Omnidirectional Sound Source Tracking Based on Sequential Updating Histogram

Yusuke SHIIKI* 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 for omnidirectional sound source tracking using a circular microphone array is proposed. The sequential updating histogram estimated every two microphones are integrated for the sound source tracking. The histogram is estimated by weighting those reliability to results obtained every adjacent microphone pair. In addition, the wrapped Cauchy distribution is used to detect the omnidirectional DOA. As a result, the accurate omnidirectional sound source tracking can be achieved. Several experimental results are shown to present the effectiveness of the proposed method.**

I. Introduction

Sound source tracking is an important technique in various applications including a hands-free communication or a video conferencing. In these applications, the multiple omnidirectional sound source tracking is often required. In a single source scenario, it is well-known that a particle filter is a powerful tool for the tracking. However, in a multiple source scenario, the particle filter often fails tracking. It causes the same source estimation problem which occurs when either sound source begins to utter after a while silent period [5]. Then, the particles persuiting the original source concentrate to the other source and can not catch the original source again. Although the PAST-IPLS method succeeded to resolve such a problem, it can be applied to just a linear array.

To avoid such a drawback, the two microphone system has been paid an attention [6]∼[8]. Among them, the sequential updating histogram based on a speech sparseness [6] has achieved the multiple sound source tracking in real time. In this method, the estimated histogram at each frame is evaluated by a reliability weight. Then, the problem estimating the same direction does not occur because the histogram indicates multiple peaks corresponding to the each sound source direction. In addition, the Cauchy distribution that is robust to the outlier is fitted to the histogram to detect the DOA (Direction-Of-Arrival) by the EM algorithm. Therefore, the high accuracy DOA estimation has been achieved by just two microphones. Although this scheme is a promising approach to us, a difference between the front and the back can not be detected in a scenario of the omnidirectional DOA estimation.

On the other hand, a circular microphone array is often used for the omnidirectional sound source tracking [9]∼[11]. In [10], the single sound source tracking has been achieved using the particle filter and Von Mises distribution. In [11], the multiple sound source localization suceeded by using the histogram of the estimated results based on the W-disjoint orthogonal (WDO) assumption. In this method, MP (Matching Pursuit) is used for the DOA estimation using the histogram. However, the multiple sound source tracking may be difficult because MP is the high computation cost. Thus, the sound source tracking for three or more sources is not attempted. In addition, the tracking accuracy is not evaluated numerically. In [9]∼[11], the GCC-PHAT (Generalized Cross-Correlation PHAse Transform) is used for the DOA estimation. However, such a method makes the estimation accuracy decrease in a noisy and a reverberant environment.

In the proposed method, the sequential updating histogram [6] is integrated for the circular microphone array. To reduce the computation cost, the DOAs are estimated every adjacent microphone pair. Then, Root-MUSIC that is the robust method against the noise and the reverberations is applied for the DOA estimation. The reliability of the estimated DOA by Root-MUSIC was evaluated by the power ratio, and thus peaks corresponding to sound source direction are enhanced. Furthermore, the wrapped Cauchy distribution is used to detect the omnidirectional DOA. Therefore, the multiple omnidirectional sound source tracking can be achieved.

Several experimental results are shown to present the effectiveness of the proposed method.

II. Problem Description

As shown in Fig. 1, two sound sources, $s_i(n)$, $i = 1, 2$, move with time, and sound signals, $x_m(n)$, $m = 1, 2, \cdots, M$, are received by the circularly-arranged M microphones. In the frequency domain, the received signal of the m-th microphone can be written as

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

where t is a frame index, k is a frequency index, $S_i(t,k)$ is complex amplitude of $s_i(n)$, ω_k is an angular frequency at k, $\Gamma_m(t,k)$ is a noise observed at m-th microphone, and $\tau_i(t)$ is the TDOA (Time-Difference-Of-Arrival) defined as below,

$$\tau_i(t) = \frac{d\cos\left(\theta_i(t) - (m-1)\alpha\right)}{c} \qquad (2)$$

where $\theta_i(t)$ is the direction of the i-th sound source, c is the velocity of sound, α is the angle between the microphone pair, $d = 2r\sin(\alpha/2)$ is the microphone width, and r is the radius

of the circular microphone array. Moreover, using the vector notation,

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

where $\boldsymbol{a}_k(\theta_i(t)) = [1, e^{-j\omega_k\tau_i(t)}, \cdots, e^{-j\omega_k(m-1)\tau_i(t)}]^T$ is a transfer-function vector and $\boldsymbol{\Gamma}(t,k) = [\Gamma_1(t,k), \cdots, \Gamma_M(t,k)]^T$ is a noise vector. The aim of sound source tracking is to estimate $\theta_i(t)$ from the received signal $\boldsymbol{X}(t,k)$.

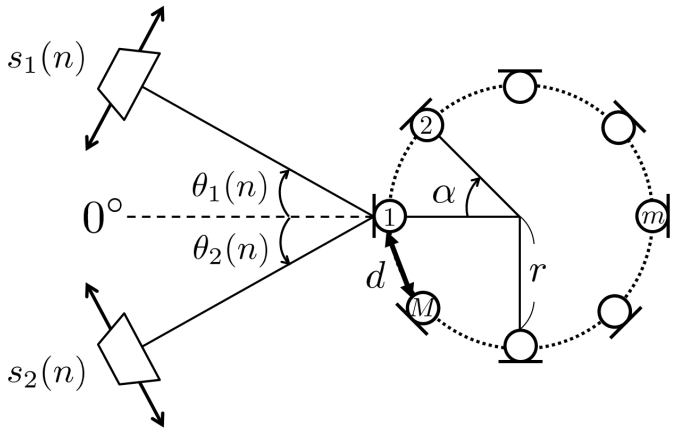

Fig. 1. Problem description.

III. The Proposed Method

A procedure of the proposed method is shown in Fig. 2. The sequential updating histogram based on a speech sparseness every two microphones is integrated for the sound source tracking as following:

1) $x_m(n)$ are transformed into the frequency domain by the DFT (Discrete Fourier Transform) and $X_m(t,k)$ is calculated.
2) The correlation matrix $\boldsymbol{R}(t,k)$ for the Root-MUSIC is calculated using $X_m(t,k)$. $\boldsymbol{R}(t,k)$ is calculated as below,

$$\boldsymbol{R}(t,k) = \boldsymbol{X}(t,k)\boldsymbol{X}^H(t,k) + \beta\boldsymbol{R}(t-1,k), \quad (4)$$

where β is a forgetting factor, and H is a Hermitian transpose.
3) DOA $\hat{\theta}_{m,m+1}(t,k)$ by the m-th and the m+1-th microphones in each time-frequency region is estimated by Root-MUSIC.
4) The reliability of $\hat{\theta}_{m,m+1}(t,k)$ is evaluated by the power ratio weight $w_p(t,k)$.
5) The reliability weighted histogram $\eta_t(C_{cell})$ is estimated from $\hat{\theta}_{m,m+1}(t,k)$.
6) $\eta_t^{'}(C_{cell})$ is sequentially updated as following,

$$\eta_t^{'}(C_{cell}) = w_u\eta_t(C_{cell}) + (1-w_u)\eta_{t-1}^{'}(C_{cell}), \quad (5)$$

where w_u is the updating weight.
7) The wrapped Cauchy mixture distribution is fitted to $\eta_t^{'}(C_{cell})$ to detect $\theta_i(t)$ by the EM algorithm.

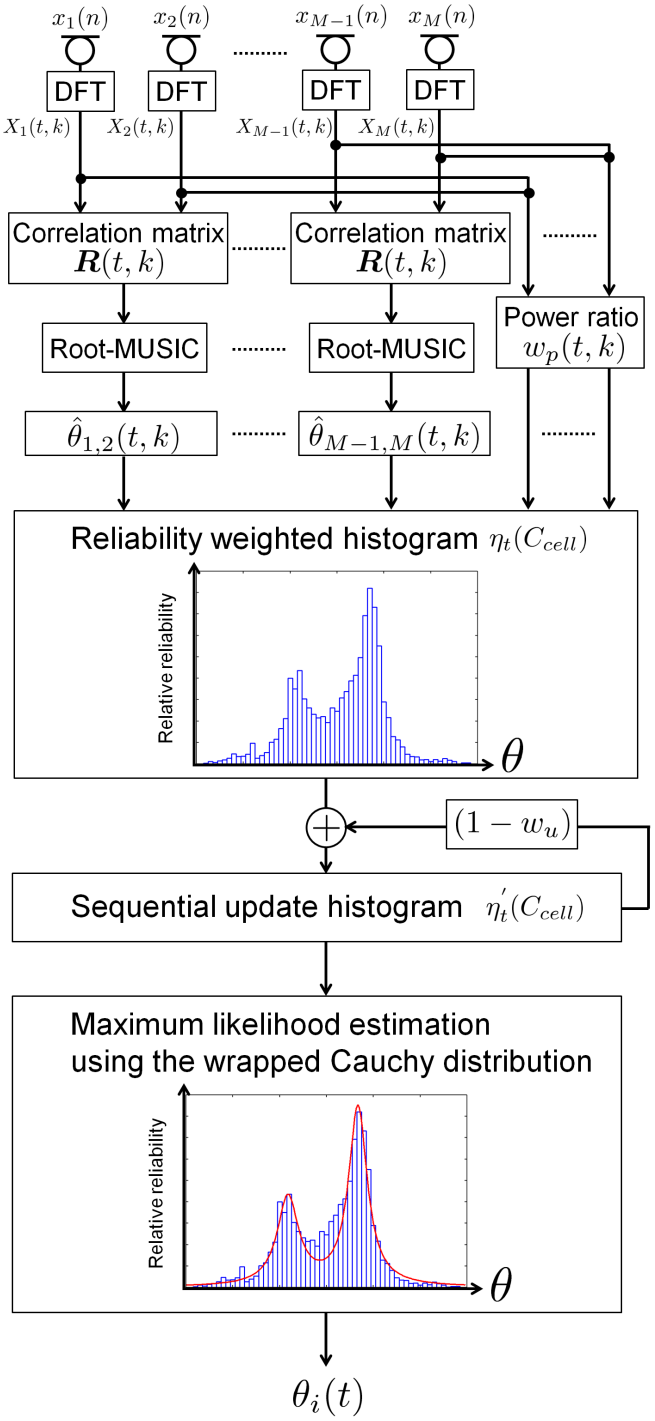

Fig. 2. A procedure of the proposed method.

In the proposed method, $\hat{\theta}_{m,m+1}(t,k)$ of all microphone pairs are used to estimate $\eta_t(C_{cell})$. Therefore, the wrong $\hat{\theta}_{m,m+1}(t,k)$ occurred by the phase ambiguity when the sound source exists behind microphone pairs are included in $\eta_t(C_{cell})$. However, because the frequencies of such $\hat{\theta}_{m,m+1}(t,k)$ are extremely low in all microphone pairs, it can be easily assumed that those results do not appear in $\eta_t(C_{cell})$.

A. The speech sparseness

A speech energy distribution of two speakers is shown in Fig. 3, and the color difference between blue and red presents the speaker difference.

As shown in Fig. 3, the each speech energies are sparsely distributed on the time-frequency plane. In addition, the distribution of the each speech energies are different every speech signal. Therefore, there exist a lot of regions which the single speech energy is dominant. In such regions, it is more likely to succeed to the DOA estimation by using two microphones.

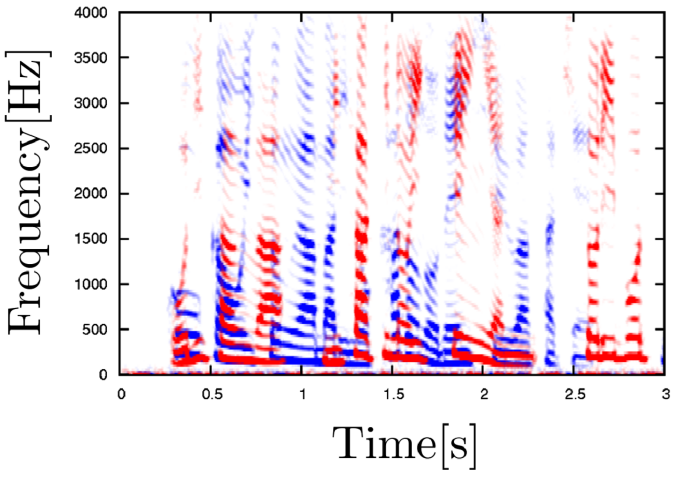

Fig. 3. The speech energy distribution.

B. Root-MUSIC

Root-MUSIC is the DOA estimation method. Root-MUSIC is based on an orthogonality between the signal subspace and the noise subspace. These subspace are calculated by the eigenvalue decomposition of $\boldsymbol{R}(t,k)$. The orthogonality between these subspaces is evaluated using following MUSIC spectrum function,

$$P_{MU}(\theta_i(t)) = \frac{\boldsymbol{a}_k^H(\theta_i(t))\boldsymbol{a}_k(\theta_i(t))}{\boldsymbol{a}_k^H(\theta_i(t))\boldsymbol{q}_k(t)\boldsymbol{q}_k^H(t)\boldsymbol{a}_k(\theta_i(t))}, \tag{6}$$

where $\boldsymbol{q}_k(t)$ is the noise subspace. (6) indicates a sharp peak around the direction corresponding to the DOA.

In Root-MUSIC, the denominator polynomial of (6) is directly solved for the DOA estimation as following,

$$\boldsymbol{a}_k^H(\theta_i(t))\boldsymbol{q}_k(t)\boldsymbol{q}_k^H(t)\boldsymbol{a}_k(\theta_i(t)) = 0. \tag{7}$$

C. Power ratio weight

In the time-frequency regions that the specific signal powers are strong, that signals are assumed to be dominant. Therefore, the reliability of the estimated DOA is high in such a region. In the proposed method, the estimated DOA is evaluated by the power ratio $w_p(t,k)$ defined by,

$$w_p(t,k) = \frac{P(t,k)}{\sum_k P(t,k)}, \tag{8}$$

where $P(t,k) = (|X_m(t,k)|^2 + |X_{m+1}(t,k)|^2)/2$. In the proposed method, the histogram of estimated DOAs are weighted by $w_p(t,k)$.

D. The wrapped Cauchy distribution

The sequential updating histogram includes several outliers because just a few estimation results are used for updating it. Therefore, the Cauchy distribution is fitted to $\eta_t'(C_{cell})$ to detect $\theta_i(t)$ by the EM algorithm. The omnidirectional DOA has to be estimated in [-180°,180°]. Then, -180° and 180° are seemed to be the same direction. When the sound source exists around 180°, $\eta_t'(C_{cell})$ tends to indicate a peak on both -180° and 180°. Therefore, these peaks have to be considered as the same direction. However, it is difficult to fit the normal mixed Cauchy distribution to $\eta_t'(C_{cell})$ because the Cauchy distribution is defined on the linear axis as shown in the following equation,

$$F(\theta) = \sum_{i=1}^{N} w_i \left[\frac{1}{\pi}\left\{\frac{\gamma}{(\theta_i - \bar{\theta}_i)^2 + \gamma^2}\right\}\right], \tag{9}$$

where w_i is a mixture ratio, $\bar{\theta}_i$ is a mode value, and γ is a half width at half maximum. As shown in Fig. 4, the normal Cauchy distribution can not detect two peaks on both sides as one peak. Therefore, we have to use the spherical distribution on the circular axis for the omnidirectional DOA estimation.

In the proposed method, the wrapped Cauchy distribution $F(\theta)$ is adopted on the circular axis. $F(\theta)$ is defined by,

$$F(\theta) = \sum_{i=1}^{N} w_i \left[\frac{1}{2\pi}\left\{\frac{1-\gamma^2}{1+\gamma^2-2\gamma\cos(\theta_i - \bar{\theta}_i)}\right\}\right]. \tag{10}$$

As shown in Fig. 5, the wrapped Cauchy distribution can be fitted to the histogram having peaks on both sides appropriately.

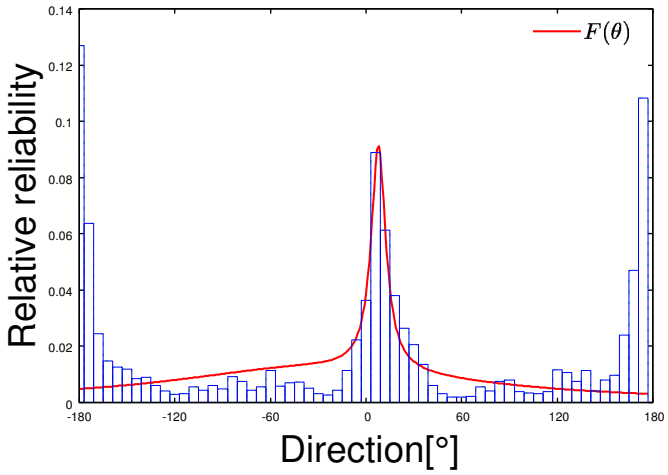

Fig. 4. The Cauchy distribution.

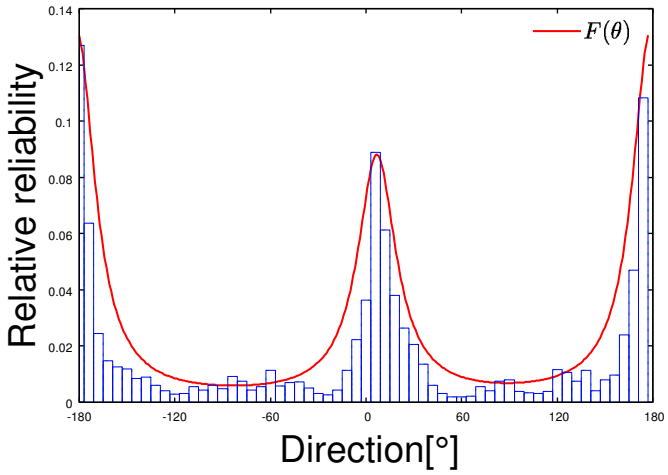

Fig. 5. The wrapped Cauchy distribution.

IV. Experiments in Real Environments

To evaluate the effectiveness of the proposed method, several experiments were conducted in real environments. The experimental conditions are listed in Table. I. The speech signals recorded in RWCP Sound Scene Database in Real Environments were used as the sound source signals. The accuracy of sound source tracking was measured by the RMSE (Root Mean Square Error). RMSE ε is calculated as below,

$$\varepsilon = \sqrt{\frac{1}{2}\sum_{i=1}^{2}\overline{(\hat{\theta}_i(t) - \theta_i(t))^2}}, \tag{11}$$

where $\bar{\cdot}$ is the time average, $\hat{\theta}_i(t)$ is the true value of the i-th sound source, and $\theta_i(t)$ is the estimated value. The average of RMSE for 12 source patterns was calculated for the evaluation. In addition, the evaluation of the real-time processing was measured by the RTF (Real Time Factor). The PC equipped with Intel Core 2 Quad 2.83[GHz] and 4[GByte] memory was used for an implementation. The proposed method was compared with the method using the normal Cauchy distribution.

TABLE I
Experimental conditions

reverberation time	0.3[s]
noise level	18.9[dB]
the number of sources	2
the number of microphones	16
microphone width	5.85[cm]
sampling frequency	8000[Hz]
flame size	512
overlap size	256
signal time	4.0-5.0[s]
frequency band for sound source tracking	500-4000[Hz]
source pattern	12

A. Tracking results in two source scenario

The tracking results for two source tracking are shown in Fig. 6, Fig. 8, and Fig. 10. As a comparison , the tracking results when the normal Cauchy distributions were used, are shown in Fig. 7, Fig. 9, and Fig. 11. For revealing a difference between the front and the back of array, the tracking results the proposed method and the comparison method are shown from Fig. 12 to Fig. 17, in which the results are depicted on the circular coordinate. The RMSEs and the RTFs for 12 patterns are listed in Table. Ⅱ.

In Fig. 7, Fig. 9, and Fig. 11, the comparison method failed the tracking around 180° because the normal Cauchy distribution could not detect the peak of the histogram on both -180° and 180°. In Fig. 6, Fig. 8, and Fig. 10, the proposed method succeeded the tracking within [-180°,180°] because the wrapped Cauchy distribution could detect the both peaks.

In Fig. 13, Fig. 15, and Fig. 17, the comparison method estimated the wrong position because the normal Cauchy distribution has failed the DOA estimation. In Fig. 12, Fig. 14, and Fig. 16, the proposed method succeeded the multiple sound source tracking even if the sound sources exist both the front and the back simultaneously.

In Table. Ⅱ, the average of RTF was 0.35, the average of RMSE of the comparison method for 12 patterns was 8.14°, and the average of RMSE of the proposed method for 12 patterns was 2.53°. Therefore, the proposed method has accurately achieved the multiple omnidirectional sound source tracking in real time for all patterns.

In addition, [9] and [10] have achieved the single omnidirectional sound source tracking but the multiple sound source tracking is untested. Among them, [10] was used the particle filter. When the particle filter is used for the multiple sound source tracking, a problem estimating the same direction occurs. In the proposed method, this problem does not occur because the estimated histogram can cluster each sound source direction.

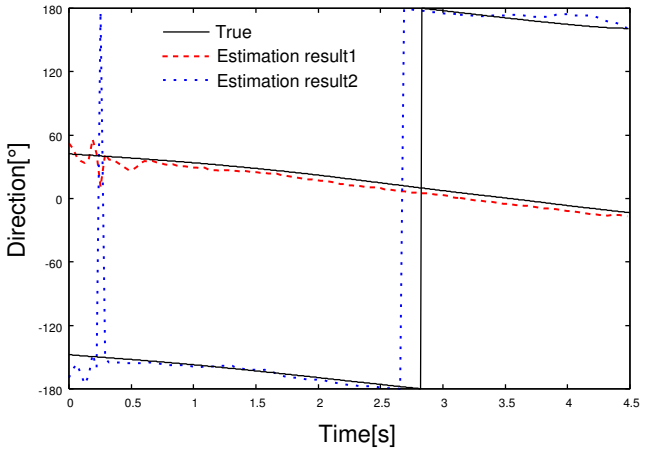

Fig. 6. Tracking results using the wrapped Cauchy distribution depicted over the direction coordinate: pattern1.

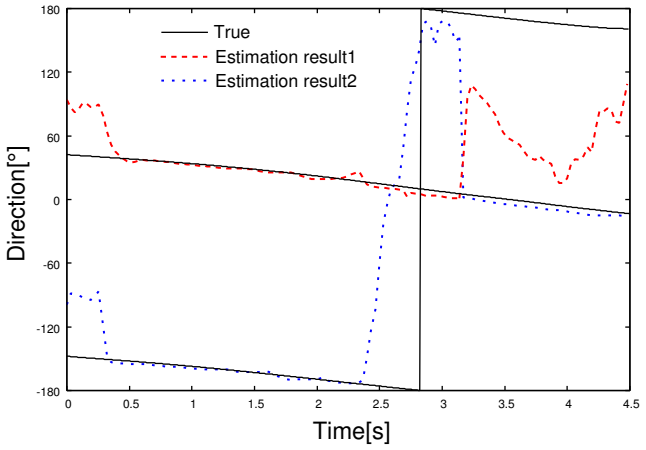

Fig. 7. Tracking results using the normal Cauchy distribution depicted over the direction coordinate: pattern1.

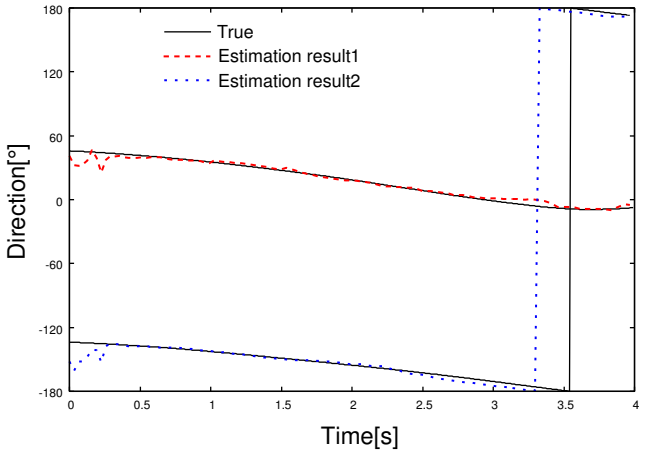

Fig. 8. Tracking results using the wrapped Cauchy distribution depicted over the direction coordinate: pattern7.

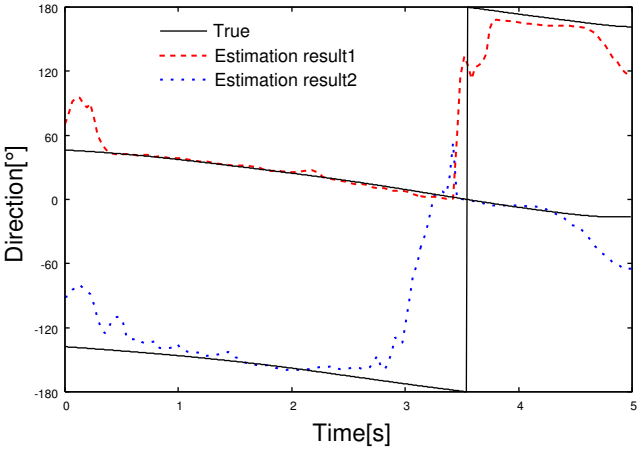

Fig. 9. Tracking results using the normal Cauchy distribution depicted over the direction coordinate: pattern7.

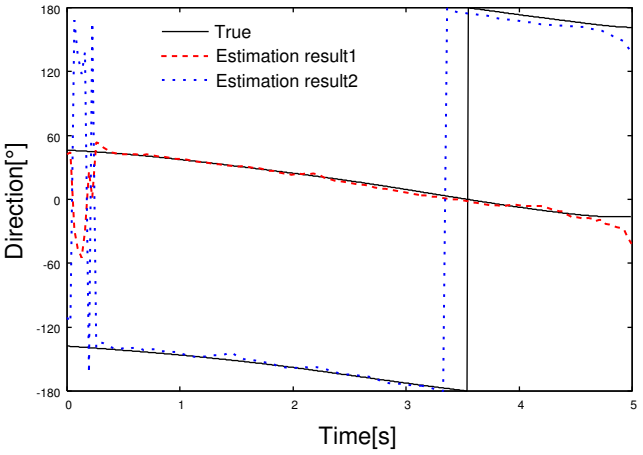

Fig. 10. Tracking results using the wrapped Cauchy distribution depicted over the direction coordinate: pattern11.

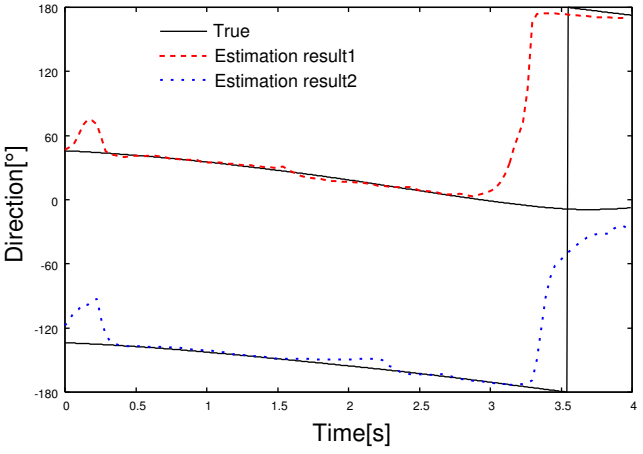

Fig. 11. Tracking results using the normal Cauchy distribution depicted over the direction coordinate: pattern11.

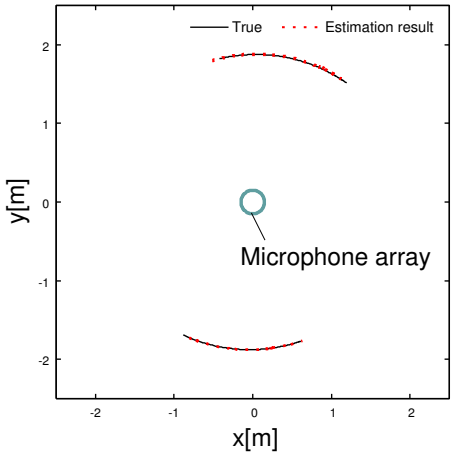

Fig. 12. Tracking results using the wrapped Cauchy distribution depicted over the circular coordinate: pattern1.

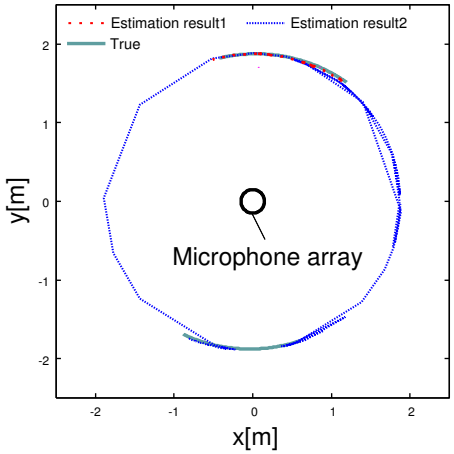

Fig. 13. Tracking results using the normal Cauchy distribution depicted over the circular coordinate: pattern1.

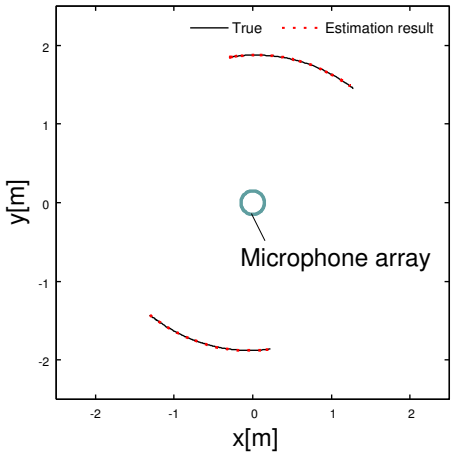

Fig. 14. Tracking results using the wrapped Cauchy distribution depicted over the circular coordinate: pattern7.

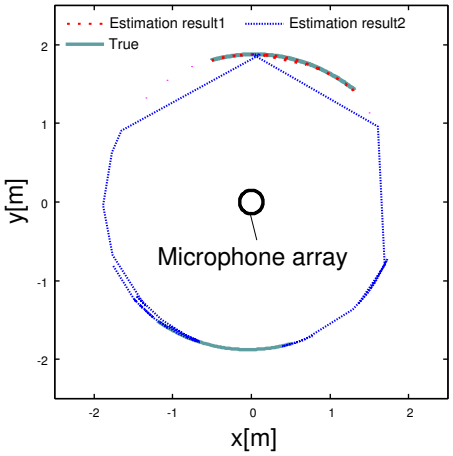

Fig. 15. Tracking results using the normal Cauchy distribution depicted over the circular coordinate: pattern7.

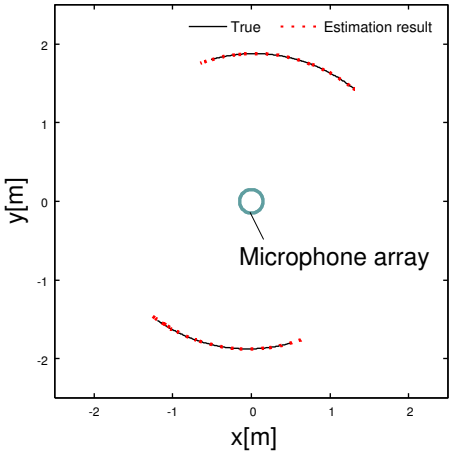

Fig. 16. Tracking results using the wrapped Cauchy distribution depicted over the circular coordinate: pattern11.

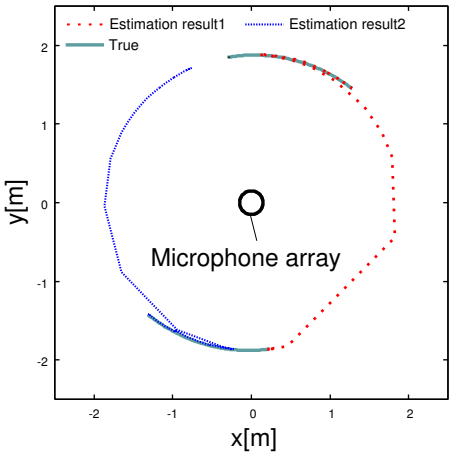

Fig. 17. Tracking results using the normal Cauchy distribution depicted over the circular coordinate: pattern11.

TABLE II
THE RESULTS OF RMSE AND RTF FOR TWO SOURCES

source pattern	RMSE[°]		RTF
	the wrapped Cauchy	the normal Cauchy	
1	3.71	9.78	0.35
2	3.66	10.40	0.35
3	2.49	11.74	0.35
4	2.59	11.87	0.35
5	1.93	3.84	0.35
6	2.09	5.56	0.35
7	2.54	16.63	0.36
8	2.67	4.91	0.36
9	2.69	4.27	0.36
10	2.66	9.20	0.36
11	1.65	3.11	0.35
12	1.65	6.35	0.35
average	2.53	8.14	0.35

B. Tracking result in multiple source scenario

To evaluate the tracking accuracy in three or more source tracking, several experiments were conducted on the experimental conditions same as Table. I. The tracking results for three sources of the proposed method are shown in Fig. 18, and the tracking results for four sources are shown in Fig. 19. The RMSE and the RTF for three and four sources are listed in Table. Ⅲ.

In Fig. 18 and Fig. 19, the proposed method succeeded the tracking for three and four sources. In Table. Ⅲ, the average of RMSE of three sources was 5.04°, and the average of RTF was 0.38. The average of RMSE of four sources was 7.56°, and the average of RTF was 0.41. Therefore, it was confirmed that the proposed method achieved the multiple omnidirectional sound source tracking in real time even for three or four sources.

In [11], the results of omnidirectional two sound source tracking are shown. However, a tracking performance is not revealed numerically.

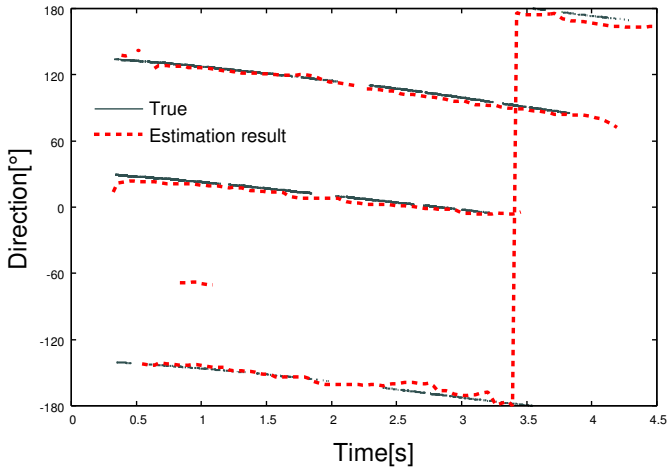

Fig. 18. Tracking results using the wrapped Cauchy distribution depicted over the direction coordinate: 3 sources.

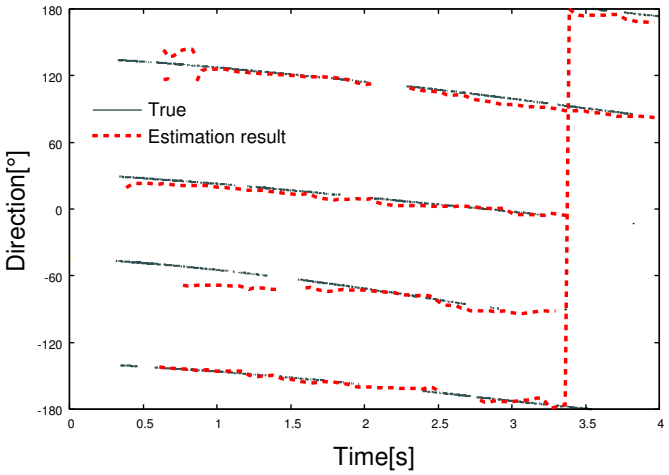

Fig. 19. Tracking results using the wrapped Cauchy distribution depicted over the direction coordinate: 4 sources.

TABLE III
THE RESULTS OF RMSE AND RTF FOR THREE OR MORE SOURCES

source pattern	RMSE[°]	RTF
3 sources (9 patterns)	5.04	0.38
4 sources (3 patterns)	7.56	0.41

V. CONCLUSIONS

In this paper, the method for the multiple omnidirectional sound source tracking based on the sequential updating histogram was proposed. In the proposed method, the reliability of the estimated DOA by Root-MUSIC was evaluated by the power ratio, and the reliabilities around the directions of sound sources were enhanced. Furthermore, the wrapped Cauchy distribution was used to detect the omnidirectional DOA. Several experimental results were shown to present the effectiveness of the proposed method.

ACKNOWLEDGMENT

This work was supported by the Grant-in-Aid for Scientific Research(C), No.15K06084, KAKENHI, JSPS.

REFERENCES

[1] D. B. Ward, E. A. Lehmann, and R. C. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE Trans. ASL,* vol. 11, no. 6, pp. 826-836, November 2003.
[2] A. Quinlan and F. Asano, "Tracking a vary number of speaker using particle filtering," *Proc. IEEE ICASSP 2008,* pp. 297-300, 2008.
[3] M. F. Fallon and S. Godsill, "Acoustic source localization and tracking using track before detect," *IEEE Trans. ASL,* vol. 18, no. 6, pp. 1228-1242, August 2010.
[4] A .Kizima, Y.Hioka, and N .Hamada, "Tracking of multiple moving sound sources using particle filter for arbitrary microphone array configurations," *Proc. IEEE ISPACS 2012,* pp. 108-113, November 2012.
[5] N. Ohwada and K. Suyama, "Multiple Sound Sources Tracking Method Based on Subspace Tracking," *Proc. IEEE WASPAA 2009,* pp. 217-220, October 2009.
[6] M.Hirakawa and K.Suyama, "Multiple sound source tracking by two microphones using PSO," *Proc. IEEE ISPACS 2013,* pp. 467-470, November 2013.

[7] Wenyi Zhang and B D.Reo, "A Two Microphone-Based Approach for Source Localization of Multiple Speech Sources," *IEEE Trans. ASL,* vol. 18, no. 8, pp. 1913-1928, November 2010.
[8] Nicoleta Roman and DeLiang Wang, "Binaural Tracking of Multiple Moving Sources," *IEEE Trans. ASL,* vol. 16, no. 4, pp. 728-739, May 2008.
[9] A. Karbasi and A. Sugiyama, "A new DOA estimation method using a circular microphone array," *Proc. EUSIPCO 2007,* pp. 778-782, 2007.
[10] Ivan Marković, and Ivan Petrović, "Speaker localization and tracking with a microphone array on a mobile robot using Von Mises distribution and particle filtering," *Robotics and Autonomous Systems,* vol. 58, no. 11, pp. 1185-1196, November 2010.
[11] Despoina Pavlidi, Anthony Griffin, Matthieu Puigt, and Athanasios Mouchtaris, "Real-Time Multiple Sound Source Localization and Counting Using a Circular Microphone Array," *IEEE Trans. ASL,* vol. 21, no. 10, pp. 2193-2206, October 2013.