A SLIDE-SAVE BASED FRAMEWORK FOR MULTI-SOURCE DOA EXTRACTION WITH CLOSELY SPACED SOURCES

Jianhua Geng, Sifan Wang and Xin Lou

ShanghaiTech University, Shanghai, China
{gengjh,wangsf,louxin}@shanghaitech.edu.cn

ABSTRACT

In adjacent sources scenarios, the low angular separation between active sources may degrade the performance of direction-of-arrival (DOA) estimation. In this work, we propose a slide-save based framework to address the problem of extracting multi-source DOAs for closely spaced sources. The basic idea is to identify the DOA estimates corresponding to the locally most dominant source within a sliding time-frequency (TF) window. Three different schemes are introduced to determine the critical DOA estimates in each TF window. The final DOAs are extracted using the retained DOA estimates by extending the histogram-based, clustering-based and Gaussian Mixture Model (GMM)-based multi-source DOA extraction methods. In addition, other intensity-based algorithms can also be incorporated into the proposed framework. Simulation results show that the proposed framework is effective to estimate multi-source DOAs in adjacent sources scenarios.

Index Terms— Direction-of-arrival (DOA) estimation, histogram, adjacent sources, Gaussian Mixture Model (GMM).

1. INTRODUCTION

Direction-of-arrival (DOA) estimation for multiple acoustic sources in a reverberant and noisy environment is a fundamental though challenging task in acoustic signal processing society [1–4]. In practical application scenarios, DOA estimation is generally required to be capable of dealing with interferences of reverberation and simultaneously active sources, especially sources with low angular separation.

Recently, a wealth of sophisticated intensity-based approaches are proposed to estimate DOAs of multiple sources in adverse environments [5–7]. Intensity-based methods are able to estimate a rough DOA at each time-frequency (TF) bin. However, it is challenging to accurately extract multi-source DOAs from the set of rough DOA estimates, when the speech sources are located close to each other. Generally, there are three types of approaches to extract multi-source DOAs, which are histogram-based [8–11], clustering-based [12–14] and model-based [15–17] approaches. In [11], a 2D histogram (azimuth and elevation) is constructed by utilizing the DOA estimates from the detected single source zones (SSZs). Thereafter, [11] extracts the final DOAs through processing the histogram which is smoothed with a 2D Gaussian filter. In [14], rough DOAs are estimated at the TF bins that passed the direct-path-dominant (DPD) test. The k-means clustering is then adopted to partition the DOA estimates and obtain the final DOAs. In [15], the Gaussian Mixture Model (GMM) is introduced to fit DOA estimates at the selected low-reverberant-single-source (LRSS) points. Thereafter, [15] extracts final DOAs from the model parameters, which are estimated via an expectation-maximization (EM) algorithm [18]. All the aforementioned multi-source DOA extraction techniques may merge the adjacent sources and misclassify other interferences as sources, resulting in extreme errors when the sources are close to each other.

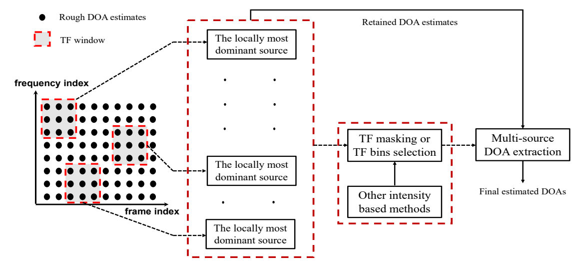

Fig. 1. Overview of the proposed framework.

In this work, we propose a framework to address the misclassification problems of multi-source DOA extraction in adjacent sources scenarios. Fig. 1 shows the overview of the proposed framework. The basic idea is to use a slide-save method, which processes the rough DOA estimates within a sliding TF window and saves the DOA estimates corresponding to the locally most dominant source. Three different schemes are introduced to determine these critical DOA estimates within each TF window. After traversing all TF windows, the final DOAs are extracted from the retained DOA estimates by improved multi-source DOA estimation algorithms. In addition, sophisticated intensity-based algorithms such as the SSZ [11], DPD [14] and LRSS [15], can also be incorporated into the proposed framework to obtain a more accurate DOA estimation.

This work was supported in part by the Natural Science Foundation of China under Grant 61801292 and in part by the Shanghai Rising-Star Program under Grant 21QC1401400.

2. REVIEW OF THE HISTOGRAM-BASED AND GMM-BASED TECHNIQUES

For histogram-based methods, let $\mathbf{C}(\psi, \phi)$ denote the histogram with azimuth and elevation candidate (ψ, ϕ). To well distinguish the irregular peaks, spatial smoothing is applied on the histogram as

$$\mathbf{C}_s(\psi, \phi) = \sum_{a=-N}^{N} \sum_{b=-N}^{N} \mathbf{h}_s(a, b)\mathbf{C}(\psi - a, \phi - b), \quad (1)$$

where $\mathbf{h}_s(a, b) = \frac{1}{2\pi\sigma_s^2} \exp \frac{-(a^2+b^2)}{2\sigma_s^2}$ is a 2D Gaussian filter with size $(2N+1) \times (2N+1)$ and $\mathbf{C}_s(\psi, \phi)$ is the smoothed histogram. Given the number of sources J, the final DOAs are selected as

$$\{(\psi_i, \phi_i)\}_{i=1}^{J} = \underset{(\psi,\phi)}{\arg\max\mathrm{k}}(\mathbf{C}_s(\psi, \phi), J). \quad (2)$$

Here $\mathrm{maxk}(\mathbf{A}, k)$ denotes the top k maxima of $\mathbf{A}$.

For GMM-based techniques, let d represent a DOA estimate. Then the probability density function of the GMM can be expressed as

$$p(d) = \sum_{i=1}^{J} w_i \mathcal{N}(d \mid \mu_i, \sigma_i), \quad (3)$$

where the mean μ_i, variance σ_i^2 and weight w_i are usually estimated via an EM algorithm [18]. The final DOA cues are derived from the model parameters.

3. THE PROPOSED FRAMEWORK

3.1. Analysis of Multiple DOAs Extraction Techniques

According to (1), spatial smoothing is applied to emphasize the peaks of $\mathbf{C}(\psi, \phi)$. However, a strong smoothing may result in the close peaks corresponding to different sources erroneously being merged in adjacent sources scenarios. Fig.2(a) and Fig.2(b) show the comparison between the original histogram and the strongly smoothed histogram in adjacent sources scenarios. As can be seen from Fig. 2(b), the peaks are merged into one after strong smoothing. On the other hand, a weak smoothing may result in irregular peaks corresponding to one active source being identified as multiple sources in the cases of widely separated sources. Fig. 2(c) and Fig. 2(d) show the distribution of peaks in this case. As we can see from Fig. 2(d), the existence of the top two peaks A and B corresponding to the same actual source may result in the neglect of peak C (corresponding to an actual source). For clustering-based techniques such as k-means clustering, although they guarantee the required number of sources, classifying rough DOA estimates by directions may merge two adjacent sources into one cluster and misclassify the other outliers as the second cluster, when the sources are spatial closely separated. This case of misclassification can be seen from Fig. 3(a), whereas Fig. 3(b) shows the relatively correct classification using the proposed clustering-based framework. For GMM-based techniques, the performances also deteriorate in adjacent sources scenarios, since it is difficult to distinguish the mean directions μ_i from the rough DOA estimates. The essential reason of these misclassifications is that, in adjacent sources scenarios, these multi-source DOAs extraction techniques classify DOA estimates according to their directions, while too many outliers far from the actual DOAs are incorrectly retained. These outliers will be constrainedly identified as speech sources to compensate the defect in the number of sources caused by the merger of adjacent sources.

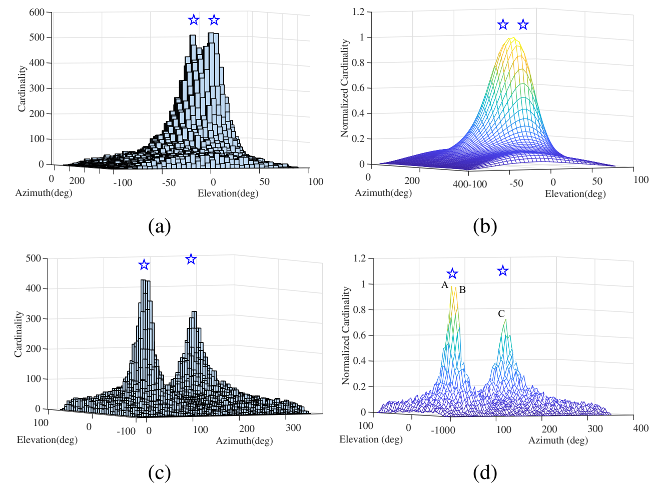

Fig. 2. Comparison of (a) original and (b) strong smoothing histogram in two adjacent sources scenarios, (c) original and (d) weak smoothing histogram in two well separated sources scenarios. The stars indicate the actual DOAs and σ_s is set to 2.5 and 0.05 for (b) and (d), respectively.

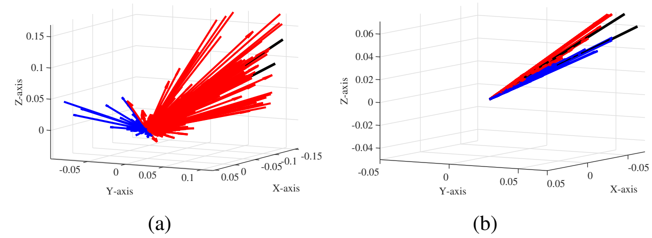

Fig. 3. An example of DOA estimates in Cartesian coordinates. (a) Misclassification and (b) correct classification. The arrow lines represent the vectors of DOA estimates and the black lines indicate the actual DOAs.

3.2. Overview of the Proposed Framework

To address the problem of misclassification, a slide-save based framework is proposed. Let $d(t, f)$ be the DOA estimate at TF bin (t, f). Given a TF window $\mathcal{W}_i$, let us define the set of DOA estimates close to d_{ci} as

$$\mathcal{D}(d_{ci}, \mathcal{W}_i) = \{d(t, f) | \measuredangle\{d(t, f), d_{ci}\} \leq \theta, (t, f) \in \mathcal{W}_i\}, \quad (4)$$

where $\measuredangle\{\cdot, \cdot\}$ measures the angle between two directions and d_{ci} is the core direction corresponding to the most dominant source within $\mathcal{W}_i$. After retrieving $\mathcal{W}_i$ and determining the core direction d_{ci} for $i = 1, ..., M$, the set of critical DOA estimates corresponding to all the locally most dominant sources can be obtained by

$$\Lambda = \mathcal{D}(d_{c1}, \mathcal{W}_1) \cup \mathcal{D}(d_{c2}, \mathcal{W}_2) \cup ... \cup \mathcal{D}(d_{cM}, \mathcal{W}_M). \quad (5)$$

Here M is the total number of sliding windows and $\cup$ denotes the union of two sets. We process $d \in \Lambda$ to extract the final DOAs by extending the conventional multi-source DOAs extraction techniques.

Various methods can be applied to determine the core direction d_{ci} and implement the proposed framework. Without loss of generality, we introduce the histogram-based, clustering-based and GMM-based schemes to determine d_{ci} and extract final DOAs with the proposed framework.

3.3. Implementation of the Proposed Framework

3.3.1. Implementation based on histogram

For histogram-based scheme, the core direction d_{ci} is determined by

$$d_{ci} = \arg\max_{(\psi,\phi)} \mathbf{C}_{si}(\psi, \phi), \quad (6)$$

where the smoothed histogram $\mathbf{C}_{si}(\psi, \phi)$ is constructed by the rough DOAs estimated from TF bins within $\mathcal{W}_i$. Substituting d_{ci} into (4) and (5), we obtain the set of critical DOA estimates Λ. Instead of searching J maxima according to (2), we search J highest peaks of $\mathbf{C}_s(\psi, \phi)$ by an iterative procedure, where $\mathbf{C}_s(\psi, \phi)$ is constructed based on $d \in \Lambda$. Let δ_m be the contribution of the mth detected peak (ψ_m, ϕ_m). Similar to [11], the contribution δ_m is calculated by

$$\delta_m = \mathbf{C}_s^m(\psi, \phi) \odot \mathbf{h}_r(\psi - \psi_m, \phi - \phi_m), \quad (7)$$

where $\odot$ denotes element-wise multiplication and $\mathbf{h}_r$ is a 2D Gaussian filter with zero mean and variance σ_r^2. The contribution from the smoothed histogram are then removed, i.e., $\mathbf{C}_s^{m+1}(\psi, \phi) \leftarrow \mathbf{C}_s^m(\psi, \phi) - \delta_m$. We proceed to detect (ψ_{m+1}, ϕ_{m+1}) and calculate δ_{m+1} from $\mathbf{C}_s^{m+1}(\psi, \phi)$ until reach the number of sources J. Afterwards, $\{(\psi_m, \phi_m)\}_{m=1}^J$ are saved as the final DOAs.

3.3.2. Implementation based on clustering

For clustering-based approaches, we set $d_{ci} = (\psi_k, \phi_k)$, where (ψ_k, ϕ_k) represents the direction of the center of the kth cluster. Let $\mathcal{C}_k$ denote the kth cluster formed by DOA estimates within $\mathcal{W}_i$, the index k is determined by

$$k = \arg\max_k \{Card(\mathcal{C}_1), ..., Card(\mathcal{C}_k), ..., Card(\mathcal{C}_J)\}, \quad (8)$$

where $Card(\cdot)$ counts the number of elements in each cluster. Given Λ by substituting d_{ci} into (5), we then adopt k-medoids clustering [19] to partition $d \in \Lambda$ into J clusters and output the corresponding J centers as the final DOAs.

Table 1. The Values of Parameters in Simulation Setup.

Notation	Description of parameters	Value
δ_s	the standard deviation of 2D filter $\mathbf{h}_s$	1
N	the parameter to control the size of $\mathbf{h}_s$	4
$\mathbf{C}_s$	the size of 2D histogram $\mathbf{C}_s$	73×37
$\mathcal{W}_i$	the size of TF windows $\mathcal{W}_i$	125×6
δ_r	the standard deviation of 2D filter $\mathbf{h}_r$	6.5
θ	the threshold in (4)	$5°$

3.3.3. Implementation based on GMM

For GMM-based approaches, we derive d_{ci} from μ_j, where μ_j is the mean direction of jth component and the index j is given by

$$j = \arg\max_j \left\{ \frac{w_1}{\sqrt{2\pi}\sigma_1}, ..., \frac{w_j}{\sqrt{2\pi}\sigma_j}, ..., \frac{w_J}{\sqrt{2\pi}\sigma_J} \right\}. \quad (9)$$

The set Λ is then obtained by substituting d_{ci} into (5). Different from (3), we model $d \in \Lambda$ with $J + 1$ Gaussian distributions, i.e., an extra Gaussian distribution is introduced to describe the outlier component. According to [15], the GMM can be expressed as

$$\begin{aligned} p(d) &= \sum_{i=1}^{J} w_i \mathcal{N}(d \mid \mu_i, \sigma_i) + w_o \mathcal{N}(d \mid \mu_o, \sigma_o) \\ &= \sum_{i=1}^{J+1} w_i \mathcal{N}(d \mid \mu_i, \sigma_i), \end{aligned} \quad (10)$$

where μ_o, σ_o and w_o are mean, standard variance and weight of the outlier component, respectively. Let $\{\mu_i, \sigma_i, w_i\}_{i=1}^{J+1}$ denote the parameter set of $J + 1$ components derived via an EM algorithm. The index corresponding to the outlier component can be determined by

$$o = \arg\min_i \left\{ \frac{w_1}{\sqrt{2\pi}\sigma_1}, ..., \frac{w_i}{\sqrt{2\pi}\sigma_i}, ..., \frac{w_{J+1}}{\sqrt{2\pi}\sigma_{J+1}} \right\}. \quad (11)$$

After removing the outlier component, we output other mean directions $\{\mu_i\}_{i=1}^J$ as the final estimated DOAs.

4. SIMULATION RESULTS

This section presents simulation results to show the effectiveness and accuracy of the proposed framework. For simulation setup, a size of $8\text{m} \times 6\text{m} \times 4\text{m}$ virtual room with an acoustic vector sensor (AVS) located at $(4\text{m}, 3\text{m}, 1.5\text{m})$ is simulated. The room impulse response (RIR) is generated using image methods [20]. Speech sources sampled from the TIMIT database [21] with a sampling frequency of 16 kHz are mounted 1.5m away from the AVS. The frame length of STFT is set to 128 samples and there is 75% overlap between frames. Table. 1 lists the parameters used in this work.

In this work, the original histogram-based, clustering-based and GMM-based multi-source DOAs extraction algorithms, labeled by "HIS", "CLU" and "GMM", respectively, are compared with the corresponding proposed framework, labeled by "FWHIS", "FWCLU" and "FWGMM", respectively. For CLU and FWCLU, the well-established k-medoids

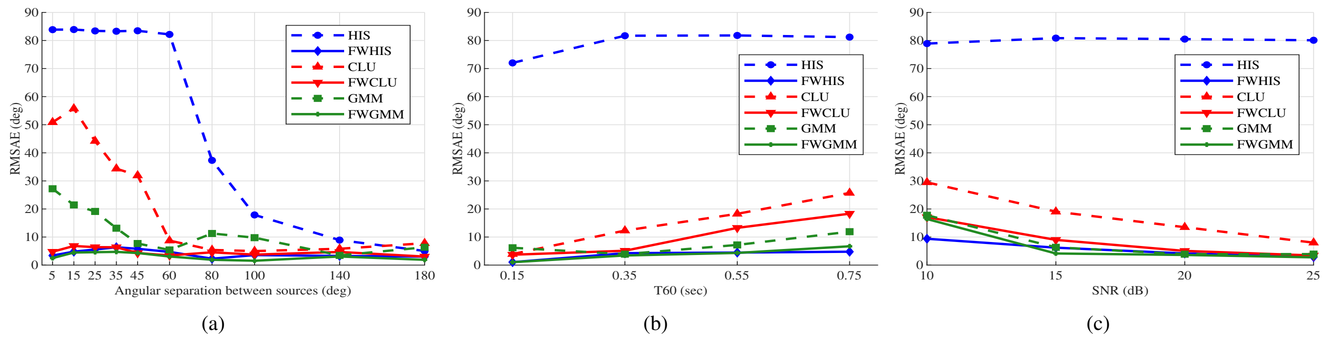

Fig. 4. RMSAE versus (a) angular separation, (b) T_{60}, (c) SNR. The RMSAE is averaged over 100 Monte Carlo trials. Without additional instructions, two active sources are separated at $60°$ and the SNR and T_{60} are set to $20\mathrm{dB}$ and $0.35\mathrm{s}$, respectively.

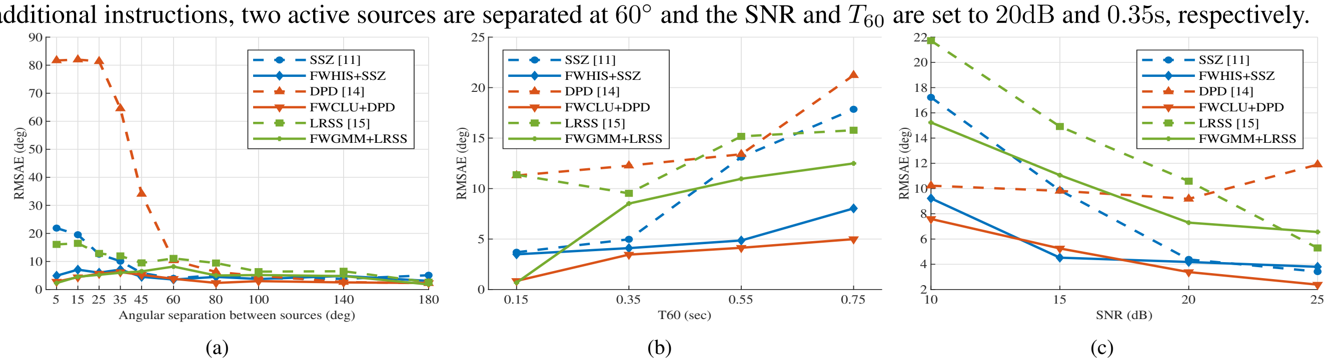

Fig. 5. RMSAE versus (a) angular separation, (b) T_{60}, (c) SNR. The RMSAE is averaged over 100 Monte Carlo trials. Without additional instructions, two active sources are separated at $60°$ and the SNR and T_{60} are set to $20\mathrm{dB}$ and $0.35\mathrm{s}$, respectively.

algorithm [19] is employed. In addition, we also compare the performances of SSZ algorithm [11], DPD test algorithm [14] and LRSS algorithm [15], with these algorithms incorporated into the proposed framework, represented as "FWHIS+SSZ", "FWCLU+DPD" and "FWGMM+LRSS". For J final DOAs $\{d_i\}_{i=1}^{J}$, accuracy is evaluated using the average angular error $e = \frac{1}{J}\sum_{i=1}^{J} \measuredangle\{d_i, (\psi_i, \phi_i)\}$. The root-mean-square angular error (RMSAE), defined as $\sqrt{\mathbb{E}\{e^2\}}$, is used to quantify the DOA performance across all trails, where $\mathbb{E}$ denotes expectation operator.

Fig. 4 shows the RMSAE of the compared methods in various scenarios. As we can see from Fig. 4(a), HIS has the worst performance and cannot even properly work when the angular separation is smaller than $60°$. Although the CLU and GMM achieve fairly good performance when angular separation is greater than $60°$, the RMSAE of these two methods increases significantly as the active sources become closer. Fig. 4(a) also shows that the proposed frameworks, i.e., FWHIS, FWCLU and FWGMM, work relatively stable with an average error less than $10°$. It can be observed from Fig. 4(b) and Fig. 4(c) that the performances of all methods degrade with increasing reverberation time and noise level. Compared with HIS, CLU and GMM, the errors of the corresponding framework FWHIS, FWCLU and FWGMMA are all reduced.

Fig. 5 shows the comparison of the existing algorithms and their corresponding counterparts where these algorithms are incorporated into the proposed framework. As can be seen from Fig. 5(a), all algorithms perform well when the angular separation is greater than $60°$. When two active sources become closer, the DPD algorithm works worst and the performances of SSZ and LRSS also degrade. Whereas, the proposed frameworks perform stable in almost all cases. Fig. 5(b) and Fig. 5(c) show that with these algorithms incorporated into the proposed framework, the errors of DOA estimation become lower. These results verify that the proposed framework is effective for multi-source DOA estimation.

5. CONCLUSIONS

We propose a framework to extract multi-source DOAs from a set of rough DOA estimates. The proposed framework first identifies the DOA estimates corresponding to the locally most dominant source within a sliding TF window. Three schemes, which are based on histogram, clustering and GMM, are introduced to determine the critical DOA estimates in each TF window. Based on the retained DOA estimates, final DOAs are obtained by extended multi-source DOA extraction approaches. Simulation results show that the proposed framework is effective, especially in adjacent sources scenarios. In addition, we show that more accurate DOA estimations can be achieved by incorporating other intensity-based approaches, such as SSZ algorithm, DPD algorithm and LRSS algorithm, into the proposed framework.

6. REFERENCES

[1] Itay Yehezkel Karo, Tsvi G Dvorkind, and Israel Cohen, "Source localization with feedback beamforming," *IEEE Trans. Signal Process.*, vol. 69, pp. 631–640, 2020.

[2] Shoko Araki, Tomohiro Nakatani, Hiroshi Sawada, and Shoji Makino, "Blind sparse source separation for unknown number of sources using gaussian mixture model fitting with dirichlet prior," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2009, pp. 33–36.

[3] Yiteng Huang, Jingdong Chen, and Jacob Benesty, "Immersive audio schemes," *IEEE Signal Process. Mag.*, vol. 28, no. 1, pp. 20–32, 2010.

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

[5] Areeb Riaz, Xiyu Shi, and Ahmet Kondoz, "Adaptive blind moving source separation based on intensity vector statistics," *Speech Comm.*, vol. 113, pp. 1–14, 2019.

[6] Jerome Daniel and Sran Kitic, "Time domain velocity vector for retracing the multipath propagation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 421–425.

[7] Kai Wu, Vaninirappuputhenpurayil Gopalan Reju, and Andy WH 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.

[8] Sina Hafezi, Alastair H Moore, and Patrick 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.

[9] Sina Hafezi, Alastair H Moore, and Patrick A Naylor, "Multiple source localization in the spherical harmonic domain using augmented intensity vectors based on grid search," in *Proc. 24th Eur. Signal Process. Conf.*, 2016, pp. 602–606.

[10] Anthony Griffin, Despoina Pavlidi, Matthieu Puigt, and Athanasios Mouchtaris, "Real-time multiple speaker doa estimation in a circular microphone array based on matching pursuit," in *Proc. 20th Eur. Signal Process. Conf.*, 2012, pp. 2303–2307.

[11] Despoina Pavlidi, Symeon Delikaris-Manias, Ville Pulkki, and Athanasios 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.

[12] Christine Evers, Alastair H Moore, and Patrick A Naylor, "Multiple source localisation in the spherical harmonic domain," in *Proc. 14th Int. Workshop Acoust. Signal Enhancement*, 2014, pp. 258–262.

[13] Sina Hafezi, Alastair H Moore, and Patrick 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.

[14] AlastairH Moore, Christine Evers, Patrick A Naylor, David L Alon, and Boaz 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] Maoshen Jia, Yuxuan Wu, Changchun Bao, and Christian Ritz, "Multi-source doa estimation in reverberant environments by jointing detection and modeling of time-frequency points," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 379–392, 2020.

[16] Xiaoyi Chen, Wenwu Wang, Yingmin Wang, Xionghu Zhong, and Atiyeh Alinaghi, "Reverberant speech separation with probabilistic time–frequency masking for b-format recordings," *Speech Commun.*, vol. 68, pp. 41–54, 2015.

[17] Banu Gunel, Huseyin Hacihabiboglu, and Ahmet M Kondoz, "Acoustic source separation of convolutive mixtures based on intensity vector statistics," *IEEE Trans, Audio, Speech, Lang. Process.*, vol. 16, no. 4, pp. 748–756, 2008.

[18] Arthur P Dempster, Nan M Laird, and Donald B Rubin, "Maximum likelihood from incomplete data via the em algorithm," *J. Roy. Statist. Soc. Ser. B, Methodol.*, vol. 39, no. 1, pp. 1–22, 1977.

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

[20] Emanuel AP Habets, "Room impulse response generator," *Technische Universiteit Eindhoven, Tech. Rep*, vol. 2, no. 2.4, pp. 1, 2006.

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