Reliable Intensity Vector Selection for Multi-source Direction-of-Arrival Estimation Using a Single Acoustic Vector Sensor

Jianhua Geng, Sifan Wang, Juan Li, JingWei Li and Xin Lou

School of Information Science and Technology, ShanghaiTech University, China

{gengjh, wangsf, lijuan1, lijw, louxin}@shanghaitech.edu.cn

Abstract

In the context of multi-source direction of arrival (DOA) estimation using a single acoustic vector sensor (AVS), the received signal is usually a mixture of noise, reverberation and source signals. The identification of the time-frequency (TF) bins that are dominated by the source signals can significantly improve the robustness of the DOA estimation. In this paper, a TF bin selection based DOA estimation pipeline is proposed. The proposed pipeline mainly involves three key steps: key frame identification, TF bin selection and DOA extraction. We identify the key frames by frame-wisely examining the effective rank. Subsequently, the geometric medians of the selected key frames are extracted to alleviate the impact of extreme outliers. The simulation results show that the accuracy and the robustness of the proposed pipeline outperform the state-of-the-art (SOTA) techniques.

Index Terms: Direction of arrival (DOA), intensity vector (IV), time-frequency (TF) bins, reverberation

1. Introduction

Direction of arrival (DOA) estimation of multiple acoustic sources in a reverberant and noisy environment has long been of great interest to the signal processing society [1–3]. It is a fundamental building block in various applications such as video conferencing, smart robot and hearing aids [4–8]. In these applications, an array of microphones is usually used to capture the acoustic signals. DOA cues are extracted from the time delay information between different microphone elements. In practical application scenarios, reverberation and interference between simultaneously active sources are major challenges for accurate DOA estimation [9, 10].

Recently, pseudo-intensity vector (PIV)-based TF bin selection methods using spherical microphone array (SMA) are proposed for DOA estimation [11–15]. The PIV-based direct path dominant (DPD) test proposed in [14] identifies appropriate TF bins by examining whether the effective rank of the estimated covariance matrix is unity or not. One of the methods to determine the effective rank is to examine the ratio of the largest and second-largest singular values of the estimated covariance matrix. Another PIV-based estimation consistency (EC) algorithm is proposed in [15]. In the EC algorithm, the higher consistency between PIVs in the source-dominated TF regions is exploited for TF bin selection. More specifically, the EC weight for a TF bin is obtained by multiplying the corresponding frame weight and the frequency weight. Subsequently, the TF bins corresponding to large EC weights will be selected. Unfortunately, the performance of the aforementioned algorithms degrades with the increase of simultaneous active sources.

It has been noted that, in addition to the PIV-based approaches, the intensity vector (IV)-based approaches are also able to produce DOA estimation from each TF bin [16]. By
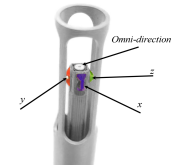

Figure 1: *Acoustic vector sensor structure. The three orthogonal directional microphones are indicated by x, y and z.*

exploiting the sound-field directivity of acoustic vector sensors (AVSs), several IV-based approaches for selecting reliable TF regions/bins have been proposed [17–21]. Two IV-based low-reverberant-single-source (LRSS) detection algorithms using a single AVS are proposed in [20] and [21], where the algorithm in [20] works at TF region level and the algorithm in [21] works at TF bin level. In [20], the TF region with low covariance rank is identified as LRSS region. Since the number of such TF regions decreases with increasing reverberation, the LRSS detection algorithm at TF region level is further extended to TF bin level. In [21], the TF bin is selected by comparing the absolute direction between the real and imaginary parts of the observations with a pre-defined threshold. However, it does not work properly in adjacent sources scenarios [21]. The reason is that low angular separation between sources also results in similar absolute directions between the real and imaginary parts, regardless it is a LRSS or non-LRSS point.

In this work, an IV-based DOA estimation pipeline which selects appropriate TF bins dominated by a single source and rejects those contaminated by reverberation and noise, is proposed. The proposed pipeline consists of three building blocks, which are key frame identification block, TF bin selection block and DOA extraction block. To identify the key frames, we propose to extend the DPD test to the frame level by adjusting the length of frequency smoothing. The TF bin selection block exploits the geometric median of the IVs to improve the robustness to extreme outliers. The DOA extraction block partitions the remaining vectors into several clusters according to directions and outputs the centroid of each cluster as the final estimated DOA.

2. Mathematical Model

Figure 1 shows the structure of an AVS, consisting of one monopole pressure sensor element and three orthogonally oriented dipole elements. The frequency-independent array manifold of an AVS is defined as

$$\mathbf{a} := \begin{bmatrix} 1 \\ \cos\psi\cos\phi \\ \cos\psi\sin\phi \\ \sin\psi \end{bmatrix} = \begin{bmatrix} 1 \\ u_x(\psi,\phi) \\ u_y(\psi,\phi) \\ u_z(\psi) \end{bmatrix} = \begin{bmatrix} 1 \\ \mathbf{u} \end{bmatrix}, \qquad (1)$$
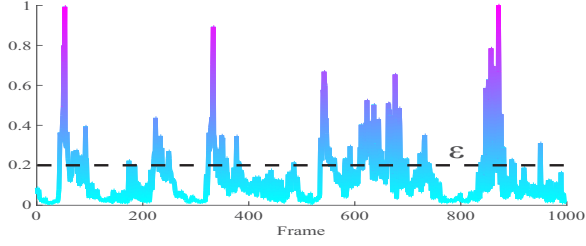
Figure 2: *An example of key frame identification. The Y-axis is the normalized ratio between the largest and second-largest eigenvalues and the dotted line is the threshold.*

Figure 3: *Schematic diagram of geometric median and center of mass. The black arrow lines represent IVs.*

where the variables $\phi \in (-\pi, \pi]$ and $\psi \in [-\pi/2, \pi/2]$ represent the azimuth and elevation angles, respectively. As we can see from (1), the unit vector $\mathbf{u}$ points towards the source direction. In this work, we assume far-field plane waves with band-limited spectrum travel in a homogeneous space. Let us consider J active acoustic sources received by a single AVS in a noisy and reverberant scenario. According to [21], using short-time Fourier transform (STFT) [22], the received signal in the TF domain can be expressed as

$$\mathbf{x}(k, l) = \sum_{i=1}^{J} \sum_{l'=0}^{L} \mathbf{h}_i(k, l') s_i(k, l - l') + \bar{\mathbf{n}}(k, l), \quad (2)$$

where k and l are frequency-bin and frame-bin index, respectively. The received signal $\mathbf{x}(k, l) = [x_p(k, l), \mathbf{x}_v^T(k, l)]^T$ consists of the outputs of the omni-directional and the three orthogonally-directional elements. The vector $\mathbf{h}_i(k, l')$ is STFT coefficients of the acoustic transfer function (ATF), which describes acoustic field between the ith source and the AVS. L denotes the maximum frame index of the ATF. The second term can be expressed as $\bar{\mathbf{n}}(k, l) = \mathbf{e}(k, l) + \mathbf{n}(k, l)$, where $\mathbf{e}(k, l)$ represents the approximation error introduced by the STFT modeling and $\mathbf{n}(k, l)$ is the additive noises.

In general, the ATF is formed by the summation of multiple propagation paths, consisting of direct path, early echoes and reverberations. Note that the latter two terms are also known as multi-path signals [1]. The ATF of direct-path ($l' = 0$) and multi-path ($l' = 1, ..., L$) with respect to the ith source can be modeled as

$$\mathbf{h}_i(k, 0) = e^{-j\omega_k \tau_i} \mathbf{a}_i, \quad (3)$$

$$\mathbf{h}_i(k, l') = \sum_{r=1}^{N_{mp}^{l'}} \alpha_i^{(l', r)} e^{-j\omega_k \tau_i^{(l', r)}} \mathbf{a}_i^{(l', r)}, \quad (4)$$

where ω_k is the discrete angular frequency, τ_i is the sample delay of the direct-path impulse. The variable $N_{mp}^{l'}$ is the number of reflections within the l'th frame and $\alpha_i^{(l', r)}$ is the attenuation coefficient. The superscript (l', r) denotes the rth reflection path in the l'th frame. The array manifold is $\mathbf{a}_i = [1, \cos \psi_i \cos \phi_i, \cos \psi_i \sin \phi_i, \sin \psi_i]^T$. The objective of 2-D multi-source DOA estimation is to find out the azimuth ϕ_i and elevation ψ_i for $i = 1, ..., J$.

3. The Proposed Pipeline

In this work, we focus on the context of multi-source DOA estimation using a single AVS, where the number of sources is known as *a priori*.
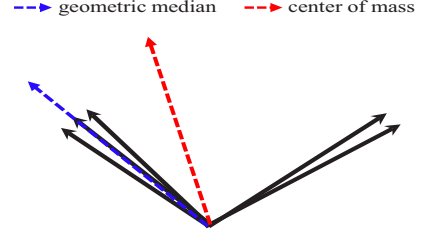
According to [23], the IV in the TF domain (also known as active IV) can be defined as

$$\mathbf{I}(k, l) := \mathcal{R}\{x_p^\star(k, l) \mathbf{x}_v(k, l)\} \in \mathbb{R}^{3 \times 1}, \quad (5)$$

where $\mathcal{R}\{\cdot\}$ and $(\cdot)^\star$ denote the real parts of a complex number and the complex conjugate operation, respectively. The IV reveals the flux of energy in the sound field by recording the direction and magnitude of the propagating wavefront [24]. Note that, the direction of IV can be regarded as an estimated DOA associated with each TF bin.

In this section, the proposed IV-based DOA estimation pipeline for selecting appropriate TF bins dominated by a single source is presented.

3.1. Key frame identification block

For frames with significant contribution from a single source, DOA estimation is expected to have low estimation error [15]. Therefore, the identification of key frames that dominated by a single source is critical for accurate DOA estimation. It has been noted that, time averaging and frequency smoothing are able to increase the effective rank of the correlation matrix up to the total number of active sources (coherent or not) [11]. Therefore, for the frames that dominated by a single active source, the effective ranks are expected to be unity. According to [11], the set of key frames can be identified by

$$\mathcal{L} = \{ l \mid erank(\hat{\mathbf{R}}(l)) = 1 \}, \quad (6)$$

$$erank(\hat{\mathbf{R}}(l)) = 1, \quad if \ \frac{\lambda_1(l)}{\lambda_2(l)} \geq \varepsilon. \quad (7)$$

Here $erank(\cdot)$ represents the effective rank, ε is a threshold and $\lambda_1(l)$ and $\lambda_2(l)$ are the largest and second-largest singular values of the sample correlation matrix $\hat{\mathbf{R}}(l)$, respectively. By smoothing over frequency and averaging L' adjacent time frames, the correlation matrix is estimated as

$$\hat{\mathbf{R}}(l) = \frac{1}{KL'} \sum_{k=1}^{K} \sum_{l'=0}^{L'-1} \mathbf{x}(k, l - l') \mathbf{x}^H(k, l - l'). \quad (8)$$

Here KL' is the size of the TF window, K is the length of Fourier transform (FT) and $\{\cdot\}^H$ represents Hermitian transposition. The frequency smoothing can be achieved by averaging over frequency due to the frequency-independent property of the array manifold in (1). Note that, the key frame identification is a special case of the DPD test under the assumption that single source dominates the key frame.

Figure 2 shows an example of the proposed key frame identification. The Y-axis is the normalized ratio of the largest to the second largest eigenvalue and the dotted line represents the
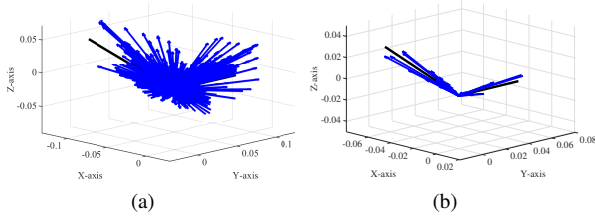
(a)   (b)

Figure 4: *An example of the distribution of the IVs (a) before and (b) after selecting single source dominated TF bins. The blue arrow lines represent IVs and the black lines indicate the true DOAs.*

Table 1: *The average angular error between IVs and the closest true DOA and the RMSAE of DOA estimation.*

	Angular separation between sources (deg)				
	25	45	60	100	180
Angular error without selection	$60.62°$	$57.11°$	$54.75°$	$49.01°$	$45.13°$
Angular error with selection	$6.31°$	$6.54°$	$6.16°$	$5.57°$	$5.40°$
RMSAE of the DOA estimation	$4.59°$	$5.22°$	$4.69°$	$3.09°$	$2.48°$

$Note$: Each angular separation is evaluated by 100 random source-position configurations. The SNR and T_{60} are set to $20 dB$ and $0.35 s$, respectively.

threshold. As can be seen, the frames that are greater than ε are considered to have an effective rank of unity and are therefore selected as key frames.

3.2. TF bin selection block

3.2.1. Geometric median extraction of key frame

Although most of the IVs in the key frames are concentrated, there are still extreme outliers caused by reverberation. The property of geometric median is exploited in the proposed TF bin selection block to alleviate the impact of extreme outliers, improving the robustness to reverberation.

Given the set $\mathcal{L}$, the geometric median [25] in the ℓth key frame is defined as

$$\tilde{\mathbf{u}}_{gm}(\ell) = \arg\min_{\tilde{\mathbf{y}}} \sum_{k=1}^{K} \|\tilde{\mathbf{I}}(k,\ell) - \tilde{\mathbf{y}}\|_2, \quad \forall \ell \in \mathcal{L}, \quad (9)$$

where $\tilde{\mathbf{I}}(k,l) = \mathbf{I}(k,l)/\|\mathbf{I}(k,l)\|$ is the normalized IV. The normalization ensures that the geometric median is only dependent on the directions of IVs. Figure 3 shows an example of the geometric median and the center of mass of a set of IVs. As can be seen, the geometric median is determined locally, while the center of mass depends on all global IVs. Given the geometric median $\tilde{\mathbf{u}}_{gm}(\ell)$, the IVs close to $\tilde{\mathbf{u}}_{gm}(\ell)$ are labeled as inliers and kept, while others are removed. The set of inliers can be obtained as

$$\Gamma = \{\tilde{\mathbf{I}}(k,\ell) \mid \angle\{\tilde{\mathbf{I}}(k,\ell), \tilde{\mathbf{u}}_{gm}(\ell)\} \leq \theta, \forall \ell \in \mathcal{L}\}. \quad (10)$$

Here $\angle\{\cdot, \cdot\}$ denotes the inter-angle between two vectors and θ is a pre-defined neighbourhood threshold.

3.2.2. Outliers removal

Given the set Γ, as well as the number of source J, $\tilde{\mathbf{I}}(k,\ell) \in \Gamma$ can be classified into J groups according to their directions, i.e., the normalized IVs with similar direction will be classified into the same group. In this work, the k-medoids algorithm is used to solve the classification problem [26]. After the clustering operation, J classified groups as well as J corresponding centroids, denoted by $\{\mathcal{G}_i\}_{i=1}^J$ and $\{\tilde{\mathbf{c}}_i\}_{i=1}^J$, respectively, are generated. Within each cluster, the IVs pointing away from the cluster centroid, i.e., $\forall \tilde{\mathbf{I}}(k,\ell) \in \mathcal{G}_i$ satisfies

$$\angle\{\tilde{\mathbf{I}}(k,\ell), \tilde{\mathbf{c}}_i\} \geq \varphi_{thr} \quad (11)$$

will be identified as outliers and removed. Here φ_{thr} is a pre-defined threshold. The existence of outliers is caused by erroneous key frames. After the outlier removal, the set that contains all remaining inliers is denoted by $\mathcal{S}$, and the final multi-source DOA estimation is performed based on $\mathcal{S}$.

Figure 4 shows an example of the distribution of the IVs before and after selecting single source dominated TF bins. Compared with Figure 4(a), the distribution of selected IVs in Figure 4(b) suggests that the proposed TF bin selection technique is capable of eliminating contaminated IVs.

3.3. DOA extraction block

There are two types of methods to extract the final multiple DOAs from the set of DOA estimates, namely the peak detection technique and clustering-based technique [27]. The peak detection technique usually require evaluating a cost function over a grid of candidate azimuth and elevation. The clustering-based technique classifies the set of DOA estimates into several clusters and output their centroids as the final DOAs. The effect of reverberation and noise may yield erroneous DOA estimates. Therefore, it is necessary to select the appropriate TF bins. In this work, given the selected $\mathcal{S}$, we cluster $\tilde{\mathbf{I}}(k,l) \in \mathcal{S}$ and output the J centroids as the final estimated DOAs.

4. Simulation Results

This section presents simulation results to demonstrate the correctness and effectiveness of the proposed method. For the simulation setup, the room impulse response (RIR) is generated using image methods [28] with room dimensions of $8m \times 6m \times 4m$ and AVS position at $(4m, 3m, 1.5m)$. Speech sources are mounted $1.5m$ away from the AVS. These source signals are formed by male and female speeches sampled from the TIMIT database [29] with a sampling frequency of 16 kHz. The frame length of STFT is set to 128 samples and there is 75% overlap between frames. For the thresholds, $\varepsilon = 0.3$ is set for key frame identification, $\theta = 5°$ and $\varphi_{thr} = 10°$ are used in TF bin selection block.

The proposed pipeline is compared with two state-of-the-art (SOTA) algorithms: the DPD test algorithm [14] and LRSS points detection algorithm [21]. The well-known MUltiple SIgnal Classification (MUSIC) algorithm [30] is also included as a baseline. To have a fair comparison, the empirical thresholds of these algorithms are set according to the corresponding papers. For J estimated $\{\hat{\mathbf{u}}_i\}_{i=1}^J$, accuracy is evaluated using the average angular error

$$e = \frac{1}{J} \sum_{i=1}^{J} 2\sin^{-1} \frac{\|\hat{\mathbf{u}}_i - \mathbf{u}_i\|}{2}. \quad (12)$$

The root-mean-square angular error (RMSAE), defined as $RMSAE = \sqrt{\mathbb{E}\{e^2\}}$, is used to quantify the DOA performance, where $\mathbb{E}$ denotes expectation operator.

Table 1 quantitatively shows the accuracy of the proposed pipeline. As can be seen from Table 1, the average angular error
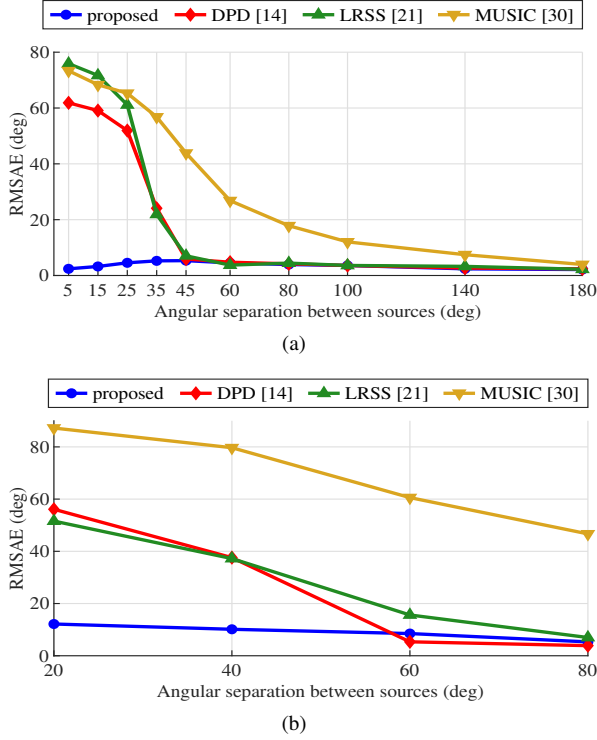
(a)

(b)

Figure 5: *RMSAE versus angular separation between sources in (a) two active sources, (b) three active sources. Each angular separation is evaluated by 100 random source-position configurations. The SNR and T_{60} are set to $20dB$ and $0.35s$, respectively.*

(a)

(b)

Figure 6: *RMSAE versus (a) T_{60} with SNR is set to $20dB$, (b) SNR with $T_{60} = 0.35s$. Two active sources are separated at an angle of $60°$. The results are generated by 100 random source-position configurations.*

is greatly reduced with the proposed single source dominated TF bin selection pipeline, leading to a better DOA estimation performance.

Figure 5 shows the DOA performance of the proposed pipeline and other SOTA algorithms for two and three simultaneously active sources scenarios. It can be observed from Figure 5(a) that the proposed pipeline achieves the lowest error compared to the SOTA algorithms as well as the MUSIC algorithm. Although the SOTA algorithms achieve fairly good performance when the angular separation is greater than $45°$, the RMSAE of these algorithms increases significantly as the angular separation decreases, especially when the angular difference is smaller than $35°$. The reason for the poor performance of LRSS algorithm in adjacent sources scenarios is that the similar directions of adjacent sources may fool the LRSS detection rule. The accuracy and robustness of the proposed pipeline in adjacent sources scenarios are benefited from the key frame identification and the subsequent geometric median extraction. Similar trends can be found in Figure 5(b). As the number of sources increase, although the performance of all algorithms degrades, the proposed pipeline is relatively stable, and outperforms all the other algorithms.

Figure 6 shows the results of DOA performance versus different T_{60} and SNR settings. In general, the performances of all algorithms degrade with increased reverberation time and noise level. Figure 6(a) shows how the RMSAE varies with T_{60}. It can be observed that the proposed pipeline achieves the lowest error compared to the other three algorithms, i.e., the proposed pipeline is less sensitive to reverberation. This is because both the extraction of geometric median and the removal
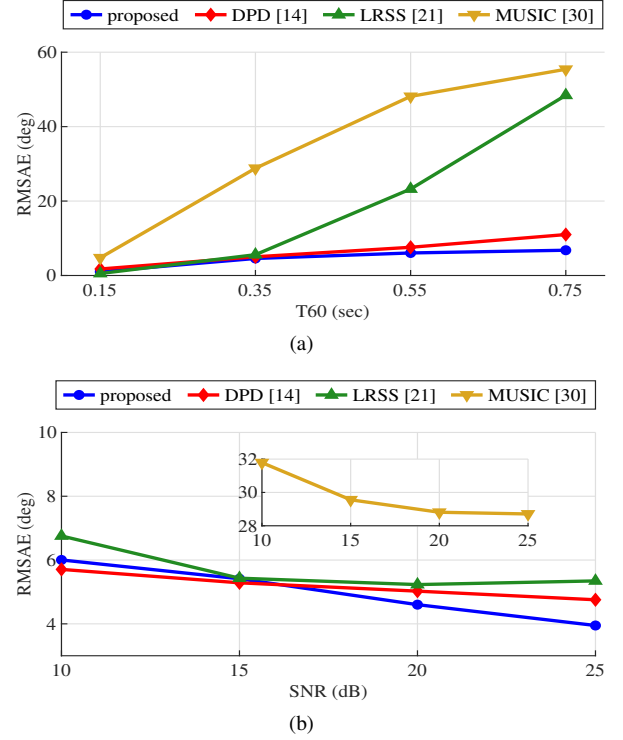
of outliers improve the robustness to reverberation. The reason for the significant performance degradation of LRSS algorithm in strong reverberation condition is the dramatic decrease of detected LRSS points. From Figure 6(b), we can see that the performance of all algorithms become better with increasing SNR and the proposed pipeline outperforms the other three algorithms. The robustness to noise of the proposed pipeline results from the inliers selection and outlier removal based on directions.

5. Conclusions

We propose a IV-based DOA estimation pipeline for selecting appropriated TF bins. The proposed pipeline mainly consists of three parts: key frame identification block, TF bin selection block and DOA extraction block. The key frame identification block identifies the key frames by frame-wisely examining the effective rank. The TF bin selection block exploits the geometric median of IVs to improve the robustness to extreme outliers. The DOA extraction block partitions the remaining vectors into several clusters and outputs the centroid of each cluster as the final estimated DOA. Simulation results show that the proposed pipeline is effective, especially in adjacent sources scenario. The robust identification ability of the proposed pipeline in adjacent sources scenarios is due to the key frame identification and the subsequent geometric median extraction. The robustness to reverberation results from both the extraction of geometric median and the removal of outliers.

6. Acknowledgements

7. References

[1] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 4, pp. 692–730, 2017.

[2] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone arrays.* Springer, 2001, pp. 157–180.

[3] C. Zhang, D. Florêncio, D. E. Ba, and Z. Zhang, "Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings," *IEEE Trans. Multimedia*, vol. 10, no. 3, pp. 538–548, 2008.

[4] Y. Huang, J. Chen, and J. Benesty, "Immersive audio schemes," *IEEE Signal Process. Mag.*, vol. 28, no. 1, pp. 20–32, 2010.

[5] I. A. McCowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Trans. Speech, Audio Process.*, vol. 11, no. 6, pp. 709–716, 2003.

[6] J. G. Desloge, W. M. Rabinowitz, and P. M. Zurek, "Microphone-array hearing aids with binaural output. i. fixed-processing systems," *IEEE Trans. Speech, Audio Process.*, vol. 5, no. 6, pp. 529–542, 1997.

[7] I. Potamitis, H. Chen, and G. Tremoulis, "Tracking of multiple moving speakers with multiple microphone arrays," *IEEE Trans. Speech, Audio Process.*, vol. 12, no. 5, pp. 520–529, 2004.

[8] A. Schwarz and W. Kellermann, "Coherent-to-diffuse power ratio estimation for dereverberation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 6, pp. 1006–1018, 2015.

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

[10] J. Cao, J. Liu, J. Wang, and X. Lai, "Acoustic vector sensor: reviews and future perspectives," *IET Signal Process.*, vol. 11, no. 1, pp. 1–9, 2016.

[11] O. Nadiri and B. Rafaely, "Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 10, pp. 1494–1505, 2014.

[12] S. Hafezi, A. H. Moore, and P. A. Naylor, "3d acoustic source localization in the spherical harmonic domain based on optimized grid search," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 415–419.

[13] D. P. Jarrett, E. A. Habets, and P. A. Naylor, "3d source localization in the spherical harmonic domain using a pseudo intensity vector," in *Proc. 18th Eur. Signal Process. Conf.*, 2010, pp. 442–446.

[14] A. Moore, C. Evers, P. A. Naylor, D. L. Alon, and B. Rafaely, "Direction of arrival estimation using pseudo-intensity vectors with direct-path dominance test," in *Proc. 23rd Eur. Signal Process. Conf.*, 2015, pp. 2296–2300.

[15] S. Hafezi, A. H. Moore, and P. A. Naylor, "Multiple source localization using estimation consistency in the time-frequency domain," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 516–520.

[16] S. Mohan, M. E. Lockwood, M. L. Kramer, and D. L. Jones, "Localization of multiple acoustic sources with small arrays using a coherence test," *J. Acoust. Soc. Amer.*, vol. 123, no. 4, pp. 2136–2147, 2008.

[17] M. Aktas and H. Ozkan, "Acoustic direction finding using single acoustic vector sensor under high reverberation," *Digital Signal Process.*, vol. 75, pp. 56–70, 2018.

[18] V. G. Reju, S. N. Koh, and Y. Soon, "An algorithm for mixing matrix estimation in instantaneous blind source separation," *Signal Process.*, vol. 89, no. 9, pp. 1762–1773, 2009.

[19] D. Pavlidi, S. Delikaris-Manias, V. Pulkki, and A. Mouchtaris, "3d localization of multiple sound sources with intensity vector estimates in single source zones," in *Proc. 23rd Eur. Signal Process. Conf.*, 2015, pp. 1556–1560.

[20] K. Wu, V. Reju, and A. W. Khong, "Multi-source direction-of-arrival estimation in a reverberant environment using single acoustic vector sensor," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 444–448.

[21] K. Wu, V. G. Reju, and A. W. Khong, "Multisource doa estimation in a reverberant environment using a single acoustic vector sensor," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1848–1859, 2018.

[22] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, no. 2, pp. 236–243, 1984.

[23] F. J. Fahy and V. Salmon, "Sound intensity," 1990.

[24] V. Pulkki, S. Delikaris-Manias, and A. Politis, *Parametric time-frequency domain spatial audio.* Wiley Online Library, 2018.

[25] S. P. Fekete, J. S. Mitchell, and K. Beurer, "On the continuous fermat-weber problem," *Operations Research*, vol. 53, no. 1, pp. 61–76, 2005.

[26] E. Schubert and P. J. Rousseeuw, "Faster k-medoids clustering: improving the pam, clara, and clarans algorithms," in *Proc. Int. conf. similarity search applicat.*, 2019, pp. 171–187.

[27] S. Hafezi, A. H. Moore, and P. A. Naylor, "Multi-source estimation consistency for improved multiple direction-of-arrival estimation," in *Proc. Hands-free Speech Commun. Microphone Arrays*, 2017, pp. 81–85.

[28] E. A. Habets, "Room impulse response generator," *Technische Universiteit Eindhoven, Tech. Rep*, vol. 2, no. 2.4, p. 1, 2006.

[29] J. S. Garofolo, "Timit acoustic phonetic continuous speech corpus," *Linguistic Data Consortium*, 1993.

[30] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas propag.*, vol. 34, no. 3, pp. 276–280, 1986.