon_{nm}(l), \tag{11}$$

where

$$\begin{aligned} \epsilon_{nm}(l) &= \widehat{\Phi}_{\gamma_{nm}\alpha_{00}}(l) - \Phi_{\gamma_{nm}\alpha_{00}}(l) \\ \widehat{\Phi}_{\alpha_{nm}\alpha_{00}}(l) &= \frac{1}{2l_0+1} \sum_{l'=l-l_0}^{l'=l+l_0} \alpha_{nm}(l')\alpha_{00}^*(l') \\ \widehat{\Phi}_{\alpha_{00}\alpha_{00}}(l) &= \frac{1}{2l_0+1} \sum_{l'=l-l_0}^{l'=l+l_0} \alpha_{00}(l')\alpha_{00}^*(l') \end{aligned} \tag{12}$$

where $2l_0 + 1 = R$, then (11) can be reformulated in a matrix form within R time frames,

$$\begin{bmatrix} \widehat{\Phi}_{\alpha_{nm}\alpha_{00}}(1) \\ \widehat{\Phi}_{\alpha_{nm}\alpha_{00}}(2) \\ \vdots \\ \widehat{\Phi}_{\alpha_{nm}\alpha_{00}}(R) \end{bmatrix} = \begin{bmatrix} \widehat{\Phi}_{\alpha_{00}\alpha_{00}}(1) & 1 \\ \widehat{\Phi}_{\alpha_{00}\alpha_{00}}(2) & 1 \\ \vdots & \vdots \\ \widehat{\Phi}_{\alpha_{00}\alpha_{00}}(R) & 1 \end{bmatrix} \begin{bmatrix} \lambda_{nm}(l) \\ \Phi_{\gamma_{nm}\alpha_{00}}(l) \end{bmatrix} + \begin{bmatrix} \epsilon_{nm}(1) \\ \epsilon_{nm}(2) \\ \vdots \\ \epsilon_{nm}(R) \end{bmatrix} \tag{13}$$

The optimal solution of RHC can be obtained by applying a weighted least-squares to the above set of overdetermined equation [27],

$$\widehat{\lambda}_{nm}(l) = \frac{(\overline{\widehat{\Phi}_{\alpha_{nm}\alpha_{00}}(l)\widehat{\Phi}_{\alpha_{nm}\alpha_{00}}(l)}) - (\overline{\widehat{\Phi}_{\alpha_{00}\alpha_{00}}(l)})(\overline{\widehat{\Phi}_{\alpha_{nm}\alpha_{00}}(l)})}{(\overline{\widehat{\Phi}^2_{\alpha_{00}\alpha_{00}}(l)}) - (\overline{\widehat{\Phi}_{\alpha_{00}\alpha_{00}}(l)})^2}, \tag{14}$$

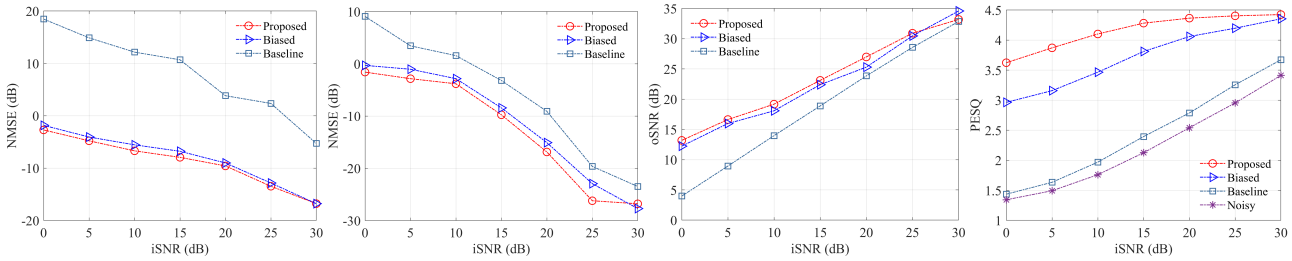

Figure 2: *The results under different iSNR, where the first two subgraphs are the NMSE of RHC vector and source signal, the last two subgraphs are the oSNR and PESQ scores of the enhanced signal.*

where

$$(\overline{\hat{\Phi}(l)}) = \frac{1}{2l_0+1}\sum_{l'=l-l_0}^{l'=l+l_0}\hat{\Phi}(l'). \tag{15}$$

Hence, an estimated N-order RHC vector can be obtained,

$$\widehat{\boldsymbol{\lambda}}_N(l,k) = [1, \widehat{\lambda}_{1,-1}(l,k), ..., \widehat{\lambda}_{NN}(l,k)]^T, \tag{16}$$

where $(\cdot)^T$ denotes the transpose operator.

3.2. Estimation of source signal

This subsection estimates the source signal at the array origin, which coincidentally denotes the zero-order SH coefficient [19]. Hence, the desired source signal can be estimated by employing a SH-domain beamformer, such as maximum signal-to-noise ratio, wiener, and minimum variance distortionless response (MVDR) filters. Here, we select the MVDR filter as the unbiased RHC vector has been given in (16). To minimize the zero-order SH coefficient of residual noise power while constraining the distortion of the desired signal as minimally as possible, the MVDR beamforer is modeled as [15, 28],

$$\min_{\mathbf{h}} \mathbf{h}^H\mathbf{R}_{vv}\mathbf{h} \ \text{ s.t. } \ \mathbf{h}^H\boldsymbol{\lambda}_N = 1. \tag{17}$$

With Lagrange multiplier, the solution of (17) is given as,

$$\mathbf{h} = \frac{\mathbf{R}_{vv}^{-1}\boldsymbol{\lambda}_N}{\boldsymbol{\lambda}_N^H\mathbf{R}_{vv}^{-1}\boldsymbol{\lambda}_N}, \tag{18}$$

where $(\cdot)^{-1}$ and $(\cdot)^H$ denote the inverse and conjugate transpose operators of a matrix, respectively.

According to (18), we know that the beamformer relies on both noise covariance matrix and RHC vector. The former term can typically be estimated using a reliable noise estimator developed in [29] or [30], while this is beyond the scope by this paper. We directly compute the matrix from the noise only signal using a recursive method [31],

$$\mathbf{R}_{vv}(l,k) = \xi\mathbf{R}_{vv}(l-1,k) + (1-\xi)\boldsymbol{v}_N(l,k)\boldsymbol{v}_N^H(l,k), \tag{19}$$

where $\xi \in (0,1)$ is the forgetting factor, which manages the impact of prior data samples on the current estimate, and an optimal parameter is determined through experiments in section 4. We then estimate the source signal by,

$$\widehat{\widetilde{\alpha}}_{00}(l,k) = \mathbf{h}^H(l,k)\boldsymbol{\alpha}_N(l,k), \tag{20}$$

where

$$\begin{aligned}\boldsymbol{\alpha}_N(l,k) &= [\alpha_{00}(l,k), \alpha_{1,-1}(l,k), ..., \alpha_{NN}(l,k)]^T \\ \boldsymbol{v}_N(l,k) &= [v_{00}(l,k), v_{1,-1}(l,k), ..., v_{NN}(l,k)]^T\end{aligned} \tag{21}$$

3.3. Desired signal estimation

Multiplying the estimated RHC vector in (16) with the enhanced source signal in (20), we can obtain the SH coefficient vector up to the entire N-th order, i.e.,

$$\begin{aligned}\widehat{\boldsymbol{\alpha}}_N(l,k) &= \widehat{\boldsymbol{\lambda}}_N(l,k)\widehat{\widetilde{\alpha}}_{00}(l,k) \\ &= [\widehat{\alpha}_{00}(l,k), \widehat{\alpha}_{1,-1}(l,k), ..., \widehat{\alpha}_{NN}(l,k)]^T\end{aligned} \tag{22}$$

Then, the clean sound pressure at (l,k) is reconstructed as,

$$\widehat{X}_j(l,k) = \sum_{n=0}^{N}\sum_{m=-n}^{n}\widehat{\alpha}_{nm}(l,k)\beta_n(kr)Y_{nm}(\varphi_j,\vartheta_j). \tag{23}$$

Repeat the above procedures at bin level over entire STFT domain, the desired sound pressure of any channel, as well as the soundfield radiated by sound source can be reconstructed. Without losing generality, this paper estimates the enhanced speech signal of the first microphone by employing the inverse STFT to recover time domain signal.

4. Simulations

This section validates the performance of the proposed spatial acoustic enhancement method under diverse noisy environments. We simulate an anechoic room with size of $6\ \text{m} \times 4\ \text{m} \times 3\ \text{m}$, where an open spherical microphone array with 32 channels (radius 0.042 m) is placed 1 m away from the speaker, the room impulse response (RIR) is generated using an available tool based on image-source method [32]. The clean speech signals are randomly selected from the Nippon Telegraph and Telephone (NTT) corporation database [33], which are downsampled from 16 kHz to 8 kHz. Subsequently, the recordings are obtained by convolving RIRs with the speech signals, and Gaussian noise is added to each individual channel. In the parameter setting of converting time domain signal to frequency domain using STFT, the length of hamming window is 256 samples, the overlap rate is 75%, and the number of samples for the fast Fourier transform (FFT) is the same as the window size. Note that l_0 in (12) is set to 2 in this study.

Table 1: *The results of different forgetting factor for the enhanced signal*

ξ	0.1	0.3	0.5	0.7	0.9
NMSE (dB)	-1.46	-5.81	-8.02	-8.58	-8.12
oSNR (dB)	4.45	13.27	17.78	19.18	19.69
PESQ	2.75	3.63	3.97	4.10	3.46

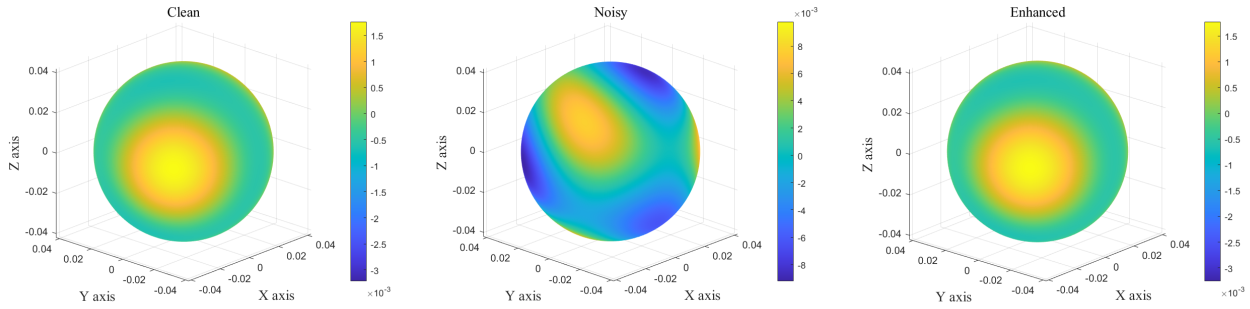

Figure 3: *Clean, noisy and enhanced 3-D soundfiled, where iSNR = 5 dB and* $f = 2\ kHz$.

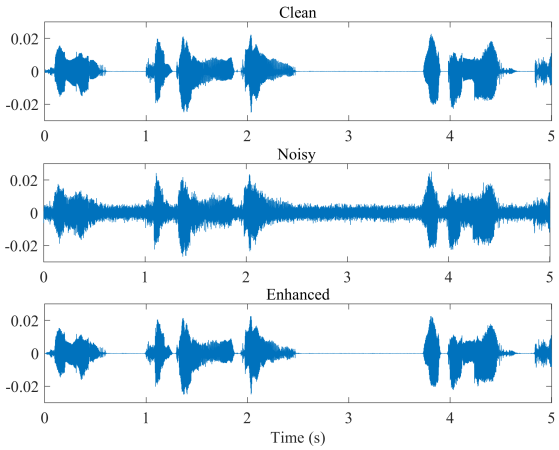

Figure 4: *Clean, noisy and enhanced time-domain speech signal in the case of 5 dB noise.*

We compare the performance of the proposed method with two reference methods, one is the biased RHC estimation-based method proposed in [19], and the method without RHC estimation (i.e., RHC is obtained directly by dividing the noise signal) as baseline. The normalized mean squared error (NMSE) is used to measure the accuracy of the original and estimated values.

$$\text{NMSE} = 10\log_{10}\frac{||S-\widehat{S}||^2}{||S||^2}, \tag{24}$$

where $||\cdot||$ represents the l_2 norm, S stands for clean RHC vector, clean source signal and clean recording, and $\widehat{S}$ denotes their estimates. Additionally, we use the output Signal-to-Noise Ratio (oSNR) to evaluate the noise reduction result of the proposed method [34], and the Perceptual Evaluation of Speech Quality (PESQ) as a measure of speech intelligibility [35].

We first study the impact of different forgetting factors on NMSE, oSNR and PESQ scores for the enhanced speech signal in the case of 10 dB noise, and the results are shown in Table 1. It is observed that a smaller or larger ξ typically results in a decline in the performance of the enhanced signal. According to the results, we set ξ to 0.7 as an optimal threshold for the following experiments.

Figure 2 depicts the respective results under various iSNRs of NMSE for the RHC vector and the source signal, as well as the oSNR and PESQ scores for the enhanced signal. Observing the first two subgraphs, we see the NMSEs of both the RHC vector and source signal of all methods decrease with the decrease of iSNR, while the proposed method achieves the minimal error compared to the reference methods under a lager noise level. Both the proposed and biased-based methods realize similar accuracy in a higher iSNR. Furthermore, from the oSNR results of the enhanced signal presented in the third subgraph, it can be seen that their oSNR significantly improves, i.e., the noise has been mitigated effectively, and the noise reduction effect of the proposed method is better than the other methods when iSNR $\leq$ 20 dB. However, the noise reduction performance of the proposed method seems to be less advantageous at a higher iSNR. While reducing noise, it is also necessary to ensure the intelligibility and quality of the enhanced signal. As shown in the last subgraph of Figure 2, all these three methods improve on PESQ score. We see that the RHC estimation-based methods achieve a satisfied speech intelligibility and quality, and the proposed method obtains a more prominent perceptual effect, while baseline method shows a significant difference compared to the RHC estimation-based methods.

Practically, the enhancement of SH coefficients is equivalent to reducing the noise in the soundfield as it is usually represented by a set of SH coefficients [18, 19]. In order to show the enhancement performance more obviously, Figure 3 demonstrates the respective clean, noisy and enhanced 3-D soundfield where iSNR = 5 dB ($f = 2$ kHz). Figure 4 plots the corresponding time-domain signal waves with the same noise level. We observe that the noise in both the soundfield and time-domain signal has been significantly eliminated, with few distortions, due to two critical factors: (i) an accurate estimation of RHC; and (ii) the covariance matrix computed directly from the noise signal is combined with the estimated RHC vector, so that the MVDR filter can estimate the source signal efficiently.

5. Conclusion

This paper presents a spatial acoustic enhancement approach using unbiased relative harmonic coefficients, which is mainly divided into two steps of estimating the spatial clues and source signal. It can enhance both the received recording and the whole soundfield, rather than solely extracting the source signal without any spatial information. This is particularly meaningful for spatial audio reproduction systems. Extensive simulation results validate the enhancement performance of the proposed method. A potential future work is to extend this approach to the more challenging multi-source scenario.

6. Acknowledgements

This work was supported by the Beijing Natural Science Foundation under Grants L233032 and L223033, and the National Natural Science Foundation of China under Grants 61971015.

7. References

[1] G. Huang, J. Benesty, T. Long, and J. Chen, "A family of maximum snr filters for noise reduction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 2034–2047, 2014.

[2] S. Ruiz, T. van Waterschoot, and M. Moonen, "Cascade multichannel noise reduction and acoustic feedback cancellation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 676–680.

[3] J. Benesty, J. Chen, and E. A. Habets, *Speech enhancement in the STFT domain*. Springer Science & Business Media, 2011.

[4] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2007.

[5] G. Huang, J. Benesty, and J. Chen, "Study of the frequency-domain multichannel noise reduction problem with the householder transformation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 486–490.

[6] R. C. Hendriks and T. Gerkmann, "Noise correlation matrix estimation for multi-microphone speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 223–233, 2011.

[7] S. Ruiz, T. van Waterschoot, and M. Moonen, "Cascade multichannel noise reduction and acoustic feedback cancellation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 676–680.

[8] G. Huang, J. Benesty, and J. Chen, "On the design of frequency-invariant beampatterns with uniform circular microphone arrays," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 1140–1153, 2017.

[9] Y. Hu, P. N. Samarasinghe, S. Gannot, and T. D. Abhayapala, "Decoupled multiple speaker direction-of-arrival estimator under reverberant environments," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 3120–3133, 2022.

[10] B. Rafaely, "Analysis and design of spherical microphone arrays," *IEEE Transactions on speech and audio processing*, vol. 13, no. 1, pp. 135–143, 2004.

[11] I. Balmages and B. Rafaely, "Open-sphere designs for spherical microphone arrays," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 727–732, 2007.

[12] Y. Peled and B. Rafaely, "Method for dereverberation and noise reduction using spherical microphone arrays," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010, pp. 113–116.

[13] S. K. Yadav and N. V. George, "Sparse distortionless modal beamforming for spherical microphone arrays," *IEEE Signal Processing Letters*, vol. 29, pp. 2068–2072, 2022.

[14] D. P. Jarrett, E. A. Habets, J. Benesty, and P. A. Naylor, "A tradeoff beamformer for noise reduction in the spherical harmonic domain," in *IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2012, pp. 1–4.

[15] D. P. Jarrett, E. A. Habets, and P. A. Naylor, "Spherical harmonic domain noise reduction using an mvdr beamformer and doa-based second-order statistics estimation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 654–658.

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

[17] L. McCormack, A. Politis, R. Gonzalez, T. Lokki, and V. Pulkki, "Parametric ambisonic encoding of arbitrary microphone arrays," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2062–2075, 2022.

[18] C. Borrelli, A. Canclini, F. Antonacci, A. Sarti, and S. Tubaro, "A denoising methodology for higher order ambisonics recordings," in *IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018, pp. 451–455.

[19] Y. Hu, P. N. Samarasinghe, and T. D. Abhayapala, "Acoustic signal enhancement using relative harmonic coefficients: Spherical harmonics domain approach." in *INTERSPEECH*, 2020, pp. 5076–5080.

[20] Y. Hu and S. Gannot, "Closed-form single source direction-of-arrival estimator using first-order relative harmonic coefficients," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 726–730.

[21] Y. Hu, W. Wang, Z. Gu, T. Mao, X. Zhu, and J. Jin, "Closed-form multiple source direction-of-arrival estimator under reverberant environments," *The Journal of the Acoustical Society of America*, vol. 154, no. 4, pp. 2349–2364, 2023.

[22] A. Fahim, T. D. Abhayapala, and P. N. Samarasinghe, "PSD estimation of multiple sound sources in a reverberant room using a spherical microphone array," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017, pp. 76–80.

[23] A. Fahim, P. N. Samarasinghe, and T. D. Abhayapala, "PSD estimation and source separation in a noisy reverberant environment using a spherical microphone array," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1594–1607, 2018.

[24] E. G. Williams and J. A. Mann III, "Fourier acoustics: sound radiation and nearfield acoustical holography," 2000.

[25] O. Shalvi and E. Weinstein, "System identification using nonstationary signals," *IEEE transactions on signal processing*, vol. 44, no. 8, pp. 2055–2063, 1996.

[26] I. Cohen, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 451–459, 2004.

[27] S. Gannot, D. Burshtein, and E. Weinstein, "Relative transfer function identification using speech signals," *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, 2001.

[28] J. Zhou, C. Bao, X. Zhang, W. Xiong, and M. Jia, "Design of a robust MVDR beamforming method with low-latency by reconstructing covariance matrix for speech enhancement," *Applied Acoustics*, vol. 211, p. 109464, 2023.

[29] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on speech and audio processing*, vol. 9, no. 5, pp. 504–512, 2001.

[30] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Transactions on speech and audio processing*, vol. 11, no. 5, pp. 466–475, 2003.

[31] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*. Springer Science & Business Media, 2008, vol. 1.

[32] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[33] Ntt cam china. Accessed: July 1, 2022. [Online]. Available: https://www.nttdata.com

[34] G. Huang, J. Chen, and J. Benesty, "Investigation of a parametric gain approach to single-channel speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 206–210.

[35] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *IEEE international conference on acoustics, speech, and signal processing (ICASSP)*, 2001, pp. 749–752.