Sparse Bayesian Learning for Acoustic Source Localization

by

Ruchi Pandey, Santosh Nannuru, Aditya Siripuram

Report No: IIIT/TR/2021/-1

Centre for Communications
International Institute of Information Technology
Hyderabad - 500 032, INDIA
June 2021

ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9413960

SPARSE BAYESIAN LEARNING FOR ACOUSTIC SOURCE LOCALIZATION

Ruchi Pandey[†]*, Santosh Nannuru*[†]*, and Aditya Siripuram*[*]

[†] IIIT Hyderabad, SPCRC, Hyderabad, India, [*] Indian Institute of Technology Hyderabad, India

ABSTRACT

The localization of acoustic sources is a parameter estimation problem where the parameters of interest are the direction of arrivals (DOAs). The DOA estimation problem can be formulated as a sparse parameter estimation problem and solved using compressive sensing (CS) methods. In this paper, the CS method of sparse Bayesian learning (SBL) is used to find the DOAs. We specifically use multi-frequency SBL leading to a non-convex optimization problem, which is solved using fixed-point iterations. We evaluate SBL along with traditional DOA estimation methods of conventional beamforming (CBF) and multiple signal classification (MUSIC) on various source localization tasks from the open access LOCATA dataset. The comparative study shows that SBL significantly outperforms CBF and MUSIC on all the considered tasks.

Index Terms— DOA estimation, MUSIC, Compressive sensing, Sparse Bayesian learning, LOCATA challenge.

1. INTRODUCTION

Sound source localization and tracking using sensor arrays has applications in many areas including advanced driver assistant systems, hearing aids, smart home appliances, drones for rescue operations, etc. Various DOA estimation algorithms have been proposed in the literature such as conventional beamforming (CBF [1]), minimum variance distortionless response (MVDR [2]), generalized cross correlation phase transform (GCC-PHAT [3]), and multiple signal classification (MUSIC [4]) along with their multiple variants. CBF is robust to noise but has poor resolution and hence can fail to localize closely spaced sources. Though MUSIC is a high resolution method, it requires large number of snapshots. In environments posing challenges such as noise and reverberation, it is desired to have high resolution methods which work with fewer snapshots.

Compressive sensing (CS) or sparse signal processing is a technique to solve sparse problems by using fewer measurements [5]. Basis pursuit (BP) [6], orthogonal matching pursuit (OMP) [7] and focal underdetermined system solver (FOCUSS) [8] are some of the popular convex optimization based CS methods. Sparse Bayesian learning (SBL [9, 10]) is a CS method derived within Bayesian framework which gives fast solution to a non-convex optimization problem using fixed-point iterations. Its probabilistic formulation allows simultaneous processing of multiple snapshots [11] as well as multiple frequencies [12]. For localization of audio sources which have a rich frequency spectrum, it is advantageous to use multi-frequency SBL [13, 14].

The IEEE-AASP Challenge on sound source localization and tracking (LOCATA [15,16]) provides an open-access data corpus of indoor multi-channel audio recordings in presence of multiple mobile sources and their ground truth for performance evaluation. Since its release, various methods for localization have been applied on this dataset (see review [17]. None of the methods have explored CS based processing for DOA estimation.

In this paper, we demonstrate that SBL is a promising method for DOA estimation task using LOCATA dataset. Estimation challenges in the dataset include near-field effects, reverberation, and multiple moving sources and arrays. We implement DOA estimation techniques of CBF, MUSIC, and SBL. The algorithms are evaluated for Tasks 1, 3, 4, and 5, using 3 different microphone arrays. Both azimuth and elevation directions are estimated. The paper organization as follows: In Section 2 details of localization algorithms are given, in Section 3 localization results on LOCATA dataset are reported followed by conclusions in Section 4.

2. DOA ESTIMATION

2.1. CBF

Conventional beamforming [1] is one of the simplest DOA estimation method. The angular power spectrum for CBF is

$$\mathbf{P}_{CBF}(\theta,\phi) = \mathbf{a}^H(\theta,\phi)\,\mathbf{S_y}\,\mathbf{a}(\theta,\phi), \tag{1}$$

where $\mathbf{a}(\theta,\phi)$ is the array steering vector corresponding to a source located at azimuth ϕ and elevation θ. $\mathbf{S_y} = \frac{1}{L}\mathbf{Y}\mathbf{Y}^H$ is the sample covariance matrix computed using L snapshots arranged column-wise in the matrix $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_L]$. For an array with N sensors, the lth snapshot $\mathbf{y}_l$ is an N length complex vector. While processing multi-frequency data, the angular power spectrum is averaged across the frequencies.

2.2. MUSIC

MUSIC [4] is a high resolution method for source DOA estimation. It decomposes the signal covariance matrix ($\mathbf{S_y}$) into

two orthogonal subspace: signal subspace ($\mathbf{E}_s$) and noise subspace ($\mathbf{E}_n$). The MUSIC spectrum is then given by

$$\mathbf{P}_{mu}(\theta,\phi) = \frac{\mathbf{1}}{\mathbf{a}^H(\theta,\phi)\mathbf{E}_n\mathbf{E}_n^H\mathbf{a}(\theta,\phi)}. \quad (2)$$

When $\mathbf{a}^H(\theta,\phi)$ is orthogonal with columns of $\mathbf{E}_n$, the value of the denominator is zero (or close to zero when noise is present) and $\mathbf{P}_{mu}(\theta,\phi)$ shows a peak corresponding to source DOAs.

Eq. (1) and (2) provide expressions for narrowband CBF and MUSIC spectrum. While processing multi-frequency data, the spectra are averaged across the frequency range.

2.3. SBL

For fth frequency, the lth snapshot can be expressed as $\mathbf{y}_{fl} = \mathbf{A}_f\mathbf{x}_{fl} + \mathbf{n}_{fl}$ where the dictionary matrix $\mathbf{A}_f$ has columns consisting of steering vectors $\mathbf{a}_f(\theta,\phi)$ and (θ,ϕ) range over the 2D search grid in azimuth and elevation. The source amplitudes $\mathbf{x}_{fl}$ are assumed sparse and $\mathbf{n}_{fl}$ models the zero-mean complex Gaussian noise with covariance $\sigma_f^2\mathbf{I}$. The multi-snapshot observation is $\mathbf{Y}_f = \mathbf{A}_f\mathbf{X}_f + \mathbf{N}_f, f = 1,\ldots,F$. Under assumptions of independence across snapshots and frequencies, likelihood is given as

$$p(\mathbf{Y}_{1:F}|\mathbf{X}_{1:F}) = \prod_{f=1}^{F} p(\mathbf{Y}_f|\mathbf{X}_f) = \prod_{f=1}^{F}\prod_{l=1}^{L} p(\mathbf{y}_{fl}|\mathbf{x}_{fl}). \quad (3)$$

In multi-snapshot, multi-frequency, SBL formulation [11,12], the source amplitudes $\mathbf{x}_{fl}$ are modeled as independent, zero-mean, complex Gaussian vectors with same diagonal covariance $\mathbf{\Gamma} = \text{diag}(\boldsymbol{\gamma}) = \text{diag}([\gamma_1,\ldots,\gamma_M])$ giving the prior

$$p(\mathbf{X}_{1:F}) = \prod_{f=1}^{F} p(\mathbf{X}_f) = \prod_{f=1}^{F}\prod_{l=1}^{L} p(\mathbf{x}_{fl}). \quad (4)$$

The sparsity of source amplitude vectors is related to the sparsity of the parameter vector $\boldsymbol{\gamma}$. As prior and likelihood are assumed to be Gaussian, the evidence $p(\mathbf{Y}_{1:F})$ is also Gaussian

$$p(\mathbf{Y}_{1:F}) = \prod_{f=1}^{F} p(\mathbf{Y}_f) = \prod_{f=1}^{F}\prod_{l=1}^{L} \mathcal{CN}(\mathbf{y}_{fl};\mathbf{0},\mathbf{\Sigma}_f), \quad (5)$$

where $\mathbf{\Sigma}_f = \sigma_f^2\mathbf{I} + \mathbf{A}_f\mathbf{\Gamma}\mathbf{A}_f^H$ and $\mathcal{CN}()$ denotes complex Gaussian density function. The SBL method estimates the unknown parameter $\boldsymbol{\gamma}$ by maximizing the evidence $p(\mathbf{Y}_{1:F})$

$$\hat{\boldsymbol{\gamma}} = \arg\max_{\boldsymbol{\gamma}} p(\mathbf{Y}_{1:F}) \quad (6)$$

$$= \arg\min_{\boldsymbol{\gamma}} \sum_{f=1}^{F}\sum_{l=1}^{L}\left(\mathbf{y}_{fl}^H\mathbf{\Sigma}_f^{-1}\mathbf{y}_{fl} - \log|\mathbf{\Sigma}_f|\right). \quad (7)$$

To find the minimum of this non-convex objective function, we differentiate with respect to $\boldsymbol{\gamma}$ and equate to zero. For details about this procedure see [11, 12]. The resulting fixed-point update equation we obtain is

$$\gamma_m^{\text{new}} = \gamma_m^{\text{old}}\left(\frac{\sum_{f=1}^{F}\sum_{l=1}^{L}|\mathbf{y}_{fl}^H\mathbf{\Sigma}_f^{-1}\mathbf{a}_{fm}|^2}{\sum_{f=1}^{F}\mathbf{a}_{fm}^H\mathbf{\Sigma}_f^{-1}\mathbf{a}_{fm}}\right), \quad (8)$$

where γ_m is the mth element of $\boldsymbol{\gamma}$ and $\mathbf{a}_{fm}$ is the mth column of the dictionary matrix $\mathbf{A}_f$. At convergence, the estimate $\hat{\boldsymbol{\gamma}}$ is sparse [11, 12] which in turn enforces source amplitudes $\mathbf{x}fl$ to be sparse. The noise variance is estimated using maximum likelihood approach [11,12]. As γ_m is the source power corresponding to mth DOA, $\hat{\boldsymbol{\gamma}}$ is called the SBL power spectrum.

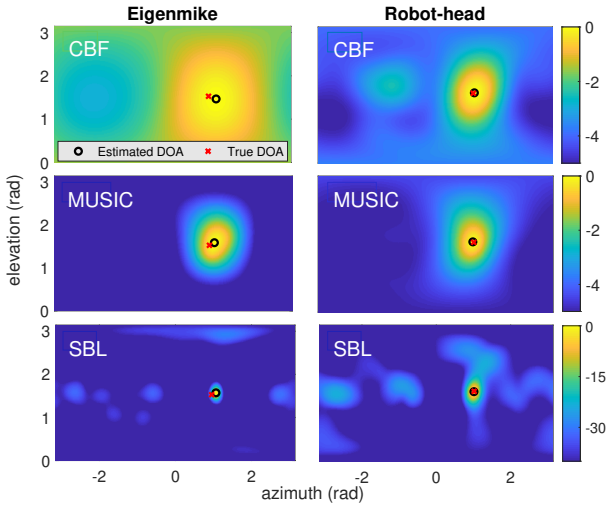

Fig. 1: Spectrum of CBF, MUSIC, and SBL for eigenmike and robot-head (Task 1, recording 1, 24^{th} block).

The power spectrum computed by CBF, MUSIC, and SBL for eigenmike and robot-head arrays are shown in Fig. 1. The spectrum are normalized to have a maximum value of 1 and plotted in log scale. For multi-frequency processing using CBF and MUSIC, an average across frequencies of their individual narrowband spectrum is performed. Since eigenmike has smaller aperture than robot-head, it shows broader peak regions. The figure clearly illustrates the resolution difference of the DOA estimation methods. CBF has a very wide peak region spread around the true DOA which reduces its resolving ability. MUSIC provides better resolution than CBF whereas SBL has the best resolution. Since SBL peak region is concentrated near true DOA, it can localize multiple near by sources with lesser ambiguity.

3. RESULTS

3.1. LOCATA dataset

The LOCATA [15] development dataset is used for performing localization. The LOCATA recordings have challenging scenarios such as near-field sources, reverberation, and

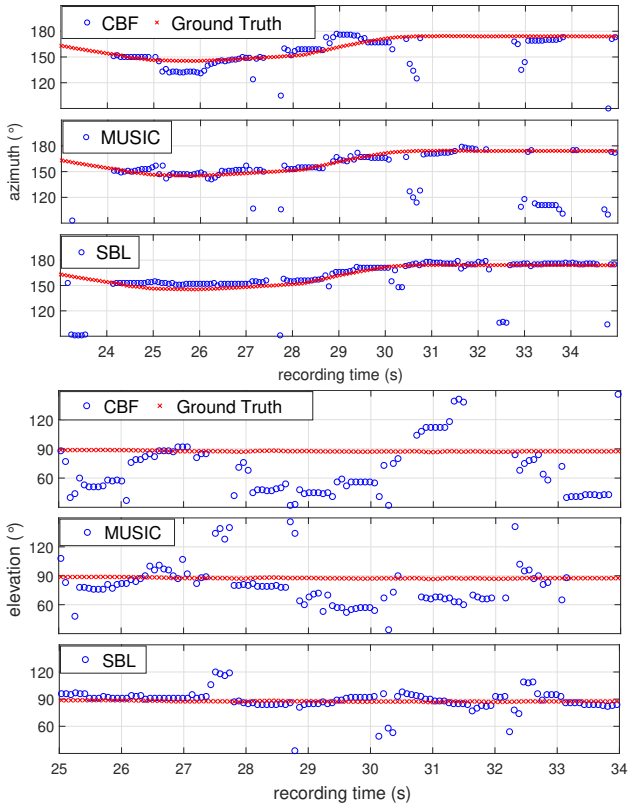

Fig. 2: DOA estimates of azimuth and elevation angles using robot-head, Task 5, recording 1.

ambient noise (from a road in front of the building). Further details and assumptions about the LOCATA dataset can be found in [15–17]. We consider a 12-microphone pseudo-spherical array named **robot-head**, 32-microphone spherical array named **eigenmike** and a 15-microphone non-uniform array named **dicit**. The arrays are non-planar and data is collected indoors. Task 1 consists of recordings of a single stationary talker, Task 3 for a moving talker, Task 4 for multiple moving talker and Task 5 for a single moving talker where the microphone array is moving as well. While processing data the following parameter values are used: FFT size of 1024, frequency range of [800, 2800] Hz, snapshot duration of 0.03 s, each block consists of 100 snapshots with 90% overlap.

3.2. Performance metrics

The DOAs are estimated as the peak location of the power spectrum computed by localization algorithms for each block. We compute spectrum with 1° resolution both in azimuth and elevation. The error between estimated and true DOAs are computed at block level during voice activity periods [18]. We compute mean absolute error, root mean square error (RMSE), and standard deviation of the estimates. All errors are averaged over all the recordings for each of the tasks.

We also compute probability of detection (P_d) for localization algorithms as $P_d(\zeta) = 1 - \frac{N_{\text{miss}}(\zeta)}{N_{\text{total}}}$, where $N_{\text{miss}}(\zeta)$ is the number of misdetections (over all the recordings in each task) and N_{total} is total number of blocks. A source is said to be misdetected if the estimated DOA is more than ζ° away from the true DOA.

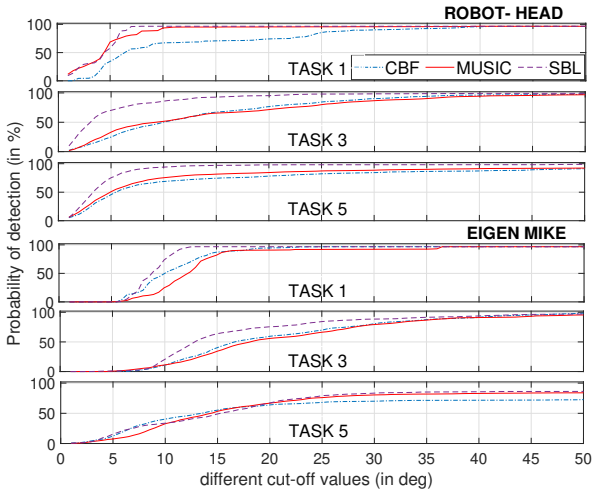

Fig. 3: Probability of detection (P_d) vs cutoff ζ.

3.3. Results for robot-head array

Task 1: Localization of single stationary talker is simplest of the tasks and all algorithms give relatively low error as seen from Table 1. The computed errors are least for SBL, followed by that of MUSIC and CBF. The probability of detection is similar for SBL and MUSIC and increases to 100% sharply (Fig. 3). In case of Task 1 using SBL, 97% of sources are detected within 10° of true DOA for robot-head.

Task 3: For Task 3, the moving talker causes its distance from the stationary microphone array to change and has poorer DOA estimation performance compared to Task 1. SBL outperforms both CBF and MUSIC in terms of localization errors as seen from Table 1. In terms of probability of detection SBL performs significantly better than CBF and MUSIC for low values of ζ (Fig. 3).

Task	Method	Mean		RMSE		Std Dev	
		az	el	az	el	az	el
T1	CBF	4.48	10.0	5.27	12.2	0.04	0.12
	MUSIC	1.96	3.94	2.20	5.91	0.01	0.06
	SBL	**1.10**	**3.57**	**1.25**	**3.72**	**0.01**	**0.01**
T3	CBF	4.37	6.09	8.37	12.3	0.12	0.18
	MUSIC	8.70	10.6	15.8	16.5	0.23	0.21
	SBL	**3.82**	**3.16**	**6.37**	**5.45**	**0.08**	**0.07**
T5	CBF	15.7	11.7	36.7	18.6	0.58	0.25
	MUSIC	8.36	9.85	23.4	17.2	0.58	0.24
	SBL	**2.98**	**3.93**	**10.5**	**7.57**	**0.17**	**0.11**

Table 1: Error performance of robot-head array for Tasks 1,3, and 5 (averaged over all recordings, in degrees)

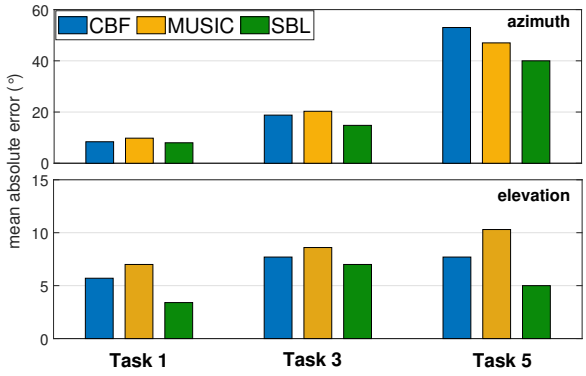

Fig. 4: Azimuth and elevation error using eigenmike for Task 1, 3 and 5 averaged over all recordings.

Task 5: In Task 5 the talker is moving as well as the microphone arrays installed on the platform are moving. The DOA estimates from a robot-head recording are shown in Fig. 2. The CBF and MUSIC estimates are often far from ground truth, whereas SBL estimates are closely aligned with the ground truth and result in lower errors (see Table 1).

3.4. Results for eigenmike array

The mean absolute error for Task 1, 3, and 5 using eigenmike is shown in Fig. 4. Due to rotations of the eigenmike within the shockmount, it is highly sensitive to scattering effects [17] which results in high azimuth error (Fig. 4, note that the two plots have different vertical range). The estimates of azimuth and elevation for an eigenmike recording from Task 3 are shown in Fig. 5. The estimates obtained from SBL are much closer to ground truth compared to CBF and MUSIC when voice activity is present (denoted by 1 in VAD plot).

3.5. Results for dicit array

We show results of localizing multiple sources from Task 4 recordings using dicit array. A 5-sensor linear subarray of dicit with 8 cm spacing is considered. The azimuth estimates for two moving sources are shown in Fig. 6. Overall, SBL estimates for both sources are more accurate compared to CBF and MUSIC for active voice period (VAD not shown). We also observe that CBF gives better estimates than MUSIC for dicit array. To avoid spatial aliasing, we processed 25 frequencies between 800–2100 Hz. It can be seen that dicit gives comparably higher error than robot-head and eigenmike.

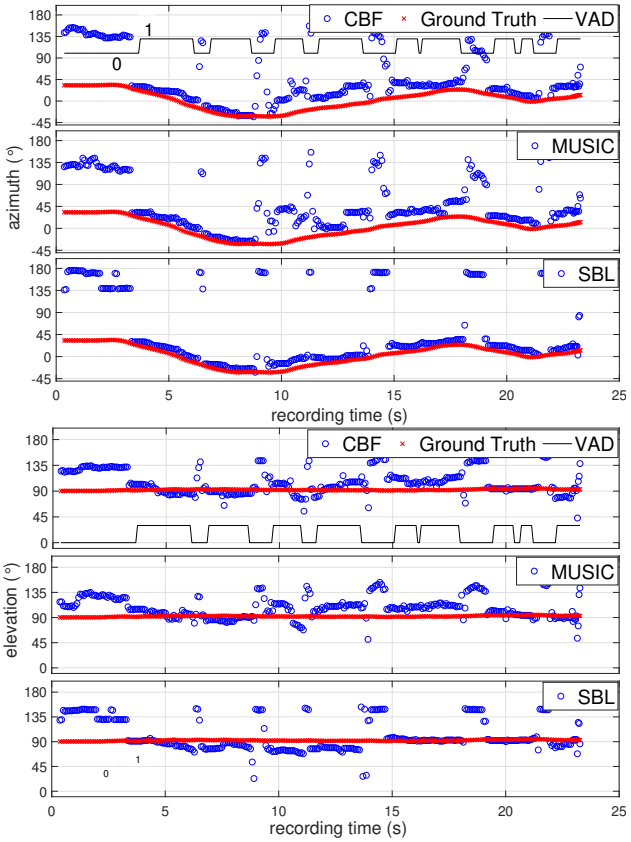

Fig. 5: Azimuth and elevation DOA estimates of single target using eigenmike array, Task 3, recording 2.

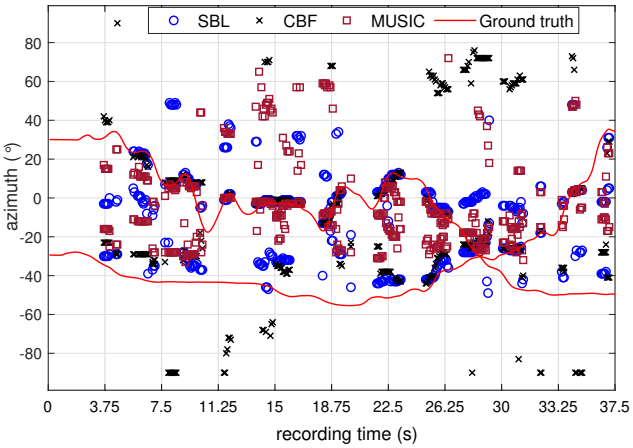

Fig. 6: Azimuth DOA estimates of two moving sources using dicit array, Task 4, recording 2.

4. CONCLUSIONS

In this paper, we have considered DOA estimation as compressive sensing problem and solved using sparse Bayesian learning algorithm. We show DOA estimation results for Tasks 1,3,4, and 5 of LOCATA dataset using three array structures. We have compared the performance of CBF, MUSIC, and SBL using various metrics. Results show that CBF and MUSIC work well for a stationary or slow moving source but are error prone in challenging tasks where source and/or array is moving. Multi-frequency SBL was observed to be robust to these challenges and performs well in all the tasks.

5. REFERENCES

[1] H. L. Van Trees, *Optimum Array Processing (Detection, Estimation, and Modulation Theory, Part IV)*, John Wiley & Sons, 2002.

[2] R. G. Lorenz and S. P. Boyd, "Robust minimum variance beamforming," *IEEE Trans. Sig. Process.*, vol. 53, no. 5, pp. 1684–1696, 2005.

[3] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust. Speech Sig. Process.*, vol. 24, no. 4, pp. 320–327, 1976.

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

[5] D. Malioutov, M. Çetin, and A. S. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Trans. Sig. Process.*, vol. 53, no. 8, pp. 3010–3022.

[6] S.S. Chen, D.L. Donoho, and M.A. Saunders, "Atomic decomposition by basis pursuit," *SIAM review*, vol. 43, no. 1, pp. 129–159, 2001.

[7] J. Tropp and A. C. Gilbert, "Signal recovery from partial information via orthogonal matching pursuit," *IEEE Trans. Inform. Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

[8] I. F. Gorodnitsky and B. D. Rao, "Sparse signal reconstruction from limited data using focuss: A re-weighted minimum norm algorithm," *IEEE Transactions on signal processing*, vol. 45, no. 3, pp. 600–616, 1997.

[9] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, Jun. 2001.

[10] D. P. Wipf and B. D. Rao, "An empirical Bayesian strategy for solving the simultaneous sparse approximation problem," *IEEE Trans. Sig. Process.*, vol. 55, no. 7, pp. 3704–3716, 2007.

[11] P. Gerstoft, C. F. Mecklenbräuker, A. Xenaki, and S. Nannuru, "Multisnapshot sparse Bayesian learning for DOA," *IEEE Sig. Process. Lett.*, vol. 23, no. 10, pp. 1469–1473, Oct. 2016.

[12] S. Nannuru, K. L. Gemba, P. Gerstoft, W. S. Hodgkiss, and C. F. Mecklenbräuker, "Sparse Bayesian learning with multiple dictionaries," *Sig. Process.*, vol. 159, pp. 159–170, 2019.

[13] K. L. Gemba, S. Nannuru, and P. Gerstoft, "Multi-frequency sparse Bayesian learning for matched field processing in non-stationary noise," *J. Acoust. Soc. Am.*, vol. 144, no. 3, pp. 1943–1943, 2018.

[14] K. L. Gemba, S. Nannuru, and P. Gerstoft, "Robust ocean acoustic localization with sparse Bayesian learning," *IEEE J. Sel. Topics Sig. Process.*, vol. 13, no. 1, pp. 49–60, 2019.

[15] H. W. Löllmann, C. Evers, A. Schmidt, H. Mellmann, P.A. Naylor H. Barfuss and, and W. Kellermann, "The LOCATA challenge data corpus for acoustic source localization and tracking," in *IEEE Sensor Array Multichannel Sig. Process. Workshop*, 2018, pp. 410–414.

[16] H. W. Löllmann, C. Evers, A. Schmidt, H. Mellmann, H. Barfuss, P.A. Naylor, and W. Kellermann, "LOCATA challenge-evaluation tasks and measures," 2018, pp. 565–569.

[17] C. Evers, H. W. Löllmann, A. Schmidt, H. Barfuss, P. A. Naylor, and W. Kellermann, "The locata challenge: Acoustic source localization and tracking," *IEEE/ACM Trans. on Audio, Speech, and Language Process.*, 2020.

[18] M. Van Segbroeck, A Tsiartas, and S. Narayanan, "A robust frontend for VAD: exploiting contextual, discriminative and spectral cues of human voice.," in *INTERSPEECH*, 2013, pp. 704–708.