Multiple Sound Source Tracking via An Avoidance of Spatial Aliasing

Katsuya Nakazawa and Kenji Suyama
School of Engineering, Tokyo Denki University,
5 Senju Asahi-cho, Adachi-ku, Tokyo 120–8551, Japan

Abstract: In this paper, a method of multiple sound source tracking using microphone array is studied. MUltiple SIgnal Classification (MUSIC) is known as a high spatial resolution method. However, it requires high computation cost and thus a real time processing is prevented. For reducing the cost, Projection Approximation Subspace Tracking Interior Point Least Square (PAST-IPLS) based on MUSIC has been proposed. Although an evaluation function of IPLS tends to indicate the high spatial resolution by expanding a microphone width, several local minimums tend to appear. Therefore, we propose an avoidance method of spatial aliasing by using a multiplication of evaluation functions of whole frequency bands. The effectiveness of the proposed method is shown by several experimental results.

1. Introduction

Sound source tracking using microphone array is an important technique for various acoustic interfaces such as a robot hearing and a teleconferencing. Various methods for sound source tracking have been studied [1]-[3]. MUSIC is one of the high spatial resolution methods. However, high computation cost is required. PAST-IPLS [6] is a method for reducing the computation cost based on two successive algorithm, i.e., PAST and IPLS. In general, widening a microphone width, an estimation accuracy can be improved by the high spatial resolution. However, in the frequencies that a spatial sampling theorem is not kept, failure peaks may appear in the evaluation function by a spatial aliasing.

In the proposed method, the spatial aliasing is avoided by a multiplication of evaluation functions of whole frequency bands. In addition, because the same source estimation problem is involved in multiple source tracking methods, a method for resolving the problem is also studied. Several experimental results are shown to evaluate the effectiveness of the proposed method.

2. Problem formulation

A model of multiple sound source tracking is shown in Figure 1, where n is discrete time, $\theta_i(n)$, $i = 1, 2$ is the i-th source direction, $x_m(n)$, $m = 1, \cdots, M$ is the signal received by the m-th microphone. Two sound sources, $s_i(n)$, move with time. They are received by multiple microphones arranged linearly. $x_m(n)$ can be written in a frequency domain as below,

$$X_m(t,k) = \sum_{i=1}^{2} S_i(t,k) e^{-j\omega_k (m-1)\tau_i(t)} + \Gamma_m(t,k), \quad (1)$$

$$\tau_i(\theta(t)) = d \sin(\theta(t))/c, \quad (2)$$

where t is a frame index, k is a frequency index, $\tau_i(\theta(t))$ is the time difference of arrival of $s_i(n)$, d is a microphone width,

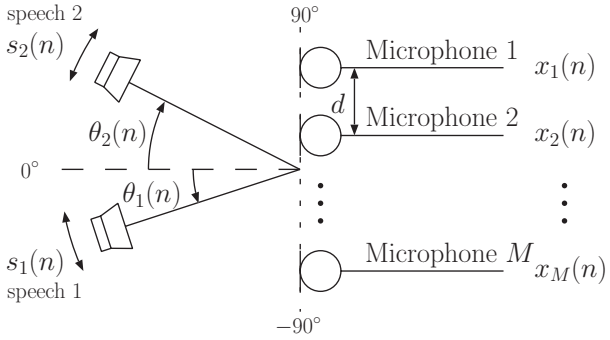


Figure 1. A model of sound source tracking

c is the velocity of sound, $S_i(t,k)$ is a complex amplitude of $s_i(n)$, and $\Gamma_m(t,k)$ is the observed noise at the m-th microphone. In our system, it is assumed that d is set to 0.16[m]. (2) can be rewritten in a vector domain notation as below,

$$\boldsymbol{X}(t,k) = \sum_{i=1}^{2} S_i(t,k)\boldsymbol{a}_k(\tau_i(t)) + \boldsymbol{\Gamma}(t,k), \quad (3)$$

$$= [X_1(t,k), \cdots, X_M(t,k)]^T, \quad (4)$$

where $\boldsymbol{a}_k(\tau_i(t))$ is a steering vector, $\boldsymbol{\Gamma}(t,k)$ is the observed noise vector, T denotes the transposition. They are defined by following,

$$\boldsymbol{a}_k(\tau_i(t)) = \left[1, \cdots, e^{-j\omega_k M \tau_i(t)}\right]^T, \quad (5)$$

$$\boldsymbol{\Gamma}(t,k) = [\Gamma_1(t,k), \cdots, \Gamma_M(t,k)]^T. \quad (6)$$

The aim of multiple sound source tracking is to estimate $\tau_i(t)$ from $X_m(t,k)$ every t.

3. Proposed method

A procedure of the proposed method is shown in Figure 2. and described as below.

1. $\boldsymbol{x}(n)$ is transformed into the frequency domain by Discrete Fourier Transform (DFT).
2. $\boldsymbol{Q}_S(t,k)$ is updated by the PAST. The updating algorithm follows [6].
3. $J_{\mathrm{I},k}(\tau)$ is evaluated every k.
4. $J_{\mathrm{I}}(\tau)$ is evaluated from the multiplication of $J_{\mathrm{I},k}(\tau)$.
5. $\tau_i(t)$ is estimated by IPLS at t.
6. When an active sound source number has increased, feasible regions $\Omega_{i,t}$ are reset.

3.1 MUSIC

PAST-IPLS is the sound source tracking method based on MUSIC. MUSIC spectrum is defined as below,

$$P_{\mathrm{MU}}(k,\tau) = \frac{\boldsymbol{a}_k^H(\tau)\boldsymbol{a}_k(\tau)}{\boldsymbol{a}_k^H(\tau)\boldsymbol{Q}_N(t,k)\boldsymbol{Q}_N^H(t,k)\boldsymbol{a}_k(\tau)}, \quad (7)$$

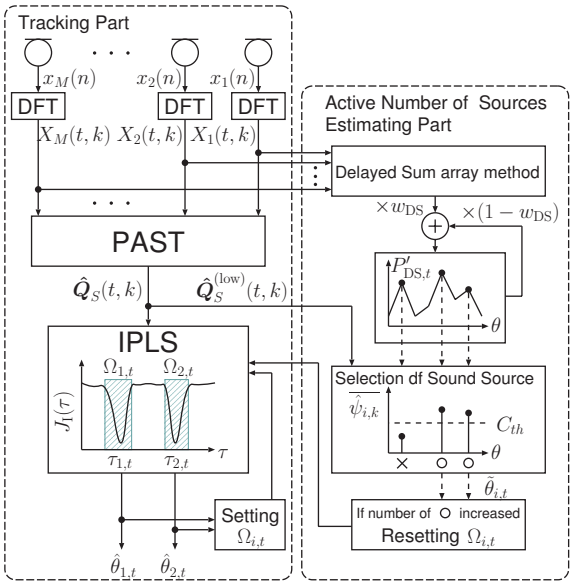


Figure 2. A procedure of the proposed method

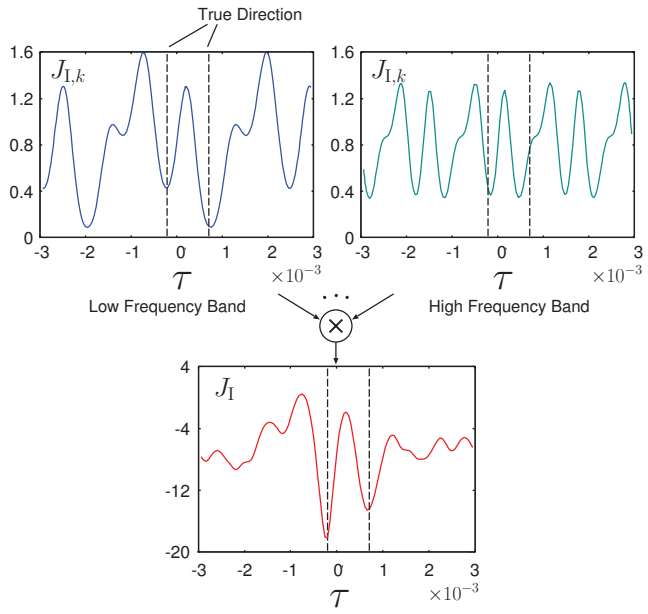


Figure 3. Peak enhancement by multiplied function

where $\boldsymbol{Q}_N(t,k)\boldsymbol{Q}_N^H(t,k)=\boldsymbol{I}-\boldsymbol{Q}_S(t,k)\boldsymbol{Q}_S^H(t,k)$, $\boldsymbol{Q}_N(t,k)$ is the noise subspace, $\boldsymbol{Q}_S(t,k)$ is the signal subspace, $\boldsymbol{I}$ is the identity matrix. $P_{\mathrm{MU}}(k,\tau)$ has peaks corresponding to the sound directions. Although the spectrum indicates high spatial resolution, high computation cost is required for following two procedures. One is the eigenvalue decomposition for calculating $\boldsymbol{Q}_N(t,k)$, the other is the peak search of $P_{\mathrm{MU}}(k,\tau)$. In the proposed method, PAST is used for the successive eigenvalue decomposition and IPLS is used for the peak search.

3.2 IPLS

An evaluation function of IPLS, $J_{\mathrm{I},k}(\tau)$, is the denominator of (7) and is represented as.

$$J_{\mathrm{I},k}(\tau)=\boldsymbol{a}^H(k,\tau)\boldsymbol{Q}_N(t,k)\boldsymbol{Q}_N^H(t,k)\boldsymbol{a}(k,\tau). \tag{8}$$

The sound source direction can be estimated by searching τ which minimizes $J_{\mathrm{I},k}(\tau)$. Because $J_{\mathrm{I},k}(\tau)$ is a multi-modal function, the number of peaks increases in the high frequency band. Moreover, false peaks by the spatial aliasing tend to appear in the frequencies where the spatial aliasing theorem can not be kept. It should be noted that those directions are different every frequency. Therefore, in the proposed method, multiplied evaluation function $J_{\mathrm{I}}(\tau)$ is used to enhance the true peaks while suppressing the effect of the spatial aliasing (Figure 3). $J_{\mathrm{I}}(\tau)$ is evaluated as,

$$J_{\mathrm{I}}(\tau)=\prod_k J_{\mathrm{I},k}(\tau) \tag{9}$$

For restricting searching areas, the feasible regions $\Omega_{i,t}$ are set to the evaluation function and a log barrier function $\phi_{i,t}(\tau)$ is defined to guarantee the existence of the analytic center. $\phi_{i,t}(\tau)$ is represented by following equation.

$$\phi_{i,t}(\tau)=-\log(\zeta_{i,t}-J_{\mathrm{I}}(\tau))-\log(\eta^2-\tau^2), \tag{10}$$

where $\zeta_{i,t}$ is the upper limit of $J_{\mathrm{I}}(\tau)$, η is the range which $|\tau|$ is available, $\eta=d/c$. In the proposed method, $\hat{\tau}_{i,t}$ is updated by following Newton single updating,

$$\hat{\tau}_{i,t}=\hat{\tau}_{i,t-1}-\frac{\nabla\phi_{i,t}(\hat{\tau}_{i,t-1})}{\nabla^2\phi_{i,t}(\hat{\tau}_{i,t-1})}. \tag{11}$$

$\nabla\phi_{i,t}(\tau)$, $\nabla^2\phi_{i,t}(\tau)$ is updated as,

$$\nabla\phi_{i,t}(\hat{\tau}_{i,t-1}) = \frac{\nabla\phi_{i,t}(\hat{\tau}_{i,t-1})}{u}+\frac{2\hat{\tau}_{i,t-1}}{v}, \tag{12}$$

$$\nabla^2\phi_{i,t}(\hat{\tau}_{i,t-1}) = \frac{(\nabla J_{\mathrm{I}}(\hat{\tau}_{i,t-1}))^2}{u^2}+\frac{\nabla^2 J_{\mathrm{I}}(\hat{\tau}_{i,t-1})}{u} +\frac{4\hat{\tau}_{i,t-1}}{v^2}+\frac{2}{v}. \tag{13}$$

u, v is defined as,

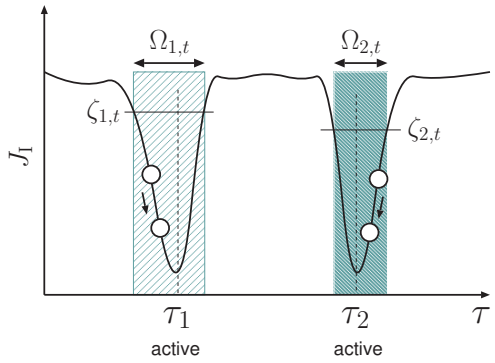
$$u = \zeta_{i,t}-J_{\mathrm{I}}(\hat{\tau}_{i,t-1}), \tag{14}$$

$$= \mu\frac{\eta}{\sqrt{2}}|\nabla J_{\mathrm{I}}(\hat{\tau}_{i,t-1})|$$

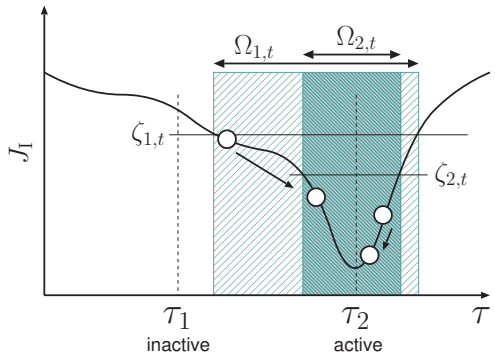
$$v = \eta^2-|\hat{\tau}_{i,t-1}|^2. \tag{15}$$

3.3 Resetting of feasible regions

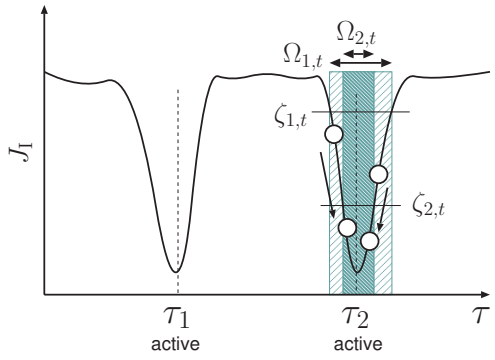
In the multiple sound source tracking, the same source estimation often occurs. The evaluation function $J_{\mathrm{I}}(\tau)$ has peaks corresponding to each sound source when two talkers are active simultaneously (Figure 4(a)). However, if one talker becomes inactive, $J_{\mathrm{I}}(\tau)$ has one peak corresponding to only the active sound source and estimation results converges to the same direction (Figure 4(b)). After that, the talker returns active and the peak corresponding to the talker direction appears again, however estimation results remain around the other talker direction because the feasible regions are not appropriate (Figure 4(c)). For avoiding this problem, feasible regions



(a) when two sources are active



(b) when one source is inactive



(c) when two sources return to be active

Figure 4. Same source estimation problem

should be needed to be reset. In the proposed method, feasible regions are set based on the delayed-sum-array method and a sound source selection.

The delayed-sum-array method calculates the spatial power spectrum. The summed spectrum is calculated as,

$$P_{\mathrm{DS}}(\theta) = \sum_k \hat{P}_{\mathrm{DS}}(k,\theta), \tag{16}$$

$$\hat{P}_{\mathrm{DS}}(k,\theta) = \left|\boldsymbol{a}_k^H(\theta)\boldsymbol{X}_t(k)\right|^2. \tag{17}$$

Moreover, the sound signal is generally non-stationary. Therefore, a sequential updating of $P_{\mathrm{DS}}(\theta)$ is applied as following,

$$P'_{\mathrm{DS},t}(\theta) = w_{\mathrm{DS}} P_{\mathrm{DS},t}(\theta) + (1 - w_{\mathrm{DS}}) P'_{\mathrm{DS},t-1}(\theta), \tag{18}$$

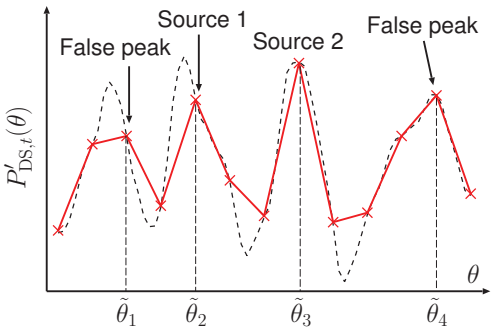


Figure 5. The rough peak search of summed spectrum

where, w_{DS}, $0 < w_{\mathrm{DS}} \leq 1$ is an updating weight. For reducing the computation cost, $P'_{\mathrm{DS},t}(\theta)$ is searched roughly. Then, the peak is picked as candidates of the sound direction.

Those candidates involves false peaks due to the spatial aliasing. Then, a sound source selection is implemented. In the sound source selection, the inner product

$$\hat{\psi}_k(\tilde{\theta}_i(t)) = \langle \hat{\boldsymbol{Q}}_S^{(\mathrm{low})}(k,t), \boldsymbol{a}_k(\theta_i(t)) \rangle \tag{19}$$

is calculated in the low frequency band where the spatial aliasing does not appear. An average of the inner product $\psi_k(\tilde{\theta}_i(t))$ is evaluated. If $\psi_k(\tilde{\theta}_i(t)) > C_{th}$, it is considered that $\theta_i(t)$ is the active source direction, where C_{th} is a threshold value specified in advance. If multiple $\psi_k(\tilde{\theta}_i(t))$ exceed C_{th}, the higher two $\psi_k(\tilde{\theta}_i(t))$ are adopted. Then, the feasible regions are reset according to the number of active sound sources $\hat{L}_t$.

1. $\hat{L}_{t-1} = 0, 1$ and $\hat{L}_t = 2$
   Resetting of the feasible regions are implemented.
2. The other
   Resetting of the feasible regions are not implemented.

4. Experiments

Several experiments were conducted in an actual room to evaluate the effectiveness of the proposed method. The room size was $9.0[\mathrm{m}] \times 26.0[\mathrm{m}] \times 2.5[\mathrm{m}]$, the reverberation time was $0.51[\mathrm{ms}]$ and the noise level was $38.4[\mathrm{dB}]$. The signal length was $20[\mathrm{s}]$ and two speech signals were uttered from the moving speakers. Experiments were conducted for 4 source signal patterns. Table 1 shows the parameters used in the experiments. Value of η, μ, w_{DS}, C_{th} were decided from preliminary experimental results. Table 2 shows the frequency bands used in the proposed method.

The accuracy of tracking was measured by the Root Mean Square Error (RMSE) as,

$$\mathrm{RMSE} = \sqrt{\frac{1}{2}\sum_{i=1}^{2} |\theta_{i,t} - \hat{\theta}_{i,t}|^2}, \tag{20}$$

where $\theta_{i,t}$ is a true direction, $\hat{\theta}_{i,t}$ is an estimation result, i is the source index, $\overline{\cdot}$ means averaging over time. The average RMSE was 3.590. The processing speed was evaluated by

Table 1. Experimental conditions

Sampling frequency f_s [kHz]	8
Frame length N	512
Microphone width d [m]	0.16
Width of peak search [°]	6
Forgetting factor β	0.2
Step size parameter μ	0.01
Threshold value C_{th}	0.61
Updating weight w_{DS}	0.3

Table 2. Frequency bands used in the proposed method

PAST-IPLS	$1 \sim 4$ [kHz]
Delayed-sum-array method	$3 \sim 4$ [kHz]
Peak selection operation	$1 \sim 2$ [kHz]

Table 3. Results of RMSE

Pattern	1	2	3	4
RMSE [°]	4.169	2.382	2.891	4.919

Real Time Factor (RTF) as,

$$\mathrm{RTF} = \frac{\text{computational time}}{\text{signal length}}. \tag{21}$$

Table 3 shows the RMSE every pattern. From Table 3, the proposed method succeeded in the multiple sound source tracking with a high accuracy. The RTF average was 0.059. It is shown that the proposed method realized real time processing.

Figure 6 shows an example of the failed tracking results of pattern 3, where resetting of feasible regions is not applied. It is confirmed that the false peaks are appeared in the evaluation function because of spatial aliasing. Figure 7 and Figure 8 show the tracking results of the proposed method. From Figure 7 and Figure 8, it is shown that the same source estimation could be avoided. and the estimation result was reset to the true direction.

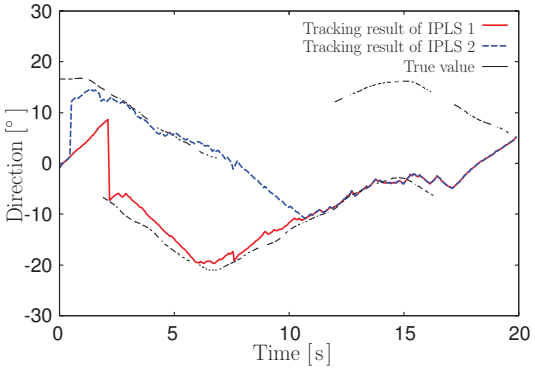


Figure 6. Failed tracking results of pattern 3

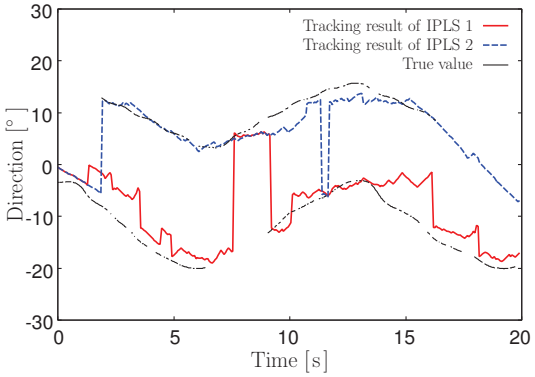


Figure 7. Tracking results of pattern 2

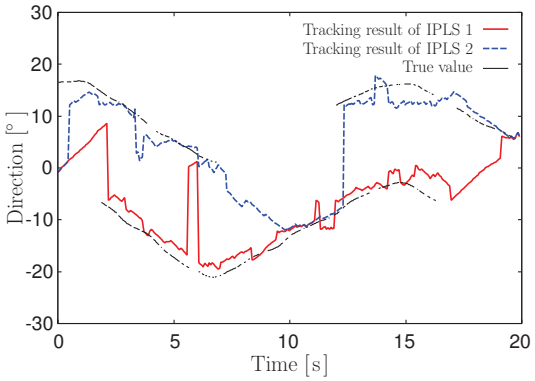


Figure 8. Tracking results of pattern 3

5. Conclusion

In this paper, the avoidance of spatial aliasing in multiple sound source tracking was studied. The spatial aliasing was avoided by the multiplication of the evaluation functions of whole frequency bands. The experimental results in an actual room showed the effectiveness of the proposed method.

Acknowledgment

This work was supported by the Research Institute for Science and Technology, Tokyo Denki University, Q16J-03.

References

[1] A. Quinlan and F. Asano, "Tracking a vary number of speaker using particle filtering," Proc.of IEEE ICASSP 2008, pp.297-300, February 2008.

[2] F. Talantzis, "An acoustic source localization and tracking framework using particle filtering and information theory," IEEE Trans., Audio, Speech, and Language Process., vol.18, no.7, pp.1806-1817, July 2010.

[3] N. Kikuma, "Iterative DOA estimation using subspace tracking methods and adaptive beamforming," IEICE Trans. Commun., vol.E88-B, no.5, pp.1818-1828, May 2005.

[4] B. Yang, "Projection approximation subspace tracking," IEEE Trans. SP., vol.43, no.1, pp.95-107, January 1995.

[5] K.H. Afkhamie, Z–Q. Luo, and K.M. Wong, "Adaptive linear filtering using interior point optimization techniques," IEEE Trans. SP., vol.48, no.6, pp.1637-1648, June 2000.

[6] N. Ohwada and K. Suyama, "Multiple Sound Source Tracking Method Based on Subspace Trackng," Proc. of IEEE WASPPA2009, pp.217-220, October 2009.