INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Sound Source Localization with Majorization Minimization

Masahito Togami and Robin Scheibler

LINE Corporation, Tokyo, Japan

masahito.togami@linecorp.com

Abstract

We propose a sound source localization technique that estimates a speech source location without precise grid searching. The source location is estimated in a parameter optimization manner to minimize the steered-response power (SRP) function with the near-field assumption. Because there is no closed-form solution for the SRP function, we introduce an auxiliary function of the SRP function based on the majorization-minimization (MM) algorithm. Parameters are updated iteratively to minimize the auxiliary function with alternate execution of time-difference-of-arrival (TDOA) estimation and range-difference (RD) based localization. When TDOA estimation and RD-based localization are performed in a cascade manner, the estimation accuracy of the source location is strongly affected by the estimation accuracy of the TDOA. On contrary, the proposed method corrects the estimated TDOA by referring to the estimated source location in the previous iteration. Thus, it is expected for the proposed method to be robust against TDOA estimation error which occurs under reverberant environments. Experimental results show that the proposed method outperforms conventional techniques under a reverberant environment.

Index Terms: TDOA estimation, SRP-PHAT, speech source localization, reverberation

1. Introduction

Sound source localization (SSL) is an essential function for teleconference systems, speaker diarization systems, acoustic event detection systems [1], humanoid robots [2, 3], and so on. As a frontend system of automatic speech recognition systems, SSL enables to identify a target talker. SSL is also useful for speech source separation [4–8].

SSL with multiple microphones has been studied for a long time [9–14]. Steered-response power (SRP) based methods [15–17], multiple signal classification (MUSIC) [18–20], and estimation of signal parameters via rotational invariance techniques (ESPRIT) [21] are frequently utilized. These techniques search for the speech source location by grid searching. SSL accuracy depends on the preciseness of grids. However, precise grid-searching is not preferable from the computational cost perspective, and it is highly required to perform SSL without precise grid-searching.

Cascade connection of time-difference-of-arrival (TDOA) estimation [22] based on the generalized cross-correlation (GCC) methods [23] and range-difference (RD) based localization [24, 25] is an alternative for SSL which does not utilize precise grid-searching. However, the former TDOA estimation typically fails to estimate the TDOA under reverberant environments. When the former TDOA estimation part fails, it is hard for the latter RD based localization part to estimate the correct location of the speech source.

Recently, direction-of-arrival (DOA) estimation methods that do not require precise grid searching have been proposed [26]. Instead of the precise grid searching, the DOA is estimated in a parameter optimization way with the majorization-minimization (MM) algorithm [27]. It is guaranteed that the cost function monotonically decreases by updating an auxiliary function. This MM-based algorithm is based on the far-field assumption that the speech source is far from the microphones. In the SSL problem, it is needed to estimate not only DOA but also the distance of the speech source. Thus, it is needed to derive a MM-based SSL algorithm based on the near-field assumption.

In this paper, we derive a MM-based SSL algorithm based on the near-field assumption. Parameter optimization is done based on the majorization-minimization (MM) algorithm [27] similarly to the sub-sample time delay estimation [28] and the MM-based DOA estimation algorithm [26]. We derive an auxiliary function with the near-field assumption, and we reveal that the proposed optimization algorithm contains a weighted version of the RD based localization step. The proposed method performs the TDOA estimation and the weighted version of the RD based localization alternately. Unlike the cascade-connection based algorithms, the TDOA estimation result is improved based on the estimated speech source location. Experimentally, it is shown that this scheme leads to overcome the shortage of the cascade-connection based algorithm.

2. Problem statement

A multi-channel microphone input signal is modeled in the time-frequency domain as $\boldsymbol{x}_{lk} \in \mathbb{C}^{N_m}$ (l is the frame index, k is the frequency index, N_k is the number of frequency bins, and N_m is the number of the microphones). The amplitude of each element of $\boldsymbol{x}_{lk}$ is normalized to one as $\overline{x}_{lkm} = \frac{x_{lkm}}{|x_{lkm}|}$. Sound source localization is performed with the normalized input signal $\overline{\boldsymbol{x}}_{lk} = \{\overline{x}_{lkm}\}$. Steered-Response Power (SRP) [15–17] estimates a three-dimensional location vector of a speech source $\overline{\boldsymbol{p}}$ as follows:

$$\overline{\boldsymbol{p}} = \arg\max_{\boldsymbol{p}} \left(f(\boldsymbol{p}) = \sum_{lk} \left| \boldsymbol{a}_{\boldsymbol{p}k}^H \overline{\boldsymbol{x}}_{lk} \right|^2 \right), \tag{1}$$

where $\boldsymbol{a}_{\boldsymbol{p}k}$ is the steering vector at the assumed source location $\boldsymbol{p}$ and H is the operator of the Hermitian transpose of a matrix/vector. The function $f(\boldsymbol{p})$ can be expanded as $f(\boldsymbol{p}) = \sum_k \boldsymbol{a}_{\boldsymbol{p}k}^H \boldsymbol{R}_k \boldsymbol{a}_{\boldsymbol{p}k}$, where $\boldsymbol{R}_k = \sum_l \overline{\boldsymbol{x}}_{lk} \overline{\boldsymbol{x}}_{lk}^H$. Each element of the steering vector $\boldsymbol{a}_{\boldsymbol{p}k}$ is modeled under the near-field assumption as $a_{\boldsymbol{p}km} = \exp(-j\omega_k \tau_{\boldsymbol{p}m})$, where $j = \sqrt{-1}$, $\omega_k = 2\pi f_k$, f_k is the frequency (Hz) of the k-th frequency bin, and $\tau_{\boldsymbol{p}m}$ is the time difference between the source location $\boldsymbol{p}$ and the m-th microphone. $\tau_{\boldsymbol{p}m}$ is modeled as $\tau_{\boldsymbol{p}m} = \frac{\|\boldsymbol{p}-\boldsymbol{b}_m\|_2}{c}$, where $\boldsymbol{b}_m$ is the three-dimensional location vector of the m-th microphone and c is the sound speed. In the SRP, grid-searching is required in the $\arg\max$ calculation of (1). SSL accuracy depends on the preciseness of grids. However, it is needed to remove precise grid-searching from the computational-cost perspective.

http://dx.doi.org/10.21437/Interspeech.2021-126

3. Proposed method

3.1. Expansion of objective function

The proposed method removes precise grid-searching for SSL. Instead, the proposed method estimates the speech source location in a parameter optimization manner. We expand the objective function $f(\boldsymbol{p})$ to formulate the SSL problem as a parameter optimization problem. Each element in the matrix $\boldsymbol{R}_k$ is modeled as $r_{kmn}\exp(j\sigma_{kmn})$. The objective function $f(\boldsymbol{p})$ is $f(\boldsymbol{p}) = \sum_{k=1}^{N_k}\sum_{m=1}^{N_m}\sum_{n=1}^{N_m} r_{kmn}\exp(j\sigma_{kmn} + j\omega_k\Delta\tau_{\boldsymbol{p}mn})$, where $\Delta\tau_{\boldsymbol{p}mn} = \tau_{\boldsymbol{p}m} - \tau_{\boldsymbol{p}n}$. $f(\boldsymbol{p})$ can be further expanded by using $\sigma_{kmn} = -\sigma_{knm}$ as follows:

$$f(\boldsymbol{p}) = \sum_{k=1}^{N_k}\sum_{m=1}^{N_m}\sum_{n=1}^{N_m} r_{kmn}\cos(\sigma_{kmn} + \omega_k\Delta\tau_{\boldsymbol{p}mn}). \quad (2)$$

Because $\boldsymbol{p}$ is included in the cos function, optimization of $f(\boldsymbol{p})$ w.r.t. $\boldsymbol{p}$ is difficult.

3.2. Introduction of Majorization-Minimization algorithm

We introduce the Majorization-Minimization (MM) algorithm [27] into the parameter optimization. Similarly to [26, 28], an upper-limit of the negative objective function can be derived as $-f(\boldsymbol{p}) \leq g(\boldsymbol{p}, z, y)$, where

$$g(\boldsymbol{p}, z, y) = \sum_{k=1}^{N_k}\sum_{m=1}^{N_m}\sum_{n=1}^{N_m} \alpha_{kmn}(\Delta\tau_{\boldsymbol{p}mn} - d_{kmn})^2 + const., \quad (3)$$

$\alpha_{kmn} = \frac{\omega_k^2 r_{kmn}\sin(y_{kmn})}{2y_{kmn}}$, and $d_{kmn} = -\frac{\sigma_{kmn} + 2\pi z_{kmn}}{\omega_k}$. d_{kmn} is corresponding to the TDOA between the m-th microphone and the n-th microphone in the k-th frequency bin. Instead of optimizing the original objective function $f(\boldsymbol{p})$, we optimize the auxiliary function $g(\boldsymbol{p}, z, y)$ w.r.t. $\boldsymbol{p}$, and the auxiliary variables z and y iteratively. Thus, the update of an auxiliary variable z is corresponding to restoration of the TDOA estimation d by referring to the estimated source location $\boldsymbol{p}$.

3.3. Parameter optimization

The proposed method updates auxiliary variables and the speech source location iteratively.

3.3.1. Update of auxiliary variables

Similarly to [26, 28], the auxiliary variables are updated as follows:

$$z_{kmn} = \underset{z\in\mathbb{Z}}{\arg\min}\,|\sigma_{kmn} + \omega_k\Delta\tau_{\boldsymbol{p},mn} + 2\pi z|, \quad (4)$$

$$y_{kmn} = \sigma_{kmn} + \omega_k\Delta\tau_{\boldsymbol{p},mn} + 2\pi z_{kmn}. \quad (5)$$

3.3.2. Parameter update

Let the first term of (3) be $h(\boldsymbol{p}, z, y)$. $h(\boldsymbol{p}, z, y)$ is redefined as follows:

$$h(\boldsymbol{p}, z, y) = \sum_{k=1}^{N_k}\sum_{m=1}^{N_m}\sum_{n=1}^{N_m} \alpha_{kmn}(\tau_{\boldsymbol{p}m} - \tau_{\boldsymbol{p}n} - d_{kmn})^2. \quad (6)$$

Optimization of $h(\boldsymbol{p}, z, y)$ w.r.t. $\boldsymbol{p}$ corresponds to a weighted version of the RD based localization problem [24, 25]. In [25], the RD based localization problem is solved by the r-means based on the MM algorithm. As an extension of the r-means algorithm, we derive the weighted r-means algorithm (wr-means) to optimize $h(\boldsymbol{p}, z, y)$.

3.3.3. Weighted r-means algorithm

We derive two types of auxiliary functions for $h(\boldsymbol{p}, z, y)$. The following lemma is used to derive the first auxiliary function.

Lemma 1. *When the matrix $\boldsymbol{A}_k = \{\alpha_{kmn}\}$ is a semi-positive definite matrix,*

$$h(\boldsymbol{p}, z, y) \leq 2\sum_{k=1}^{N_k}\sum_{m=1}^{N_m} \overline{\alpha}_{km}(\tau_{\boldsymbol{p}m} - s_{1km})^2 + const., \quad (7)$$

where $s_{1km} = \frac{\sum_{n=1}^{N_m}\alpha_{kmn}(\tau_{0,n} + d_{kmn})}{\overline{\alpha}_{km}}$, $\overline{\alpha}_{km} = \sum_{n=1}^{N_m}\alpha_{kmn}$, *and* τ_{0n} *is an auxiliary variable.*

Proof. $h(\boldsymbol{p}, z, y)$ is expanded as follows:

$$\begin{aligned} h(\boldsymbol{p}, z, y) = &\sum_{k=1}^{N_k}\sum_{m=1}^{N_m}\sum_{n=1}^{N_m} 2\alpha_{kmn}\left(\tau_{\boldsymbol{p}m}^2 + d_{kmn}^2 - 4d_{kmn}\tau_{\boldsymbol{p}m}\right) \\ &- 2\sum_{k=1}^{N_k}\sum_{m=1}^{N_m}\sum_{n=1}^{N_m} \alpha_{kmn}\tau_{\boldsymbol{p}m}\tau_{\boldsymbol{p}n}. \end{aligned} \quad (8)$$

When $\boldsymbol{A}_k$ is a semi-positive definite matrix, an upper limit of the last term can be derived as follows:

$$\begin{aligned} &\sum_{k=1}^{N_k}\sum_{m=1}^{N_m}\sum_{n=1}^{N_m} \alpha_{kmn}(\tau_{\boldsymbol{p},m} - \tau_{0,m})(\tau_{\boldsymbol{p},n} - \tau_{0,n}) \geq 0 \\ &\Leftrightarrow -2\sum_{k=1}^{N_k}\sum_{m=1}^{N_m}\sum_{n=1}^{N_m} \alpha_{kmn}\tau_{\boldsymbol{p}m}\tau_{\boldsymbol{p}n} \\ &\leq -4\sum_{k=1}^{N_k}\sum_{m=1}^{N_m}\sum_{n=1}^{N_m} \alpha_{kmn}\tau_{0n}\tau_{\boldsymbol{p}m} + 2\sum_{m=1}^{N_m}\sum_{n=1}^{N_m} \alpha_{kmn}\tau_{0m}\tau_{0n}. \end{aligned} \quad (9)$$

We can derive (7) by applying (9) to (8). □

When $\boldsymbol{A}_k$ is not a semi-positive definite matrix, we can utilize $\overline{\boldsymbol{A}}_k = \boldsymbol{A}_k - \lambda_k\boldsymbol{I}$ instead of the original $\boldsymbol{A}_k$, where λ_k is the minimum eigenvalue of the matrix $\boldsymbol{A}_k$. An important point is that even when we utilize $\overline{\boldsymbol{A}}_k$, $g(\boldsymbol{p}, z, y)$ does not change. However, in our experiment, there are few cases that do not fulfill (7). We define $r_1(\boldsymbol{p})$ as $\sum_k\sum_{m=1}^{M}\overline{\alpha}_{km}(\tau_{\boldsymbol{p},m} - s_{1km})^2$. $\boldsymbol{p}$ can be updated to minimize $r_1(\boldsymbol{p})$. There is another upper-limit of $h(\boldsymbol{p}, z, y)$.

$$\begin{aligned} h(\boldsymbol{p}, z, y) &\leq \sum_{k=1}^{N_k}\sum_{m=1}^{N_m}\sum_{n=1}^{N_m} \alpha_{kmn}(2\tau_{\boldsymbol{p},m} - \tau_{0m} - \tau_{0n} - d_{kmn})^2 \\ &= 4\sum_{k=1}^{N_k}\sum_{m=1}^{N_m} \overline{\alpha}_{km}(\tau_{\boldsymbol{p}m} - s_{2km})^2, \end{aligned} \quad (10)$$

where $s_{2km} = \frac{\sum_n \alpha_{kmn}(\tau_{0m} + \tau_{0n} + d_{kmn})}{2\overline{\alpha}_{km}}$. We define $r_2(\boldsymbol{p})$ as $\sum_{k=1}^{N_k}\sum_{m=1}^{N_m}\overline{\alpha}_{km}(\tau_{\boldsymbol{p},m} - s_{2km})^2$.

To optimize $r_1(\boldsymbol{p})$ or $r_2(\boldsymbol{p})$, we derive an additional auxiliary function. In [25], the following lemma is shown.

Lemma 2. *When s_{km} is a positive scalar,*

$$\sum_{k=1}^{N_k}\sum_{m=1}^{N_m}\overline{\alpha}_{km}\left(\tau_{\boldsymbol{p},m}-s_{km}\right)^2 \leq \sum_{k=1}^{N_k}\sum_{m=1}^{N_m}\overline{\alpha}_{km}\|\boldsymbol{p}-\boldsymbol{a}_m-s_{km}\boldsymbol{e}_m\|_2^2, \tag{11}$$

where

$$\boldsymbol{e}_m=\frac{\boldsymbol{p}-\boldsymbol{a}_m}{\|\boldsymbol{p}-\boldsymbol{a}_m\|_2}. \tag{12}$$

Similarly to [25], s_{1km} and s_{2km} are almost positive scalars. Thus, $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ are optimized by minimizing the right term of the (11) as follows:

$$\boldsymbol{p}_1=\frac{\sum_k\sum_{m=1}^M\sum_{n=1}^M\alpha_{kmn}\left(\boldsymbol{a}_m+\left(\tau_{0n}+d_{kmn}\right)\boldsymbol{e}_m\right)}{\sum_k\sum_{m=1}^M\sum_{n=1}^M\alpha_{kmn}}, \tag{13}$$

$$\boldsymbol{p}_2=\frac{\sum_k\sum_{m=1}^M\sum_{n=1}^M\alpha_{kmn}\left(\boldsymbol{a}_m+\frac{\tau_{0,m}+\tau_{0n}+d_{kmn}}{2}\boldsymbol{e}_m\right)}{\sum_k\sum_{m=1}^M\sum_{n=1}^M\alpha_{kmn}}. \tag{14}$$

Generally speaking, the cost function decreases more rapidly with $\boldsymbol{p}_1$ than with $\boldsymbol{p}_2$. However, there is no justification that $\boldsymbol{A}$ is a positive definite matrix. Therefore, we utilize $\boldsymbol{p}\in\{\boldsymbol{p}_1,\boldsymbol{p}_2\}$ which minimizes the original cost function. Similarly to the r-means [25], we also utilize the Steffensen acceleration method [29] to accelerate convergence speed. t is the iteration index.

$$\boldsymbol{q}^{(t)}=\boldsymbol{p}^{(t)}+\alpha\Delta\boldsymbol{p}^{(t)}, \tag{15}$$

where

$$\Delta\boldsymbol{p}^{(t)}=\boldsymbol{p}^{(t)}-\boldsymbol{p}^{(t-1)}, \tag{16}$$

$$\Delta^2\boldsymbol{p}^{(t)}=\Delta\boldsymbol{p}^{(t)}-\Delta\boldsymbol{p}^{(t-1)}, \tag{17}$$

$$\alpha=-\frac{\Delta\boldsymbol{p}^{(t),T}\Delta^2\boldsymbol{p}^{(t)}}{\|\Delta^2\boldsymbol{p}^{(t)}\|^2}. \tag{18}$$

Finally,

$$\tilde{\boldsymbol{p}}^{(t)}=\arg\max\left(f\left(\boldsymbol{p}^{(t)}\right),f\left(\boldsymbol{q}^{(t)}\right)\right). \tag{19}$$

3.3.4. Summary of parameter update

In each iteration, auxiliary variables and the speech source location are updated as follows:

1. Update z_{kmn} based on (4).
2. Update y_{kmn} based on (5).
3. Update $\tau_{0m}=\tau_{\boldsymbol{p}m}$.
4. Update $\boldsymbol{e}_m$ based on (12).
5. Update $\boldsymbol{p}_1$, $\boldsymbol{p}_2$, $\boldsymbol{q}^{(t)}$, and $\tilde{\boldsymbol{p}}^{(t)}$ based on (13), (14), (15), and (19), respectively.

$\tilde{\boldsymbol{p}}^{(L_t)}$($L_t$ is the final-iteration index) is the output source location.

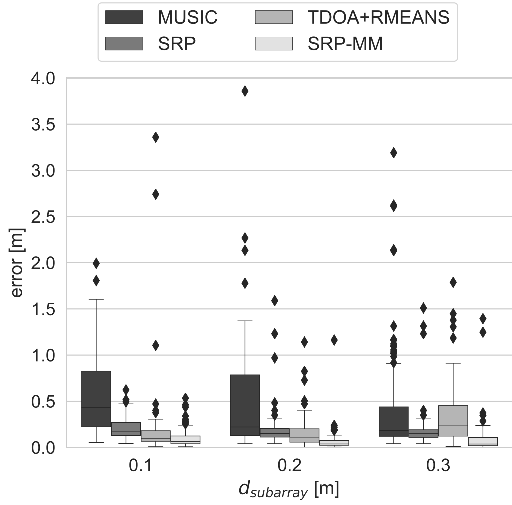

Figure 1: *Experimental results when $T_{60}=0.33$ [s]*

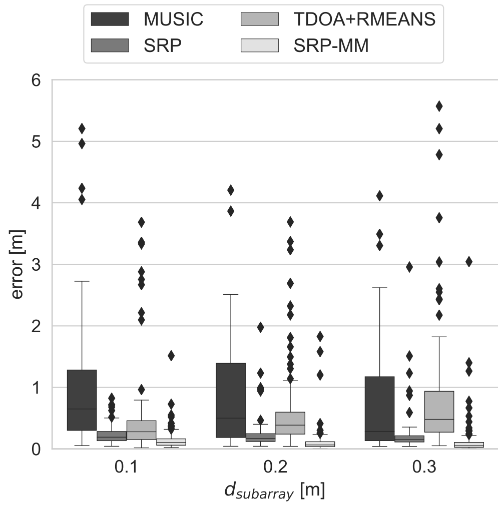

Figure 2: *Experimental results when $T_{60}=0.73$ [s]*

4. Experiment

4.1. Experimental setup

Sound source localization performance was evaluated by using `Pyroomacoustics` [30]. `Pyroomacoustics` simulates reverberant mixtures in a $5\times5\times3$ m room with two RT_{60} conditions, i.e., 0.33 [s] and 0.73 [s]. Anechoic speech sources were extracted from the CMU ARCTIC Concat15 dataset [31] which concatenates utterances extracted from the CMU Sphinx database [32]. The average speech length was approximately 17 [s]. The number of the speech sources was set to 1. 100 reverberant mixtures were simulated for each condition. The sampling rate was 16000 Hz. Signal to Noise Ratio (SNR) between a speech source and background noise was set to 10 dB. The frame size was 2048. The frame shift was 512. The total number of frequency bins was 1025. All frequency bins were utilized for localization. We utilized 5 regular tetrahedron arrays. The total number of microphones was 20. Each array was regarded as a subarray. Locations of 5 arrays were $(2.5,2.5,1.5)$, $(0.25,2.5,1.5)$, $(4.75,2.5,1.5)$, $(2.5,0.25,1.5)$, and $(2.5,4.75,1.5)$. We call

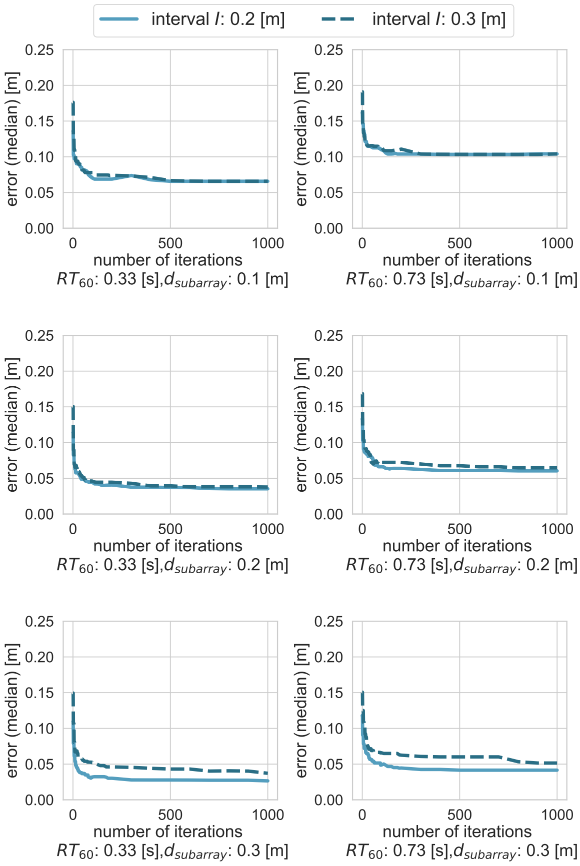

Figure 3: *Experimental results of convergence speed*

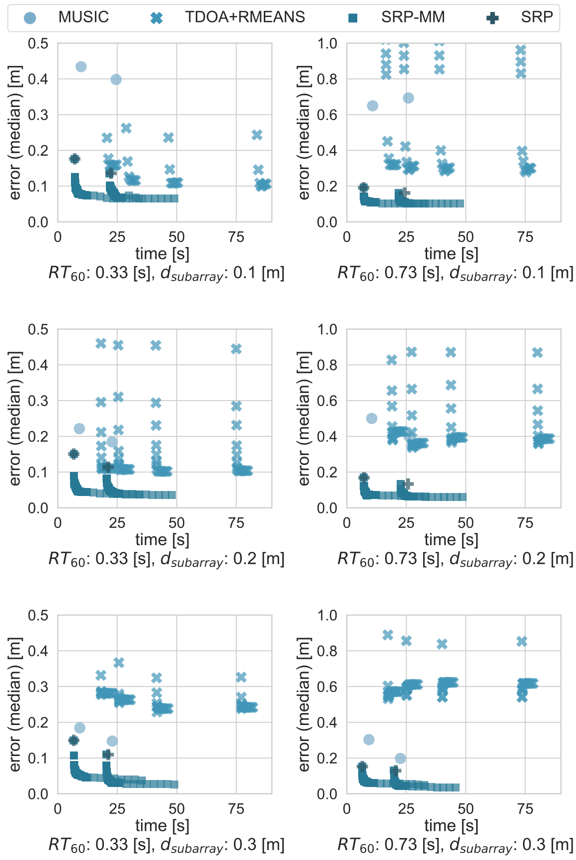

Figure 4: *Experimental results of localization performance against processing time*

the proposed method SRP-MM. The proposed SRP-MM was compared with SRP [15], MUSIC [18], and TDOA + RMEANS in which TDOA estimation is done with GCC-PHAT [23] followed by SSL with r-means [25]. We calculate the correlation between microphones only within each subarray because SSL results of all methods were better with this approximation in the preliminary experiment. This means that the covariance matrix $\boldsymbol{R}_k$ was approximated as $\boldsymbol{R}_k \approx \mathrm{diag}\{\ \boldsymbol{R}_{1k}\ \ \cdots\ \ \boldsymbol{R}_{Bk}\ \}$. In the proposed method, the initial location $\boldsymbol{p}$ was set to be a SSL result by SRP with the rough grid searching under the condition that the interval between two grids was I.

4.2. Experimental results

In Fig. 1 and Fig. 2, experimental results for $RT_{60} = 0.33$ [s] and 0.73 [s] are shown as box-whisker plots, respectively. The distance between microphones in each subarray, d_{subarray}, was set to 0.1, 0.2, and 0.3 [m]. In TDOA+RMEANS, the interpolation rate was set to 8. In the proposed SRP-MM, the interval between two grids I was set to 0.3m each. The number of the MM iterations for SRP-MM was 1000, and the number of the MM iterations for TDOA+RMEANS was 5000. In the preliminary experiment, we confirmed that SRP-MM and TDOA+RMEANS converge in these iterations. When the environment is more reverberant, i.e., $RT_{60} = 0.73$ [s], the estimation error of the TDOA-RMEANS is much bigger than the less reverberant case ($RT_{60} = 0.33$ [s]). On the other hand, the estimation error of the SRP-MM is much less than that of the TDOA+RMEANS. It is shown that the proposed alternate update of the parameters is more robust against reverberation than the cascade-connection based method. In Fig. 3, convergence speed was evaluated. I was set to be 0.2 [m] or 0.3[m]. It is shown that the median error of the SRP-MM is decreasing stably in each result. We also evaluated localization performance against processing time in Fig. 4. In the proposed method, I was set to be 0.2 [m] or 0.3 [m]. A server with Intel Xeon Silver 4210 CPU @ 2.20GHz and 32 GB RAM was used. Each dot represents the localization error for each iteration. In TDOA+RMEANS, the interpolation ratio was set to be 1, 2, 4, and 8. It is shown that the proposed SRP-MM outperformed the other methods under the condition that the processing time was equivalent under a reverberant environment.

4.3. Conclusion

We proposed a sound source localization method that does not utilize precise grid searching. The proposed method estimates a speech source location in a parameter optimization manner based on the majorization-minimization (MM) algorithm. Experimental results showed that the proposed method outperformed cascade connection of TDOA estimation and range difference based localization, especially under reverberant environments.

5. References

[1] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, “Sound event localization and detection of overlapping sources using convolutional recurrent neural networks,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48, 2019.

[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] S. Argentieri, P. Danès, and P. Souères, "A survey on sound source localization in robotics: from binaural to array processing methods," *Computer Speech and Language, Elsevier*, vol. 34, no. 1, pp. 87–112, 2015.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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&path=ASIN/3540419535

[10] A. Brutti, M. Omologo, and P. Svaizer, "Comparison between different sound source localization techniques based on a real data collection," in *2008 Hands-Free Speech Communication and Microphone Arrays*, 2008, pp. 69–72.

[11] A. Brutt, M. Omologo, and P. Svaizer, "Multiple source localization based on acoustic map de-emphasis," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, pp. 1–17, 2010.

[12] F. Keyrouz, "Binaural range estimation using head related transfer functions," in *2015 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, 2015, pp. 89–94.

[13] A. Brendel and W. Kellermann, "Distance estimation of acoustic sources using the coherent-to-diffuse power ratio based on distributed training," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018, pp. 1–5.

[14] 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.

[15] 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/

[16] 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.

[17] 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.

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

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

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

[21] 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.

[22] Y. Huang, J. Benesty, G. Elko, and R. Mersereati, "Real-time passive source localization: a practical linear-correction least-squares approach," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 8, pp. 943–956, 2001.

[23] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, Aug 1976.

[24] 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.

[25] N. Ono and S. Sagayama, "R-means localization: A simple iterative algorithm for range-difference-based source localization," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 2718–2721.

[26] M. Togami and R. Scheibler, "Sparseness-aware DOA estimation with majorization minimization," in *Proc. Interspeech*, Aug. 2020 (accepted).

[27] 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

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

[29] C. Small and J. Wang, "Numerical methods for nonlinear estimating equations," *Numerical Methods for Nonlinear Estimating Equations*, 01 2003.

[30] 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.

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

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