



Power Spectral Density Estimation for Acoustic Source Separation Using A Spherical Microphone Array

Liang Tao¹, Maoshen Jia^{1,*}, Yonggang Hu²

¹Beijing University of Technology, Beijing, China

²Shanghai branch of the Southwest Institute of Electronics and Telecommunication Technology of China, Shanghai, China

liangtaobjut@163.com, jiamashen@bjut.edu.cn,
yongganghu@outlook.com

Abstract

Acoustic source separation using microphone arrays has garnered increasing attention within the audio signal processing community. Cascading a beamformer and a post-filter for source separation is widely recognized as an effective solution, where accurate estimation of the power spectral density (PSD) is crucial for post-filter. However, estimating the PSD components is a challenging task in the presence of noise. In this paper, we propose an efficient multi-source PSD estimation method using a spherical microphone array (SMA) in noisy environments for source separation. Specifically, the orthogonality of spherical harmonics (SH) inherent in an error-free sampled SMA enables effective suppression of noise components in the spatial covariance matrix (SCM) of the observation signals, facilitating accurate PSD estimation for each individual source, thus the post-filter would be effectively constructed. The source separation problem is then addressed by applying a SH-domain beamformer along with the developed post-filter. Both simulation and experimental results demonstrate the proposed algorithm's effectiveness over the baseline methods.

Index Terms: acoustic source separation, power spectral density, spherical microphone arrays, noisy environments

1. Introduction

Acoustic source separation refers to the process of extracting one or several sources from the mixed recordings. Accurate source separation not only improves speech quality and intelligibility, but also enhances the performance of various practical applications, such as automatic speech recognition (ASR) [1]. Up to present, we have witnessed notable development of acoustic source separation in both academia and industry, particularly in scenarios using multiple microphones [2].

Specifically, microphone array offers significant advantages for source separation, as the additional spatial cues provide valuable information for distinguishing sources waving from different directions [3]. Traditional separation methods include: (i) sparsity-based methods which utilize the W-disjoint orthogonality (WDO) assumption that each time-frequency (TF) bin of the mixture in the short-time Fourier transform (STFT) domain is dominated by a single source, enabling source separation by predicting a binary mask for each TF bin [4]; (ii) beamforming methods, separating multiple sources by steering beamformer toward the direction of each source [5, 6]; (iii) independent component analysis (ICA) methods, assuming each source is independent and follows a non-Gaussian distribution, and the mixing matrix is estimated using maximizing non-Gaussian-based or information theoretic-based methods to recover the source signal [7, 8]; (iv) non-negative matrix factorization (NMF) methods, decomposing the mixed signal into

multiple independent, non-negative components, facilitating effective source separation [9, 10].

Beyond source separation, spherical microphone arrays (SMAs) have been widely investigated for various applications, such as direction-of-arrival (DOA) estimation [11], speech enhancement [12] and spatial audio rendering [13]. This popularity primarily stems from the decomposition of the sound pressure into the spherical harmonics (SH) domain, offering two main advantages: (i) the decoupled frequency- and angular-components; and (ii) enhanced spatial resolution. Recently, there arises a growing interest using spherical arrays for source separation. Epain et. al. proposed an extension of the ICA method into the SH domain, enhancing the separation performance [14]. Kalkur et.al. introduced a joint source localization and separation approach based on sparsity methods [15]. In another study [16], the authors achieved source separation using orthogonal matching pursuit (OMP) on complex-valued steered response maps. Fahim et. al. further enhanced separation quality by incorporating a wiener post-filter estimating the power spectral density (PSD) of each individual source after a fixed beamformer [17]. Sun et. al. optimized the approach developed in [18] by applying the relative harmonic coefficients (RHC), achieving improved separation effect in reverberant environments [19].

This paper presents a novel solution for estimating the PSDs of multiple sources using an SMA in noisy environments, enhancing the robustness of source separation against noise. Specifically, we highlight the SH orthogonality inherent in error-free sampled SMAs effectively suppresses noise components in the spatial covariance matrix (SCM) of the observation signals, enabling accurate PSD estimation for each individual source. This, in turn, facilitates the design of a post-filter and achieves efficient source separation. We demonstrate the superiority of the proposed method in noisy environments through evaluating numerically and perceptually.

2. System Model

Consider a higher-order SMA consisting of J omni-directional sensors to pick up Q simultaneous sound sources in a noisy environment. Since array signals are typically processed in the frequency domain, with the short-time Fourier transform (STFT), the multi-channel observations are modeled as,

$$\begin{aligned} \mathbf{p}(l, k) &= [P_1(l, k) P_2(l, k) \cdots P_J(l, k)]^T \\ &= \mathbf{x}(l, k) + \mathbf{v}(l, k), \end{aligned} \quad (1)$$

where

$$P_j(l, k) = \sum_{q=1}^Q S_q(l, k) H_{j,q}(l, k) + V_j(l, k), \quad (2)$$

l denotes the time frame index, $k = w/c$ is the wave number, $w = 2\pi f$ is the angular frequency with the temporal frequency f , and c is the sound speed, $P_j(l, k)$ is the observation signal at the j -th ($j = 1, 2, \dots, J$) channel, $S_q(l, k)$ denotes the q -th incident signal, $H_{j,q}(l, k)$ denotes the acoustic transfer function (ATF) from the q -th source to the j -th microphone, $V_j(l, k)$ denotes the noise signal at the j -th sensor, $\mathbf{x}(l, k)$ denotes the clean signal vector, $\mathbf{v}(l, k)$ denotes the noise signal vector, and $[\cdot]^T$ denotes the transpose operator. Additionally, all the signals are assumed to be zero mean, and the noise signal is also assumed to be uncorrelated with the source signal, i.e., $\mathbb{E}[\mathbf{x}(l, k)\mathbf{v}^H(l, k)] = \mathbf{0}$, where $\mathbb{E}[\cdot]$ denotes the statistical expectation operator and $[\cdot]^H$ denotes the conjugate transpose operator. For the sake of simplicity, the dependence on the time index l is omitted in subsequent chapters. Given multi-channel mixed recordings, the aim by this paper is to estimate the PSD component of each source signal for post-filter construction, followed by applying an SH-domain beamformer along with the developed post-filter to separate individual sources.

Multi-channel measurements acquired by SMAs tend to be decomposed into the SH domain due to several additional advantages and a more convenient representation. By expanding the signal vector in the form of SH while keeping the noise term constant, (1) is able to be represented as [20],

$$\mathbf{p}(k) = \sum_{n=0}^N \sum_{m=-n}^n \alpha_{nm}(k) b_n(kr) \mathbf{T}_{nm} + \mathbf{v}(k), \quad (3)$$

where $N = \lceil kr \rceil$ is the maximum truncated SH order due to the high-pass nature of the spherical Bessel functions [21], and $\lceil \cdot \rceil$ denotes the ceiling operator. Notably, the sound field order is closely related to the number of sensors, i.e., $J \geq (N+1)^2$ must be satisfied to prevent spatial aliasing. $\alpha_{nm}(k)$ is the SH coefficient that characterizes the sound field information within the interior region. $b_n(kr)$ is the radial function and various types of sphere has different expression,

$$b_n(kr) = \begin{cases} 4\pi i^n j_n(kr), & \text{open sphere} \\ 4\pi i^n \left[j_n(kr) - \frac{j'_n(kr)}{h'_n(kr)} h_n(kr) \right], & \text{rigid sphere} \end{cases} \quad (4)$$

where $(\cdot)'$ denotes the first derivative of a function, $j_n(\cdot)$ and $h_n(\cdot)$ are the first kind of spherical Bessel and second kind of spherical Hankel functions, respectively. \mathbf{T}_{nm} denotes an SH function vector related to the position of microphone, i.e.,

$$\mathbf{T}_{nm} = [Y_{nm}(\Psi_1), Y_{nm}(\Psi_2), \dots, Y_{nm}(\Psi_J)]^T, \quad (5)$$

and $Y_{nm}(\Psi_j)$ is the SH function of order n and degree m , with $\Psi_j = (\theta_j, \phi_j)$ being the direction of the j -th sensor,

$$Y_{nm}(\Psi_j) = \sqrt{\frac{(2n+1)(n-m)!}{4\pi(n+m)!}} P_{nm}(\cos\theta_j) e^{im\phi_j}, \quad (6)$$

where $P_{nm}(\cdot)$ is the associated Legendre polynomial, and $(\cdot)!$ denotes the factorial operator. In the sound field sampling scheme, uniform sampling is the most popular solution by arranging multiple microphones uniformly on the surface of a sphere [22], such as the sensor layout of Eigenmike [23], thus the following orthogonal property is satisfied,

$$\mathbf{T}_{nm}^H \mathbf{T}_{n'm'} = \begin{cases} \frac{J}{4\pi}, & n = n', m = m' \\ 0, & \text{orther} \end{cases} \quad (7)$$

3. Proposed Method

Motivated by directional-gain beamforming [24] and the orthogonal property described in (7), the following matrix is defined as,

$$\mathcal{F}(k) = [\mathbf{f}_{1,-1}(k), \mathbf{f}_{1,0}(k), \dots, \mathbf{f}_{NN}(k)], \quad (8)$$

where

$$\mathbf{f}_{nm}(k) = \frac{1}{b_n^*(kr)} \frac{4\pi}{J} \mathbf{T}_{nm}. \quad (9)$$

The vectors in (8) are denoted by primary ones, and the secondary vector is defined as that of zeroth-order, in which $(\cdot)^*$ denotes the conjugate operator. Applying the primary and secondary vectors on the SCM of the received signals, we have,

$$\mathcal{F}^H(k) \Phi_{\mathbf{p}}(k) \mathbf{f}_{00}(k) = \mathcal{F}^H(k) \Phi_{\mathbf{x}}(k) \mathbf{f}_{00}(k), \quad (10)$$

where

$$\Phi_{\mathbf{p}}(k) = \mathbb{E}[\mathbf{p}(k)\mathbf{p}^H(k)] = \Phi_{\mathbf{x}}(k) + \Phi_{\mathbf{v}}(k), \quad (11)$$

$\Phi_{\mathbf{x}}(k)$ and $\Phi_{\mathbf{v}}(k)$ are respectively the SCMs of $\mathbf{x}(k)$ and $\mathbf{v}(k)$, which are defined similarly to $\Phi_{\mathbf{p}}(k)$. From (10), we see that the noise covariance matrix is absent due to the orthogonality between the primary and secondary vectors. Specifically, rewrite the covariance matrix $\Phi_{\mathbf{v}}(k)$ as,

$$\Phi_{\mathbf{v}}(k) = \phi_V(k) \mathbf{\Gamma}, \quad (12)$$

where $\phi_V(k)$ is the variance of noise that is the same at all sensors, $\mathbf{\Gamma}$ is denoted by the pseudo-coherence matrix of noise. One can easily prove that the following equality is satisfied in the case of both spatially white noise and spherically isotropic (diffuse) noise field,

$$\mathcal{F}^H(k) \Phi_{\mathbf{v}}(k) \mathbf{f}_{00}(k) = \phi_V(k) \mathcal{F}^H(k) \mathbf{\Gamma} \mathbf{f}_{00}(k) = \mathbf{0}. \quad (13)$$

Therefore, substitute (3) and (8) into (10), then combine (7),

$$\mathcal{F}^H(k) \Phi_{\mathbf{p}}(k) \mathbf{f}_{00}(k) = \begin{bmatrix} \alpha_{1,-1}(k) \alpha_{0,0}^*(k) \\ \alpha_{1,0}(k) \alpha_{0,0}^*(k) \\ \vdots \\ \alpha_{NN}(k) \alpha_{0,0}^*(k) \end{bmatrix}. \quad (14)$$

Assume the multi-path effect is neglected for simplification of signal model, thus the analytic expression of SH coefficient, $\alpha_{nm}(k)$, is given by,

$$\alpha_{nm}(k) = \sum_{q=1}^Q S_q(k) Y_{nm}^*(\Phi_q), \quad (15)$$

where $\Phi_q = (\vartheta_q, \varphi_q)$ is the DOA of the q -th source. Substitute (15) into (14), it expresses as,

$$\underbrace{\begin{bmatrix} \alpha_{1,-1}(k) \alpha_{0,0}^*(k) \\ \alpha_{1,0}(k) \alpha_{0,0}^*(k) \\ \vdots \\ \alpha_{NN}(k) \alpha_{0,0}^*(k) \end{bmatrix}}_{\Delta(k)} = \underbrace{\begin{bmatrix} Y_{1,-1}^*(\Phi_1) Y_{00} & \cdots & Y_{1,-1}^*(\Phi_Q) Y_{00} \\ \vdots & \vdots & \vdots \\ Y_{N,N}^*(\Phi_1) Y_{00} & \cdots & Y_{N,N}^*(\Phi_Q) Y_{00} \end{bmatrix}}_{\Upsilon} \underbrace{\begin{bmatrix} \phi_{S_1}(k) \\ \phi_{S_2}(k) \\ \vdots \\ \phi_{S_Q}(k) \end{bmatrix}}_{\Theta(k)} \quad (16)$$

Consequently, the PSD of individual source signal can be estimated using the least square solution,

$$\Theta(k) = (\mathbf{Y}^T \mathbf{Y})^{-1} \mathbf{Y}^T \Delta(k), \quad (17)$$

where $(\cdot)^{-1}$ denotes the inverse operator. Note that the computation of (17) requires the source DOAs as well as their number to be correctly estimated for ensuring the efficiency of \mathbf{Y} . Multi-source DOA estimation in the SH domain has been extensively studied, such as multiple signal classification (MUSIC) [25], pseudo-intensity vector (PIV) [26], and RHC-based methods [27]. In this work, we adopt the improved MUSIC algorithm developed in [28] for multi-source localization and source number counting. Due to space limitations, the reader refer to [28] for further details.

In multi-channel source separation and enhancement, beamforming is one of the most effective approaches, particularly adaptive beamforming. As noted in [29], an adaptive beamformer can be decomposed into the product of a fixed beamformer and a post filter, i.e.,

$$\mathbf{h}(k) = \mathbf{h}_F(k)M(k), \quad (18)$$

where $\mathbf{h}_F(k)$ denotes a fixed beamformer, and $M(k)$ denotes a post-filter. Generally, fixed beamformers are the easiest to implement, as they solely rely on the DOAs of the sources rather than statistical parameters. The choice of a fixed beamforming technique depends on specific design criteria, such as delay-and-sum (DS) and maximum directivity (MD) [22]. Consider the aperture of the recording area is much smaller than the distance to the sources, an MD beamformer is adopted for the initial separation of source signals. The output by the beamformer toward the q -th source is given as,

$$Z_q(k) = \mathbf{h}_{F,q}^H(k) \alpha_N(k) = \frac{4\pi}{(N+1)^2} \mathbf{y}_N(\Phi_q) \alpha_N(k), \quad (19)$$

where $\mathbf{y}_N(\Phi_q)$ denotes an N -th order SH function vector associated with the DOA of the q -th source, i.e.,

$$\mathbf{y}_N(\Phi_q) = [Y_{00}(\Phi_q), Y_{1,-1}(\Phi_q), \dots, Y_{NN}(\Phi_q)], \quad (20)$$

and $\alpha_N(k)$ denotes the SH coefficient vector of the received signals, where the (n, m) -th element can be estimated using,

$$\alpha_{nm}(k) = \frac{4\pi}{J \times b_n(kr)} \mathbf{T}_{nm}^H \mathbf{p}(k). \quad (21)$$

Subsequently, the separated signal would be enhanced by a post-filter, denoted as below,

$$M_q(k) = \frac{\phi_{S_q}(k)}{\sum_{q=1}^Q \phi_{S_q}(k) + \phi_V(k)}, \quad (22)$$

where $\phi_V(k)$ is the PSD of noise, and the sum of the individual PSD components, i.e., the denominator in (22), is replaced with $\text{tr}[\Phi_{\mathbf{p}}(k)]/J$, in which $\text{tr}[\cdot]$ denotes the trace of a matrix. We emphasize that the mean of the diagonal elements of the observation signals can be well approximated by the sum of individual PSD components in the environment with low reverberation. However, the coherent components due to increased reflections in highly reverberant conditions would disrupt this principle. The q -th source signal is finally estimated by,

$$\tilde{S}_q(k) = Z_q(k)M_q(k). \quad (23)$$

Repeat the above procedures over the entire STFT domain, followed by the individual separated time-domain source signal can be recovered using the inverse STFT. Figure 1 presents a compact flowchart of the proposed algorithm.

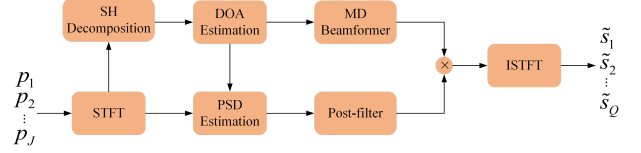


Figure 1: Flowchart of the proposed method.

4. Experiments

In this section, we present the experimental results of the proposed algorithm in both simulated and real-world scenarios, with a comparison of baseline methods such as beamforming [22] and the RHC methods [19]. The separation performance is evaluated both numerically and perceptually using three objective metrics: perceptual evaluation of speech quality (PESQ) [30], short-time objective intelligibility (STOI) and signal to interference ratio (SIR) [31].

4.1. Simulated Environments

We simulate a rectangular room with dimensions of $6\text{m} \times 4\text{m} \times 3\text{m}$ using a public toolbox¹ that based on image source method [32] to synthesize room impulse response (RIR). An open spherical array equipped with 32 capsules is positioned at the center of the room to capture the interior sound field. All source signals are assumed to be static and placed 1m away from the array origin. Dry speech signals are randomly selected from the Nippon Telegraph and Telephone (NTT) corporation database [33], and are down-sampled from 16 kHz to 8 kHz. The noise signal received at the sensors are composed of the both the Gaussian white noise and diffuse noise. We randomly select three different speakers and overlap them in the time domain to simulate multiple simultaneously active sources, with the angular separation between sources being 60° . Both the anechoic and reverberant ($T_{60} = 300$ ms) environments are considered to access the proposed method with varying signal-to-noise ratios (SNRs). The mixture signals are transformed into the frequency domain using the STFT with a Hamming window of 256 samples, a 25% overlap between frames, and a 256-sample fast Fourier transform (FFT).

Figure 2 depicts an example of the spectrograms for the mixed signal, the first clean source signal, the separated source signal and the estimated mask, respectively. It is evident that our method effectively separates the desired source signal from strong background noise, significantly alleviating most of the noise. Furthermore, the estimated mask closely matches the spectrogram of the desired signal, confirming the validity of the proposed PSD estimation for post-filter.

Table 1 presents the separation results of the proposed and baseline methods in scenarios with three simultaneous sources. We see that in both anechoic and reverberant environments, the proposed method consistently outperforms the baseline methods across diverse noise levels. Specifically, compared with the beamforming method, our approach achieves significant improvements in terms of PESQ, STOI and SIR, indicating the effectiveness of the PSD estimation for post-filter. Additionally, the proposed method demonstrates greater robustness in noisy environments than the RHC method, particularly at lower SNR levels. This is because our method effectively extracts the PSD components of each source in the presence of noise, whereas the

¹<https://www.audiolabs-erlangen.de/fau/professor/habets/software/signal-generator>

Table 1: Results comparison with baseline methods under anechoic and reverberant noisy environments. The values are specified with PESQ/STOI(%)/SIR(dB) format. **BOLD** indicates the best score in each case.

| T_{60} | Method | -5 dB | 0 dB | 5 dB | 10 dB | 15 dB |
|----------|-------------|-------------------------|-------------------------|-------------------------|-------------------------|--------------------------|
| 0 ms | Unprocessed | 1.12/53.67/-2.65 | 1.14/57.33/-2.65 | 1.35/60.14/-2.66 | 1.47/62.02/-2.66 | 1.58/63.35/-2.65 |
| | Beamforming | 1.58/70.48/4.49 | 1.84/75.06/4.96 | 2.01/78.43/5.03 | 2.12/79.87/5.07 | 2.17/81.02/5.09 |
| | RHC | 1.43/70.52/9.75 | 1.82/76.49/14.49 | 2.22/83.47/17.56 | 2.58/87.94/20.73 | 2.88/ 92.93/22.95 |
| | Proposed | 2.34/72.78/10.34 | 2.57/79.01/14.96 | 2.78/84.64/18.75 | 3.01/88.12/20.82 | 3.19/89.31/21.51 |
| 300 ms | Unprocessed | 1.03/51.55/-2.69 | 1.16/54.14/-2.69 | 1.33/56.06/-2.69 | 1.45/57.43/-2.70 | 1.51/58.31/-2.70 |
| | Beamforming | 1.53/67.21/2.95 | 1.71/70.67/3.15 | 1.84/72.85/3.19 | 1.89/73.62/3.22 | 1.92/74.60/3.24 |
| | RHC | 1.38/64.75/3.52 | 1.65/69.52/3.86 | 1.78/71.85/3.90 | 1.89/73.49/3.93 | 1.91/74.77/3.97 |
| | Proposed | 1.99/65.56/5.85 | 2.10/70.30/7.94 | 2.16/73.21/9.31 | 2.21/75.51/10.04 | 2.23/76.19/10.23 |

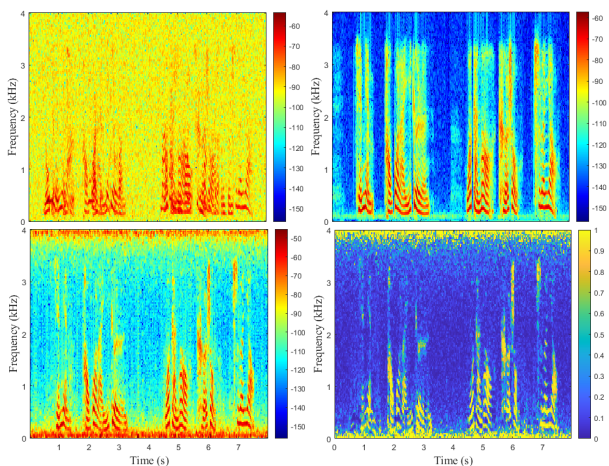


Figure 2: Spectrograms of the mixed signal, the first source signal, the separated source signal and the estimated mask in the case of 5 dB noise.

RHC method tends to neglect the impact of noise during source separation. However, we also note that the proposed method experiences significant degradation in reverberant scenarios. This is due to the fact that, in (22), the variance of the observation signal no longer represents the sum of individual PSD components but instead includes numerous coherent components. As a result, the efficiency of the post-filter is diminished, leading to a decline in separation performance.

4.2. Real-world Environments

We finally evaluate the proposed algorithm using real-world measurements recorded with an EM32 Eigenmike in a reverberant chamber (see Figure 3). The room dimensions are [7, 6, 4]m, with $T_{60} \approx 350$ ms. The background noise originates from sensor self-noise and outdoor activity. Three loudspeakers are positioned 1m away from the array at approximate directions of $\Phi_1 = (229^\circ, 98^\circ)$, $\Phi_2 = (314^\circ, 104^\circ)$ and $\Phi_3 = (59^\circ, 92^\circ)$, respectively. All other experimental settings are consistent with those used in the simulations above. The results presented in Table 2 show that the proposed method continues to achieve the optimal separation performance compared to the baseline methods, aligning with the simulation results. However, we observe a slight decline in separation quality compared to simulations



Figure 3: Experimental setup in a reverberant chamber

Table 2: Separation results in the real-world scenarios

| Method | PESQ | STOI (%) | SIR (dB) |
|-------------|-------------|--------------|-------------|
| Unprocessed | 1.39 | 54.39 | -3.27 |
| Beamforming | 1.67 | 64.42 | 3.16 |
| RHC | 1.74 | 67.72 | 3.74 |
| Proposed | 2.06 | 73.44 | 8.29 |

due to non-negligible errors encountered in practical scenarios, such as deviations in DOA estimation. Overall, the results from real-world recordings confirm that the proposed algorithm is effective in practical environments.

5. Conclusions

In this paper, we propose a source separation approach based on PSD estimation in noisy environments using a spherical array. By exploiting the SH orthogonality of error-free sampled arrays and incorporating the design principle of directional-gain beamforming, the PSD of each individual source can be effectively estimated in noisy conditions. This enables efficient source separation through a post-filter applied after beamforming. Extensive experimental results, including both simulated and real-world multi-channel recordings, validate the effectiveness of the proposed method.

However, the proposed method experiences significant performance degradation in reverberant scenarios, as the current signal model does not account for multi-path effects. A promising future research direction is to enhance the reliable estimation of the PSD for each source, as well as reverberation and noise components, thus improves source separation in challenging reverberant and noisy environments.

6. Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant No. 62471012 and Beijing Natural Science Foundation (No. L233032, L223033).

7. References

- [1] Z. Zhang, J. Geiger, J. Pohjalainen, A. E.-D. Mousa, W. Jin, and B. Schuller, "Deep learning for environmentally robust speech recognition: An overview of recent developments," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 9, no. 5, pp. 1–28, 2018.
- [2] K. Tesch and T. Gerkmann, "Multi-channel speech separation using spatially selective deep non-linear filters," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 542–553, 2023.
- [3] C. Quan and X. Li, "Spatialnet: Extensively learning spatial information for multichannel joint speech separation, denoising and dereverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1310–1323, 2024.
- [4] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on signal processing*, vol. 52, no. 7, pp. 1830–1847, 2004.
- [5] D. Li, G. Huang, Y. Lei, J. Chen, and J. Benesty, "Robust source separation with differential microphone arrays and independent low-rank matrix analysis," in *2020 28th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 291–295.
- [6] S. Araki, H. Sawada, and S. Makino, "Blind speech separation in a meeting situation with maximum snr beamformers," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, vol. 1, 2007, pp. 1–41.
- [7] S. Choi, A. Cichocki, H.-M. Park, and S.-Y. Lee, "Blind source separation and independent component analysis: A review," *Neural Information Processing-Letters and Reviews*, vol. 6, no. 1, pp. 1–57, 2005.
- [8] Y. Yang, X. Wang, W. Zhang, and J. Chen, "Independent vector analysis assisted adaptive beamforming for speech source separation with an acoustic vector sensor," in *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2022, pp. 1–5.
- [9] J. Nikunen and A. Politis, "Multichannel nmf for source separation with ambisonic signals," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018, pp. 251–255.
- [10] Y. Mitsufuji, N. Takamune, S. Koyama, and H. Saruwatari, "Multichannel blind source separation based on evanescent-region-aware non-negative tensor factorization in spherical harmonic domain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 607–617, 2020.
- [11] 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.
- [12] M. Lugasi and B. Rafaely, "Speech enhancement using masking for binaural reproduction of ambisonics signals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1767–1777, 2020.
- [13] Y. Hu, P. N. Samarasinghe, T. D. Abhayapala, and G. Dickins, "Modeling characteristics of real loudspeakers using various acoustic models: Modal-domain approaches," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 561–565.
- [14] N. Epain and C. T. Jin, "Independent component analysis using spherical microphone arrays," *Acta Acustica united with Acustica*, vol. 98, no. 1, pp. 91–102, 2012.
- [15] S. N. Kalkur, S. Reddy, and R. M. Hegde, "Joint source localization and separation in spherical harmonic domain using a sparsity based method," in *INTERSPEECH*, 2015, pp. 1493–1497.
- [16] M. B. Çöteli and H. Hacıhabiboğlu, "Acoustic source separation using rigid spherical microphone arrays via spatially weighted orthogonal matching pursuit," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018, pp. 81–85.
- [17] A. Fahim, P. N. Samarasinghe, and T. D. Abhayapala, "Psd estimation of multiple sound sources in a reverberant room using a spherical microphone array," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017, pp. 76–80.
- [18] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Source counting and separation based on simplex analysis," *IEEE Transactions on Signal Processing*, vol. 66, no. 24, pp. 6458–6473, 2018.
- [19] H. Sun, P. Samarasinghe, and T. D. Abhayapala, "Blind source counting and separation with relative harmonic coefficients," in *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.
- [20] P. N. Samarasinghe, T. D. Abhayapala, and H. Chen, "Estimating the direct-to-reverberant energy ratio using a spherical harmonics-based spatial correlation model," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 2, pp. 310–319, 2016.
- [21] D. B. Ward and T. D. Abhayapala, "Reproduction of a plane-wave sound field using an array of loudspeakers," *IEEE Transactions on speech and audio processing*, vol. 9, no. 6, pp. 697–707, 2001.
- [22] D. P. Jarrett, E. A. Habets, and P. A. Naylor, *Theory and applications of spherical microphone array processing*. Springer, 2017, vol. 9.
- [23] E. Pfanzagl-Cardone, "HOA—higher order ambisonics (eigenmike®)," in *The Art and Science of 3D Audio Recording*. Springer, 2023, pp. 189–209.
- [24] C. Pan and J. Chen, "A framework of directional-gain beamforming and a white-noise-gain-controlled solution," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2875–2887, 2022.
- [25] Y. Hu, T. D. Abhayapala, and P. N. Samarasinghe, "Multiple source direction of arrival estimations using relative sound pressure based MUSIC," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 253–264, 2020.
- [26] S. Hafezi, A. H. Moore, and P. A. Naylor, "Multiple source localization in the spherical harmonic domain using augmented intensity vectors based on grid search," in *2016 24th European Signal Processing Conference (EUSIPCO)*, 2016, pp. 602–606.
- [27] Y. Hu, P. N. Samarasinghe, T. D. Abhayapala, and S. Gannot, "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.
- [28] D. Khaykin and B. Rafaely, "Coherent signals direction-of-arrival estimation using a spherical microphone array: Frequency smoothing approach," in *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2009, pp. 221–224.
- [29] S. Gannot and I. Cohen, "Adaptive beamforming and postfiltering," *Springer handbook of speech processing*, pp. 945–978, 2008.
- [30] 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 *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, vol. 2, 2001, pp. 749–752.
- [31] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.
- [32] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.
- [33] "NTT database," 2008, accessed 2009. [Online]. Available: <http://www.ntt-at.com/product>