INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Sparseness-Aware DOA Estimation with Majorization Minimization

Masahito Togami, Robin Scheibler

LINE Corporation, Japan

masahito.togami@linecorp.com, robin.scheibler@linecorp.com

Abstract

We propose a direction-of-arrival (DOA) estimation technique which assumes that speech sources are sufficiently sparse and there is only one active speech source at each time-frequency (T-F) point. The proposed method estimates the DOA of the active speech source at each T-F point. A typical way for DOA estimation is based on grid-searching for all possible directions. However, computational cost of grid-searching is proportional to the resolution of search area. Instead of accurate grid-searching, the proposed method adopts rough grid-searching followed by an iterative parameter optimization based on Majorization-Minimization (MM) algorithm. We propose a parameter optimization method which guarantees a monotonical increase of the objective function. Experimental results show that the proposed method estimates DOAs of speech sources more accurately than conventional DOA estimation methods when computational cost of each method is almost the same.

Index Terms: DOA estimation, sparseness, auxiliary function

1. Introduction

Direction-of-Arrival (DOA) estimation techniques are fundamental and enable the tracking of an active speaker in acoustic applications such as smart speaker applications, speaker diarization systems [1], humanoid robots [2], and so on. DOA estimation is also useful in speech source separation [3–7]. Typically, these applications have multiple microphones. Multiple microphone input signals contain DOA information of the active speaker in the form of differences between input signals, i.e., amplitude ratio, time difference, or phase difference. Generally speaking, it is possible to perform more accurate DOA estimation with more microphones. However, accurate DOA estimation suffers from high computational cost. Thus, one challenge is to perform accurate DOA estimation with many microphones with lower computational cost.

For a long time, several DOA estimation techniques have been studied [8–17]. Early works have been imported from the wireless communication research field, e.g., multiple signal classification (MUSIC) [10–12] and estimation of signal parameters via rotational invariance techniques (ESPRIT) [13]. Although original methods output discrete DOA estimates for narrow-band signals, there are several advanced techniques which output continuous DOA estimates for broadband signals [18–20]. Because these methods have not been developed for speech signals, characteristics of speech sources are not fully utilized. Steering response power (SRP) based methods [21–23] are DOA estimation techniques for broad-band signals such as speech signals. Because the SRP based methods are based on fixed-beamforming, the resolution is not enough when there are multiple speech sources.

Sparseness-aware DOA estimation techniques have been also proposed [24–28]. These techniques fully utilize characteristics of speech sources, i.e., W-disjoint orthogonality [29]. These techniques assume that speech sources are highly sparse in time-frequency (T-F) domain and there is at most one active speech source per each T-F point. In DUET [24], the DOA of the active speech source at each T-F point is estimated from phase differences between two microphones under the W-disjoint orthogonality assumption. Then, DOAs of multiple speech sources are estimated by peak-searching a histogram of phase differences across T-F axes. DUET with two microphones works well under anechoic and noiseless environments. On the other hand, DOA estimation accuracy degrades under reverberant and noisy environments. Althouth DOA estimation accuracy will be higher with a longer distance between two microphones, there is an upper limit on the allowable distance between two microphones due to the spatial aliasing problem [26].

There are several sparseness-aware DOA estimation techniques which overcome the spatial aliasing problem by utilizing more than three microphones [3, 26, 30]. Modified Delay-and-Sum Beamformer (MDSBF) [3] estimated the DOA of the active speech source at each time-frequency point by grid-searching for all possible directions. Because computational cost of grid-searching is proportional to the resolution of search area, fine grid-searching is typically problematic. Instead of fine grid-searching, stepwise phase difference restoration (SPIRE) methods [26, 30] optimize an approximated objective function with the Taylor expansion [30]. However, the optimization step is heuristically derived and there is less theoretical justification.

In this paper, we propose a sparseness-aware DOA estimation method which estimates the DOA of the active speech source at each T-F point without fine grid-searching. The DOA of the active speech source is estimated with rough grid-searching followed by an iterative parameter optimization. The objective function in the parameter optimization is based on the objective function of the MDSBF. Because a closed-form solution cannot be achieved, a quadratic surrogate function is derived based on an inequality originally proposed in the context of time delay estimation [31]. A solution which optimizes the proposed surrogate function can be obtained with lower computational cost than grid-searching. It is also guaranteed that the proposed method increases the objective function monotonically based on the Majorization-Minimization (MM) algorithm [32]. Experimental results show that the proposed method can perform more accurate DOA estimation than conventional methods with similar computational cost under reverberant and noisy environments.

2. Problem statement

2.1. Signal modeling

A recorded microphone input signal is transformed from time-domain into a T-F domain via short time Fourier transform. In the T-F domain, a multi-channel microphone input signal $\boldsymbol{x}_{lk} \in \mathbb{C}^{N_m}$ (l is the frame index, k is the frequency index, and N_m is

http://dx.doi.org/10.21437/Interspeech.2020-1168

the number of the microphones) is modeled as follows:

$$\boldsymbol{x}_{lk} = \sum_{i=1}^{N_s} s_{ilk} \boldsymbol{a}_{ik} + \boldsymbol{n}_{lk}, \tag{1}$$

where N_s is the number of the speech sources, s_{ilk} is the i-th speech source signal, $\boldsymbol{a}_{ik}$ is the steering vector of the i-th speech source, and $\boldsymbol{n}_{lk}$ is the background noise signal. The steering vector $\boldsymbol{a}_{ik}$ contains DOA information of the i-th speaker and it is parameterized with the DOA of the speech source as $\boldsymbol{a}_{ik} = \boldsymbol{a}_{\theta_i, \phi_i, k}$,

where θ_i and ϕ_i are the azimuth and elevation of the ith speech source, respectively. The active speech source and the noise signal are assumed to be uncorrelated with each other. It is also assumed that the noise signal in each microphone is uncorrelated with each other and it has the same power as $E\left[\boldsymbol{n}_{lk}\boldsymbol{n}_{lk}^H\right] = P(n)\boldsymbol{I}$, where H is the Hermitian transpose of a matrix/vector, E is the expectation operator, $\boldsymbol{I}$ is the N_m-dimensional identity matrix, and $P(n)$ is the expected value of the noise power. The objective is to estimate $\{\theta_i, \phi_i\}_{1 \le i \le N_s}$ from the observed microphone input signal $\{\boldsymbol{x}_{lk}\}_{1 \le l \le L, 1 \le k \le K}$.

2.2. W-disjoint orthogonality

In [29], it is assumed that speech sources are sufficiently sparse and there is at most one active speech source per each T-F point (W-disjoint orthogonality). Under the W-disjoint orthogonality assumption, the mixture is approximated as follows:

$$\boldsymbol{x}_{lk} \approx s_{lk} \boldsymbol{a}_{lk} + \boldsymbol{n}_{lk}, \tag{2}$$

where s_{lk} and $\boldsymbol{a}_{lk}$ are redefined as the source signal of the active speech source and the steering vector at the l-th frame and the k-th frequency, respectively. Let (θ_{lk}, ϕ_{lk}) be the DOA of the active speech source at the l-th frame and the k-th frequency as well. DOA histogram based methods [3, 26, 30] estimate DOAs of all speech sources as follows:

1. Estimate θ_{lk}, ϕ_{lk} from the microphone input signal $\boldsymbol{x}_{lk}$
2. DOAs of multiple speech sources are estimated by peak-searching a histogram of $\{\theta_{lk}, \phi_{lk}\}_{1 \le l \le L, 1 \le k \le K}$ across T-F axes.

2.3. Conventional approach of DOA estimation at each time-frequency point

Modified Delay and Sum Beamforming (MDSBF) [3] estimates $\boldsymbol{a}_{lk}$ by using grid-searching for all possible directions. Let $\boldsymbol{b}_{\theta,\phi,k}$ be the normalized steering vector of a possible direction (θ, ϕ), i.e., $\|\boldsymbol{b}_{\theta,\phi,k}\|_2^2 = 1$. Square of the inner product between $\boldsymbol{b}_k$ and $\boldsymbol{x}_{lk}$ can be calculated in the form of the expected value as follows:

$$P_{\theta,\phi,l,k} = E\left[\left|\boldsymbol{b}_{\theta,\phi,k}^H \boldsymbol{x}_{lk}\right|^2\right] = P_s \left|\boldsymbol{b}_{\theta,\phi,k}^H \boldsymbol{a}_{lk}\right|^2 + P_n, \tag{3}$$

where P_s is the expected value of the signal power. It is important to note that $P_{\theta,\phi,l,k}$ is maximized when $\boldsymbol{b}_{\theta,\phi,k} = \boldsymbol{a}_{lk}$. This means that we can estimate the DOA of the active speech source (θ_{lk}, ϕ_{lk}) by maximizing $P_{\theta,\phi,l,k}$ w.r.t. (θ, ϕ). Unfortunately, a closed-form solution which maximizes $P_{\theta,\phi,l,k}$ cannot be obtained. Instead, MDSBF estimates (θ_{lk}, ϕ_{lk}) via grid-searching for all possible directions with an approximation of the expected value as follows:

$$(\theta_{lk}, \phi_{lk}) = \underset{(\theta,\phi)\in\Omega}{\arg\max}\, P_{\theta,\phi,l,k} \approx \underset{(\theta,\phi)\in\Omega}{\arg\max} \left|\boldsymbol{b}_{\theta,\phi,k}^H \boldsymbol{x}_{lk}\right|^2, \tag{4}$$

where Ω is the set of all possible directions. The steering vector $\boldsymbol{b}_{\theta,\phi,k}$ is modeled as a function of θ, ϕ with the far-field assumption as follows:

$$b_{\theta,\phi,k,m} = \frac{1}{\sqrt{N_m}} \exp\left(j 2\pi f_k \frac{\boldsymbol{q}_{\theta,\phi}^T \boldsymbol{d}_m}{c}\right), \tag{5}$$

where $b_{\theta,\phi,k,m}$ is the m-th element of $\boldsymbol{b}_{\theta,\phi,k}$, j is the imaginary unit, f_k is the frequency [Hz] of the k-th frequency bin, c is the sound speed [m/s], $\boldsymbol{d}_m$ is the three-dimensional position vector of the m-th microphone, T is the transpose operator of a matrix/vector, and $\boldsymbol{q}_{\theta,\phi}$ is the position vector of a virtual speech source which comes from the (θ, ϕ) direction on the unit sphere, which is defined as $\boldsymbol{q}_{\theta,\phi} = \begin{pmatrix}\cos\theta\cos\phi & \sin\theta\cos\phi & \sin\phi\end{pmatrix}^T$. The MDSBF framework is quite solid, but one problem is that its computational cost is proportional to the number of grid points $|\Omega|$. Thus, the computational cost is typically problematic when the resolution is precise.

3. Proposed method

The proposed method avoids the fine grid-searching for all possible directions by incorporating an iterative parameter optimization technique based on Majorization-Minimization (MM) algorithm [32]. Instead of the original objective function, the parameter is updated so as to maximize the derived surrogate function.

3.1. Derivation of objective function

The objective function of the proposed method is based on the objective function of the MDSBF. The objective function $\mathcal{F}$ is derived under the assumption that all microphone input signals have the same amplitude at each time-frequency point as follows:

$$\mathcal{F}(\theta_{lk}, \phi_{lk}, \boldsymbol{x}_{lk}) = \left|\boldsymbol{b}_{\theta_{lk},\phi_{lk},k}^H \overline{\boldsymbol{x}}_{lk}\right|^2, \tag{6}$$

where $\overline{x}_{lkm}$ is the normalized input vector defined as $\frac{x_{lkm}}{|x_{lkm}|} = \exp(j\sigma_{lkm})$ and σ_{lkm} is the phase component of the m-th microphone input signal. The inner product between $\boldsymbol{b}_{\theta_{lk},\phi_{lk},k}$ and $\overline{\boldsymbol{x}}_{lk}$ can be expanded with (5) as follows:

$$\boldsymbol{b}_{\theta_{lk},\phi_{lk},k}^H \overline{\boldsymbol{x}}_{lk} = \frac{1}{\sqrt{N_m}} \sum_{m=1}^{N_m} \exp\left(j\sigma_{lkm} - j\alpha_k \boldsymbol{q}_{\theta_{lk},\phi_{lk}}^T \boldsymbol{d}_m\right), \tag{7}$$

where $\alpha_k = \frac{2\pi f_k}{c}$. Optimization of $\mathcal{F}$ at each time-frequency point w.r.t. $\boldsymbol{q}_{\theta_{lk},\phi_{lk}}$ under the constraint that $\|\boldsymbol{q}_{\theta_{lk},\phi_{lk}}\|_2^2 = 1$ is simpler than optimization of $\mathcal{F}$ w.r.t. (θ_{lk}, ϕ_{lk}) and these two optimizations are equivalent. Thus, the objective function $\mathcal{F}$ is rewritten as a function of $\boldsymbol{q}_{\theta_{lk},\phi_{lk}}$. The objective function $\mathcal{F}(\boldsymbol{q}_{\theta_{lk},\phi_{lk}}, \boldsymbol{x}_{lk})$ can be also expanded as follows:

$$\begin{aligned}\mathcal{F}(\boldsymbol{q}_{\theta_{lk},\phi_{lk}}, \boldsymbol{x}_{lk}) &= \sum_{m=1}^{N_m}\sum_{n=1}^{N_m} \frac{\exp\left(j\overline{\sigma}_{lkmn} - j\alpha_k \boldsymbol{q}_{\theta_{lk},\phi_{lk}}^T \boldsymbol{d}_{mn}\right)}{N_m} \\ &= 1 + \sum_{m=1}^{N_m}\sum_{n=m+1}^{N_m} \frac{2\cos\left(\overline{\sigma}_{lkmn} - \alpha_k \boldsymbol{q}_{\theta_{lk},\phi_{lk}}^T \overline{\boldsymbol{d}}_{mn}\right)}{N_m},\end{aligned} \tag{8}$$

where $\overline{\boldsymbol{d}}_{mn} = \boldsymbol{d}_m - \boldsymbol{d}_n$ and $\overline{\sigma}_{lkmn} = \sigma_{lkm} - \sigma_{lkn}$. Because it is impossible to optimize $\mathcal{F}(\boldsymbol{q}_{\theta_{lk},\phi_{lk}}, \boldsymbol{x}_{lk})$ w.r.t. $\boldsymbol{q}_{\theta_{lk},\phi_{lk}}$ in a closed-form manner, the proposed method optimizes a surrogate function of $\mathcal{F}(\boldsymbol{q}_{\theta_{lk},\phi_{lk}})$ instead. An initial value of $\boldsymbol{q}_{\theta_{lk},\phi_{lk}}$ is estimated by rough grid-searching based on MDSBF.

3.2. Derivation of surrogate function

In the MM algorithm [32], it is ensured that the objection function $\mathcal{F}(\boldsymbol{q}_{\theta_{lk},\phi_{lk}})$ is monotonically increasing by iteratively optimizing a surrogate function $\mathcal{F}^{+}(\boldsymbol{q}_{\theta_{lk},\phi_{lk}}, \boldsymbol{r})$ which fulfills the following equations:

$$\mathcal{F}(\boldsymbol{q}_{\theta_{lk},\phi_{lk}}) \geq \mathcal{F}^{+}(\boldsymbol{q}_{\theta_{lk},\phi_{lk}}, \boldsymbol{r}) \tag{9}$$

$$\mathcal{F}(\boldsymbol{q}_{\theta_{lk},\phi_{lk}}) = \max_{\boldsymbol{r}} \mathcal{F}^{+}(\boldsymbol{q}_{\theta_{lk},\phi_{lk}}, \boldsymbol{r}), \tag{10}$$

where $\boldsymbol{r}$ is an auxiliary variable. A surrogate function for $\cos\left(\overline{\sigma}_{lkmn} - \alpha_k \boldsymbol{q}_{\theta_{lk},\phi_{lk}}^{T} \overline{\boldsymbol{d}}_{mn}\right)$ is derived from an inequality proposed for sub-sample time delay estimation [31] as follows:

$$\begin{aligned}&\cos\left(\overline{\sigma}_{lkmn} - \alpha_k \boldsymbol{q}_{\theta_{lk},\phi_{lk}}^{T} \overline{\boldsymbol{d}}_{mn}\right)\\ &\geq \frac{-\sin\left(y_{lkmn}\right)\left(\overline{\sigma}_{lkmn} - \alpha_k \boldsymbol{q}_{\theta_{lk},\phi_{lk}}^{T} \overline{\boldsymbol{d}}_{mn} + 2z_{lkmn}\right)^2}{2y_{lkmn}}\\ &\quad + C,\end{aligned} \tag{11}$$

where C is a term that is not a function of $\boldsymbol{q}_{\theta_{lk},\phi_{lk}}$. The auxiliary variable $\boldsymbol{r}$ is $\{y_{lkmn}, z_{lkp}\}$ ($y_{lkmn} \in \mathbb{R}^1$ and $z_{lkmn} \in \mathbb{Z}^1$). A surrogate function $\mathcal{F}^{+}(\boldsymbol{q}_{\theta_{lk},\phi_{lk}}, \boldsymbol{r})$ can be obtained as follows:

$$\begin{aligned}&\mathcal{F}^{+}(\boldsymbol{q}_{\theta_{lk},\phi_{lk}}, \boldsymbol{r})\\ &= \sum_{p} \frac{-\sin\left(y_{lkp}\right)\left(\overline{\sigma}_{lkp} - \alpha_k \boldsymbol{q}_{\theta_{lk},\phi_{lk}}^{T} \overline{\boldsymbol{d}}_{p} + 2z_{lkp}\pi\right)^2}{N_m y_{lkp}} + C,\end{aligned} \tag{12}$$

where $p = (m, n)$ is the microphone pair index and z_{lkp} is set to

$$z_{lkp} = \underset{z \in \mathbb{Z}}{\arg\min} \left|\overline{\sigma}_{lkp} - \alpha_k \boldsymbol{q}_{\theta_{lk},\phi_{lk}}^{T} \overline{\boldsymbol{d}}_{p} + 2z\pi\right|. \tag{13}$$

$\mathcal{F}(\boldsymbol{q}_{\theta_{lk},\phi_{lk}})$ is equal to $\mathcal{F}^{+}(\boldsymbol{q}_{\theta_{lk},\phi_{lk}}, \boldsymbol{r})$ if and only if the following equation is satisfied:

$$y_{lkmn} = \overline{\sigma}_{lkmn} - \alpha_k \boldsymbol{q}_{\theta_{lk},\phi_{lk}}^{T} \overline{\boldsymbol{d}}_{mn} + 2z_{lkmn}. \tag{14}$$

3.3. Parameter optimization

Minimization of the quadratic surrogate function $\mathcal{F}^{+}(\boldsymbol{q}_{\theta_{lk},\phi_{lk}}, \boldsymbol{r})$ w.r.t. $\boldsymbol{q}_{\theta_{lk},\phi_{lk}}$ under the constraint $\|\boldsymbol{q}_{\theta_{lk},\phi_{lk}}\|_2^2 = 1$ is a generalized trust region subproblem [33]. It can be tackled by the method of Lagrange multipliers. $\mathcal{G}(\boldsymbol{q}_{\theta_{lk},\phi_{lk}}, \boldsymbol{r}, \lambda_{lk}) = \mathcal{F}^{+}(\boldsymbol{q}_{\theta_{lk},\phi_{lk}}, \boldsymbol{r}) - \lambda_{lk}\left(\|\boldsymbol{q}_{\theta_{lk},\phi_{lk}}\|_2^2 - 1\right)$ is optimized by setting the derivative of $\mathcal{G}$ to $\mathbf{0}$ as follows:

$$\boldsymbol{q}_{\theta_{lk},\phi_{lk}} = \left(\boldsymbol{D}_{lk} + \lambda_{lk}\boldsymbol{I}\right)^{-1} \boldsymbol{w}_{lk} \tag{15}$$

where $\boldsymbol{D}_{lk} = \alpha_k^2 \sum_p \frac{\sin\left(y_{lkp}\right)\overline{\boldsymbol{d}}_p \overline{\boldsymbol{d}}_p^T}{y_{lkp}}$ and $\boldsymbol{w}_{lk} = \alpha_k \sum_p \frac{\sin\left(y_{lkp}\right)\left(\overline{\sigma}_{lkp} + 2z_{lkp}\pi\right)\overline{\boldsymbol{d}}_p}{y_{lkp}}$. λ_{lk} should satisfy the following equation from the constraint $\|\boldsymbol{q}_{\theta_{lk},\phi_{lk}}\|_2^2 = 1$:

$$\boldsymbol{w}_{lk}^T \left(\boldsymbol{D}_{lk} + \lambda_{lk}\boldsymbol{I}\right)^{-T} \left(\boldsymbol{D}_{lk} + \lambda_{lk}\boldsymbol{I}\right)^{-1} \boldsymbol{w}_{lk} = 1. \tag{16}$$

The proposed method utilizes a bisection method to solve λ_{lk} similar to least-squares (LS) approaches based on squared range observations (SR-LS) [34]. Because (16) includes matrix inversion, a simplified equation without matrix inversion is derived. Eigenvalue decomposition of $\boldsymbol{D}_{lk}$ can be obtained as $\boldsymbol{D}_{lk} = \boldsymbol{V}_{lk}\boldsymbol{W}_{lk}\boldsymbol{V}_{lk}^T$, where $\boldsymbol{V}_{lk}$ is a matrix which contains eigenvectors and $\boldsymbol{W}_{lk}$ is a diagonal matrix which contains eigenvalues $\{W_{lki}\}_{i=1,2,3}$ in its diagonal elements. (16) can be expanded as follows:

Table 1: *Simulation configurations*

T_{60} [s]	Max order	Absorption
0.36	17	0.35
0.70	34	0.2

$$\left(\boldsymbol{D}_{lk} + \lambda_{lk}\boldsymbol{I}\right)^{-1} = \boldsymbol{P}_{lk}\left(\boldsymbol{W}_{lk} + \lambda_{lk}\boldsymbol{I}\right)^{-1} \boldsymbol{P}_{lk}^T. \tag{17}$$

Finally, we can obtain the following simplified equation:

$$\frac{a_1^2}{(\lambda_{lk} + W_{lk1})^2} + \frac{a_2^2}{(\lambda_{lk} + W_{lk2})^2} + \frac{a_3^2}{(\lambda_{lk} + W_{lk3})^2} = 1, \tag{18}$$

where a_i is the i-th element of the vector $\boldsymbol{P}_{lk}^T \boldsymbol{w}_{lk}$. We can search for λ_{lk} efficiently without matrix inversion based on (18). The search area of λ_{lk} is limited in $(-\min_i W_{lki}, \max_i \sqrt{N}\,|a_i| - W_{lki})$ under the assumption that the matrix $\boldsymbol{D}_{lk} + \lambda_{lk}\boldsymbol{I}$ is a positive-definite matrix.

3.4. Summary of proposed method

The proposed method is summarized as follows:

1. Initilize $\boldsymbol{q}_{\theta_{lk},\phi_{lk}}$ by rough grid-searching based on MDSBF
2. Update parameters in an iterative manner
   - Update z_{lkp} based on (13)
   - Update y_{lkp} based on (14)
   - Optimize λ_{lk} based on bisection search with (18)
   - Update $\boldsymbol{q}_{\theta_{lk},\phi_{lk}}$ based on (15)
3. Extract θ_{lk}, ϕ_{lk} from $\boldsymbol{q}_{\theta_{lk},\phi_{lk}}$
4. Make a DOA histogram of $\{\theta_{lk}, \phi_{lk}\}_{1 \leq l \leq L, 1 \leq k \leq K}$ across T-F axes
5. DOAs of active sources are estimated by peak-searching the DOA histogram

3.5. Relation to prior work

The proposed surrogate function is similar to the objective function of SPIRE [30]. The object function of SPIRE is derived by using a Taylor expansion which is equivalent to $\frac{\sin y_{lkp}}{y_{lkp}} = 1$ in (12). In the SPIRE, microphone pairs which are utilized in each iteration are determined manually. On the contrary, the proposed method determines a weight for each microphone pair $\frac{\sin y_{lkp}}{y_{lkp}}$ automatically so as to increase the objective function monotonically. Thus, the proposed method is interpreted as an extension of SPIRE with the MM algorithm. In this context, we call the proposed method SPIRE-MM.

4. Experiment

4.1. Setup

DOA estimation performance of the proposed method was eval-

Table 2: *Evaluation results*

Approach	RTF	RMSE [degrees] T_{60} : 0.36 [s]/0.70 [s] $N_s = 1$	2	3	4	Accuracy [%] T_{60} : 0.36 [s]/0.70 [s] 1	2	3	4
MDSBF (full) [3]	26.48	0.66/0.66	0.69/0.71	0.64/0.72	0.72/0.78	86.0/84.0	85.0/80.5	83.0/73.0	80.2/68.2
SRP-PHAT [21]	1.50	0.83/1.15	1.77/2.22	2.53/3.16	3.47/4.23	60.0/43.0	23.5/20.5	13.0/11.0	8.5/7.5
MUSIC [10]	3.06	**0.67**/0.75	0.92/1.18	1.32/2.28	1.50/2.25	**83.0**/73.0	55.5/41.0	38.3/26.7	35.5/27.5
MDSBF (half)	13.21	0.96/0.96	0.90/0.92	0.94/0.95	0.96/0.99	52.0/52.0	57.5/56.0	57.0/54.7	54.8/51.0
MDSBF (quarter)	6.60	1.21/1.21	1.28/1.28	1.28/1.28	1.38/1.38	32.0/32.0	31.5/31.5	29.3/29.0	28.5/28.5
SPIRE-MM 1	1.46	0.82/0.94	1.01/1.38	1.12/1.69	1.59/2.61	63.0/53.0	50.0/36.0	44.0/29.3	31.0/22.5
SPIRE-MM 2	2.99	0.70/0.77	0.77/0.84	0.70/0.91	0.81/1.07	75.0/66.0	76.0/65.0	76.3/56.3	66.2/47.5
SPIRE-MM 3	11.02	0.68/0.72	**0.71**/0.76	**0.68**/0.78	0.76/**0.88**	81.0/**78.0**	80.5/**75.0**	**77.7/69.7**	73.0/**59.5**
SPIRE-MM 4	6.46	0.68/**0.70**	0.72/**0.75**	0.69/**0.77**	**0.75**/0.93	**83.0/78.0**	**83.0**/74.0	76.3/67.7	**74.5**/55.5

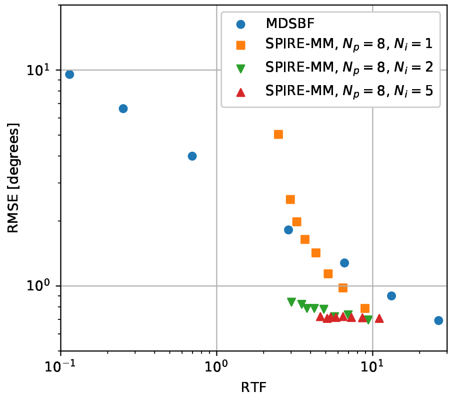

Figure 1: *Experimental results when* $T_{60} = 0.36$ *[s] and* $N_s = 2$*: Each point represents an evaluation result of median value of RMSE over 100 mixtures with a specific number of initial grid points* N_r

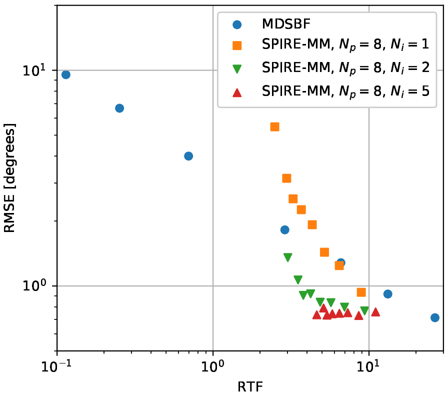

Figure 2: *Experimental results when* $T_{60} = 0.70$ *[s] and* $N_s = 2$

uated by using `Pyroomacoustics` [35]. Anechoic speech sources were extracted from CMU ARCTIC Concat15 dataset [36] which concatenates utterances extracted from the CMU Sphinx database [37]. `Pyroomacoustics` simulates reverberant mixtures in a 10 × 10 × 10 m room with two configurations shown in Table 1. The number of the speech sources N_s was set to 1, 2, 3, and 4. 100 reverberant mixtures were simulated for each condition. Sampling rate was 16000 Hz. Signal to Noise Ratio (SNR) between speech sources and background noise was set to 10 dB. The Pyramic microphone array [38] geometry ($N_m = 48$) was utilized. Frame size was 256. Frame shift was 128. The total number of frequency bins was 127. For DOA estimation, frequency bins from the 6-th bin to the 60-th bin were used. Real-Time Factor (RTF) was measured by using a mixture with 16.63 [s] duration. Distance between microphones and talkers was set to 3m. For evaluation of computational cost, a server with Intel Xeon CPU E5-2630 v4 2.20GHz CPU and 132 GB RAM was used. In the proposed method, we limit the number of the microphone pairs (m, n) as $m \in \{1 \cdots N_m\}$ and $n \in \{m + 1 \cdots \min(m + N_p, N_m)\}$.

4.2. Results

We evaluated median value of Root Mean Square Error (RMSE) [degrees] between the estimated DOA and the oracle one and percentage of estimation results within 1 degree error (Accuracy). The grid size in MDSBF (full), SRP-PHAT, and MUSIC was 180 × 90. MDSBF (full) is the upper-bound of the proposed method. The grid size in MDSBF (half) was the half of that in MDSBF (full). The grid size in MDSBF (quarter) was the quarter of that in MDSBF (full). The grid size for peak-searching the output histogram was set to 180 × 90. In the proposed method, the grid size for initial grid-searching N_r was set to a lower value. The number of MM iterations N_i was set to $1, 2, 5$. In SPIRE-MM 1,2,3,4, we adjusted N_p, N_i, and N_r to match the RTFs of SRP-PHAT, MUSIC, MDSBF (half), and MDSBF (quarter), respectively. Estimation results were shown in Table 2. The second best result was bolded. It is shown that the proposed SPIRE-MM outperformed the conventional methods whose RTFs are almost the same. Effectiveness of the proposed iterative parameter update was also evaluated in Fig. 1 and Fig. 2 with various number of initial grid points $N_r \in \{162, 450, 648, 900, 1296, 1800, 2592, 4050\}$. It is shown that RMSE is decreasing by increasing the number of the iterations N_i when the RTF is almost the same. Thus, it can be said that the proposed iterative parameter optimization based on the MM algorithm is effective.

5. Conclusions

We proposed a sparseness-aware DOA estimation method. The proposed method estimates the DOA of the active speech source at each frequency point in an iterative way. Monotonical increase of the objective function is guaranteed in the proposed method. Experimental results showed that the proposed method outperformed the conventional method when the real-time factor (RTF) is almost the same. It was also shown that DOA estimation performance increases with the proposed iterative parameter optimization.

6. Acknowledgements

Thank you for fruitful discussion with Mr. Kouei Yamaoka at Tokyo Metropolitan University.

7. References

[1] S. Araki, M. Fujimoto, K. Ishizuka, H. Sawada, and S. Makino, "A DOA based speaker diarization system for real meetings," in *2008 Hands-Free Speech Communication and Microphone Arrays*, 2008, pp. 29–32.

[2] V. Tourbabin and B. Rafaely, "Direction of arrival estimation using microphone array processing for moving humanoid robots," *IEEE/ACM Trans. ASLP*, vol. 23, no. 11, pp. 2046–2058, 2015.

[3] M. Togami, Y. Obuchi, and A. Amano, *Automatic Speech Recognition of Human-Symbiotic Robot EMIEW*. I-tech Education and Publishing, 2007, ch. 22, pp. 395–404.

[4] Y. Kawaguchi and M. Togami, "Soft masking based adaptation for time-frequency beamformers under reverberant and background noise environments," in *EUSIPCO 2010*, Aug. 2010, pp. 736–740.

[5] M. Togami, T. Sumiyoshi, Y. Obuchi, Y. Kawaguchi, and H. Kokubo, "Beamforming array technique with clustered multichannel noise covariance matrix for mechanical noise reduction," in *2010 18th European Signal Processing Conference*, 2010, pp. 741–745.

[6] M. Togami, Y. Kawaguchi, N. Nukaga, and Y. Obuchi, "Online MVBF adaptation under diffuse noise environments with MIMO based noise pre-filtering," in *2012 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA)*, 2012, pp. 292–297.

[7] S. Markovich-Golan, S. Gannot, and W. Kellermann, "Combined LCMV-TRINICON beamforming for separating multiple speech sources in noisy and reverberant environments," *IEEE/ACM Trans. ASLP*, vol. 25, no. 2, pp. 320–332, 2017.

[8] M. Brandstein and D. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*, 1st ed. Springer, Jun. 2001. [Online]. Available: http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&amp;path=ASIN/3540419535

[9] C. Evers, H. Loellmann, H. Mellmann, A. Schmidt, H. Barfuss, P. A. Naylor, and W. Kellermann, "The LOCATA challenge: Acoustic source localization and tracking," *IEEE/ACM Trans. ASLP*, pp. 1–1, 2020.

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

[11] A. Barabell, "Improving the resolution performance of eigenstructure-based direction-finding algorithms," in *Proc. IEEE ICASSP*, 1983, pp. 336–339.

[12] B. Friedlander, "The root-MUSIC algorithm for direction finding with interpolated arrays," *Signal Processing*, vol. 30, no. 1, pp. 15–29, 1993.

[13] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 7, pp. 984–995, July 1989.

[14] H. Wang and M. Kaveh, "Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wideband sources," *IEEE Trans. ASSP*, vol. 33, no. 4, pp. 823–831, 1985.

[15] J. P. Dmochowski, J. Benesty, and S. Affes, "Broadband MUSIC: Opportunities and challenges for multiple source localization," in *IEEE WASPAA*, 2007, pp. 18–21.

[16] C. T. Ishi, O. Chatot, H. Ishiguro, and N. Hagita, "Evaluation of a music-based real-time sound localization of multiple sound sources in real noisy environments," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2009, pp. 2027–2032.

[17] S. Argentieri and P. Danes, "Broadband variations of the music high-resolution method for sound source localization in robotics," in *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2007, pp. 2009–2014.

[18] F. Andersson, M. Carlsson, J.-Y. Tourneret, and H. Wendt, "On an iterative method for direction of arrival estimation using multiple frequencies," in *Proc. IEEE CAMSAP*, 2015, pp. 328–331.

[19] H. Pan, R. Scheibler, E. Bezzam, I. Dokmanic, and M. Vetterli, "FRIDA: FRI-based DOA estimation for arbitrary array layouts," in *Proc. ICASSP*, 2017, pp. 3186–3190.

[20] Y. Pan, G. Q. Luo, Z. Liao, B. Cai, and M. Yao, "Wideband direction-of-arrival estimation with arbitrary array via coherent annihilating," *IEEE Access*, 2019.

[21] J. H. Dibiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," *Ph. D. Thesis, AA (Brown University)*, 2000. [Online]. Available: https://ci.nii.ac.jp/naid/10030937868/

[22] H. F. Silverman, Ying Yu, J. M. Sachar, and W. R. Patterson, "Performance of real-time source-location estimators for a large-aperture microphone array," *IEEE Trans. SAP*, vol. 13, no. 4, pp. 593–606, 2005.

[23] M. Cobos, A. Marti, and J. J. Lopez, "A modified SRP-PHAT functional for robust real-time sound source localization with scalable spatial sampling," *IEEE Signal Processing Letters*, vol. 18, no. 1, pp. 71–74, 2011.

[24] S. Rickard and F. Dietrich, "DOA estimation of many W-disjoint orthogonal sources from two mixtures using duet," in *Proceedings of the Tenth IEEE Workshop on Statistical Signal and Array Processing (Cat. No.00TH8496)*, 2000, pp. 311–314.

[25] S. Araki, H. Sawada, R. Mukai, and S. Makino, "DOA estimation for multiple sparse sources with normalized observation vector clustering," in *ICASSP 2006*, vol. 5, May 2006.

[26] M. Togami, T. Sumiyoshi, and A. Amano, "Stepwise phase difference restoration method for sound source localization using multiple microphone pairs," in *ICASSP 2007*, vol. 1, April 2007, pp. I–117–I–120.

[27] W. Zhang and B. D. Rao, "Two microphone based direction of arrival estimation for multiple speech sources using spectral properties of speech," in *ICASSP*, 2009, pp. 2193–2196.

[28] N. T. N. Tho, S. Zhao, and D. L. Jones, "Robust DOA estimation of multiple speech sources," in *ICASSP*, 2014, pp. 2287–2291.

[29] S. Rickard and O. Yilmaz, "On the approximate W-disjoint orthogonality of speech," in *ICASSP 2002*, vol. 1, May 2002, pp. I–529–I–532.

[30] M. Togami, A. Amano, T. Sumiyoshi, and Y. Obuchi, "DOA estimation method based on sparseness of speech sources for human symbiotic robots," in *ICASSP*, 2009, pp. 3693–3696.

[31] K. Yamaoka, R. Scheibler, N. Ono, and Y. Wakabayashi, "Sub-sample time delay estimation via auxiliary-function-based iterative updates," in *WASPAA*, 2019, pp. 130–134.

[32] D. Hunter and K. Lange, "A tutorial on MM algorithms," *The American Statistician*, vol. 58, no. 1, pp. 30–37, 2004. [Online]. Available: https://doi.org/10.1198/0003130042836

[33] J. J. More, "Generalizations of the trust region problem," *Optimization Methods and Software*, vol. 2, no. 3-4, pp. 189–209, 1993. [Online]. Available: https://doi.org/10.1080/10556789308805542

[34] A. Beck, P. Stoica, and J. Li, "Exact and approximate solutions of source localization problems," *IEEE Transactions on Signal Processing*, vol. 56, no. 5, pp. 1770–1778, 2008.

[35] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *ICASSP*, 2018, pp. 351–355.

[36] R. Scheibler, "CMU ARCTIC concatenated 15s," *Zenodo*. [Online]. Available: https://doi.org/10.5281/zenodo.3066489

[37] J. Kominek and A. W. Black, "The CMU ARCTIC speech databases," in *in 5th ISCA Speech Synthesis Workshop*, 2004, pp. 223–224.

[38] R. Scheibler, J. Azcarreta, R. Beuchat, and C. Ferry, "Pyramic: Full stack open microphone array architecture and dataset," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018, pp. 226–230.