International Journal of Electrical, Electronic and Communication Sciences
ISSN: 2517-9438
Vol:5, No:12, 2011

Multiple Moving Talker Tracking by Integration of Two Successive Algorithms

Kenji Suyama, Masahiro Oshida, and Noboru Owada

Abstract—In this paper, an estimation accuracy of multiple moving talker tracking using a microphone array is improved. The tracking can be achieved by the adaptive method in which two algorithms are integrated, namely, the PAST (Projection Approximation Subspace Tracking) algorithm and the IPLS (Interior Point Least Square) algorithm. When either talker begins to speak again after a silent period, an appropriate feasible region for an evaluation function of the IPLS algorithm might not be set. Then, the tracking fails due to the incorrect updating. Therefore, if an increment of the number of active talkers is detected, the feasible region must be reset. Then, a low cost realization is required for the high speed tracking and a high accuracy realization is desired for the precise tracking. In this paper, the directions roughly estimated using the delayed-sum-array method are used for the resetting. Several results of experiments performed in an actual room environment show the effectiveness of the proposed method.

Keywords—moving talkers tracking, microphone array, signal subspace

I. Introduction

Microphone array signal processing is an important technique in several acoustic applications, including teleconferencing and human interface [1]. In many applications, it is easy to suppose that there are multiple talkers simultaneously and they are free to move. Therefore, moving sound source tracking methods have been developed during the previous decade. Several methods among them are based on the particle filters which estimate talker directions in a stochastic framework [2].

On the other hand, the method in an adaptive signal processing framework was proposed in [3]. The method is based on the MUSIC (MUltiple SIgnal Classification) method [4] which is well-known as the high resolution estimation method. In the method, the PAST (Projection Approximation Subspace Tracking) algorithm [5] and the IPLS (Interior Point Least Square) algorithm [6] are integrated to estimate talker directions successively. The method firstly estimates the signal subspace using the PAST algorithm without an eigen-decomposition of the correlation matrix of the array acquisition signals. Secondly, each the direction-of-arrival (DOA) is estimated using the IPLS algorithm without a peak search of the MUSIC spectrum function. Then, if an appropriate feasible region is set for an evaluation function of the IPLS algorithm, the method can track even if there are multiple talkers simultaneously. However, when the talkers begin to speak again after a silent period or a low speech energy period, the feasible region might not be set to the appropriate interval of the evaluation function. Then, the DOA estimation fails because of the incorrect updating of the IPLS algorithm. Thus, a resetting scheme is required in which the number of active talkers and those rough directions must be estimated fast. In [7], the histogram based method is applied in which the direction estimates by the instantaneous phase difference between two microphones are used for constructing the histogram. Although this method can operate fast, the estimation accuracy was not always high due to the low spatial resolution of the two microphones.

Department of Electrical and Electronic Engineering, School of Engineering, Tokyo Denki University, 2-2, Kandanishiki-cho, Chiyoda-ku, Tokyo, Japan, 101-8457, e-mail: suyama@cck.dendai.ac.jp,

This work was supported by the Grant-in-Aid for Scientific Research (C), No.23560468, KAKENHI, JSPS.

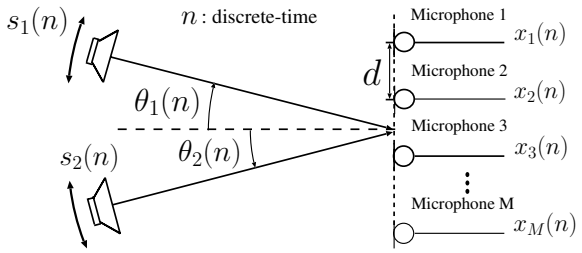

Fig. 1. Acoustic signal acquisition model using M channel microphone array.

In this paper, a delayed-sum-array method is applied for the resetting of the feasible region alternatively. The method can estimate the rough directions fast similar to the histogram based method. Moreover, the array with wide microphone width is used for improving the spatial resolution. Thus, the estimation accuracy can be improved while maintaining low computational cost. Several results of experiments performed in an actual room environment show the effectiveness of the proposed method.

II. Moving Talker Tracking Problem

As shown in Fig.1, two talkers move with time, and the speech signals are received by the equally spaced microphone array. The microphone array has M microphones. The received signal of m-th microphone can be written in the frequency domain as follows:

$$X_{t,m}(k) = \sum_{j=1}^{2} S_{j,t}(k) e^{-j\omega_k (m-1)\tau_{j,\theta(t)}} + \Gamma_m(k), \qquad (1)$$

where t is a frame index, k is a frequency index, $S_{j,t}(k)$ is a complex amplitude of $s_j(n)$, $\tau_{j,\theta(t)} = d \sin\theta_{j,t}/c$ is a time-difference-of-arrival between adjacent two microphones, $\theta_{j,t}$ is the direction of the j-th talker, c is the velocity of sound, and $\Gamma_m(k)$ is the noise at each microphone.

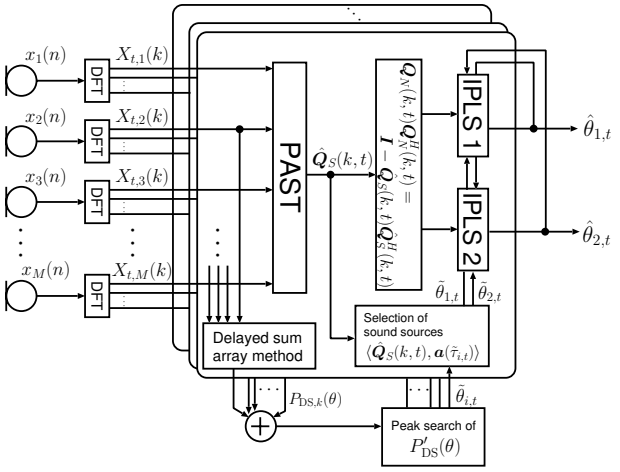

Fig. 2. The block diagram of the proposed multiple talker tracking method.

Using the vector notation

$$\boldsymbol{X}_t(k) = \sum_{j=1}^{2} S_{j,t}(k)\boldsymbol{a}(\theta_{j,t}) + \boldsymbol{\Gamma}(k) \tag{2}$$

where $\boldsymbol{a}(\theta_{j,t}) = [1, e^{-j\omega_k \tau_{j,\theta(t)}}, \cdots, e^{-j\omega_k (M-1)\tau_{j,\theta(t)}}]^T$ is a transfer-function vector, $\boldsymbol{\Gamma}(k) = [\Gamma_1(k), \Gamma_2(k), \cdots, \Gamma_M(k)]^T$ is a noise vector, and T denotes transposition.

The aim of sound sources tracking is to estimate $\theta_{j,t}$ using $\boldsymbol{X}_t(k)$.

III. Talker tracking using PAST-IPLS method

The moving talker tracking method using the PAST-IPLS method is based on the signal subspace that is obtained by the eigen-decomposition of the correlation matrix $\boldsymbol{R}_{k,t} = \overline{\boldsymbol{X}_t(k)\boldsymbol{X}_t^H(k)}$, where, $\overline{\cdot}$ represents time averaging and H represents the Hermitian transposition. The eigenvalues that are computed by the eigen-decomposition can be arranged in the following manner: $\lambda_1 \geq \lambda_2 > \lambda_3 = \cdots = \lambda_M$. The subspace that is spanned by the eigenvectors $\boldsymbol{q}_1(k,t)$ and $\boldsymbol{q}_2(k,t)$ corresponding to λ_1 and λ_2, respectively, is called the signal subspace $\boldsymbol{Q}_S(k,t)$.

The other subspace that is spanned by the other eigenvectors is called the noise subspace $\boldsymbol{Q}_N(k,t)$. The two subspaces are related via the orthocomplement. In the MUSIC algorithm, the orthogonality between $\boldsymbol{Q}_N(k,t)$ and $\boldsymbol{a}(\theta_{j,t})$ is evaluated by using MUSIC spectrum function. The objective of the DOA estimation is to search for peaks in MUSIC spectrum function.

The block diagram of the proposed method is shown in Fig.2. The main procedure is implemented in the frequency domain. First, the received signal $\boldsymbol{x}(n)$ in the time domain is transformed into $\boldsymbol{X}_t(k)$ by using the DFT (Discrete Fourier Transform). Second, for estimating the number of talkers and setting the appropriate initial value in the IPLS, the spatial power is calculated by the delayed-sum-array method. The peaks of power indicates the initial value that must be chosen. Next, $\boldsymbol{Q}_S(k,t)$ is sequentially updated using PAST[5] without performing the eigen-decomposition of $\boldsymbol{R}_{k,t}$.

Finally, the DOA is estimated by each IPLS without performing the peak-search in the MUSIC spectrum, where $\hat{\theta}_{j,t}(j = 1,2)$ is the average of the result over the entire frequency band.

A. PAST algorithm

In the proposed method, the PAST algorithm is utilized to calculate $\boldsymbol{Q}_S(k,t)$. The PAST algorithm is a sequential-update algorithm of the signal subspace whose basis vectors vary with frame. This algorithm is based on the idea that $\boldsymbol{Q}_S(k,t)$ can be calculated by minimizing the following evaluation function:

$$J_P(\boldsymbol{Q}_S(k,t)) = \sum_{i=1}^{t} \alpha^{t-i} \|\boldsymbol{X}_i(k) - \boldsymbol{Q}_S(k,t)\hat{\boldsymbol{Q}}_S^H(k,i-1)\boldsymbol{X}_i(k)\|^2,$$

where α is the forgetting factor, $\hat{\boldsymbol{Q}}_S(k,i-1)$ is the estimated value at the previous frames. The steps of the IPLS algorithm are summarized in [7].

B. IPLS algorithm

In the MUSIC spectrum function, the peak-search involves enormous computational costs because a large number of complex multiplications must be performed. The IPLS enables the DOA estimation by setting the appropriate feasible region without the peak-search. The objective of DOA estimation is to minimize the following evaluation function:

$$J_I(\tau_{j,\theta(t)}) = \boldsymbol{a}^H(\tau_{j,\theta(t)})\boldsymbol{U}_{N,j}(k,t)\boldsymbol{U}_{N,j}^H(k,t)\boldsymbol{a}(\tau_{j,\theta(t)}) \quad (j = 1,2), \tag{3}$$

where $\boldsymbol{U}_{N,j}(k,t)\boldsymbol{U}_{N,j}^H(k,t) = \boldsymbol{I} - \boldsymbol{q}_j(k,t)\boldsymbol{q}_j^H(k,t)$, $\boldsymbol{I}$ is the identity matrix, $J_I(\tau_{j,\theta(t)})$ is used to evaluate the orthogonality. Although $J_I(\tau_{j,\theta(t)})$ are non-convex functions, the IPLS algorithm can be used to track each global minimum by setting the appropriate feasible region in $J_I(\tau_{j,\theta(t)})$. Then, the constrained minimized problem is considered to be as follows:

$$\begin{aligned} &\min \quad J_I(\tau_{j,\theta(t)}), \\ &\text{sub. to} \quad J_I(\tau_{j,\theta(t)}) \leq \zeta_{j,t},\ \tau_{j,\theta(t)}^2 \leq \eta^2, \end{aligned}$$

where $\eta = d/c$. The feasible region $\Omega_{j,t}$ of this problem is defined as

$$\Omega_{j,t} = \{\tau_{j,\theta(t)} \in \mathbb{R} \,|\, J_I(\tau_{j,\theta(t)}) \leq \zeta_{j,t}, \tau_{j,\theta(t)}^2 \leq \eta^2\}. \tag{4}$$

For $\Omega_{j,t}$, the following logarithmic barrier function $\phi(\tau_{j,\theta(t)})$ is defined:

$$\phi(\tau_{j,\theta(t)}) = -\log(\zeta_{j,t} - J_I(\tau_{j,\theta(t)})) - \log(\eta^2 - \tau_{j,\theta(t)}^2), \tag{5}$$

$\phi(\tau_{j,\theta(t)})$ diverges to ∞ at the boundary of $\Omega_{j,t}$. Therefore, the existence of an analytic center is guaranteed in $\Omega_{j,t}$. The objective of the DOA estimation performed using the IPLS is to estimate the $\tau_{j,\theta(t)}$ that minimizes $\phi(\tau_{j,\theta(t)})$. For this purpose, the initial value $\tau_{j,\theta(t-1)}$ needs to be set appropriately for setting the appropriate feasible region. Moreover, $\Omega_{j,t}$ can be adjusted according to the following:

$$\zeta_{j,t} = J_I(\tau_{j,\theta(t-1)}) + \beta\frac{\eta}{\sqrt{2}}\left|\nabla J_I(\tau_{j,\theta(t-1)})\right|, \tag{6}$$

where $\beta > 0$ is a step-size parameter. The constraint $J_I(\tau_{j,\theta(t)}) \leq \zeta_{j,t}$ is varied in proportion to $|\nabla J_I(\tau_{j,\theta(t-1)})|$, and $\Omega_{j,t}$ is modified with time. The analytic center of $\phi(\tau_{j,\theta(t-1)})$ is updated by a single Newton step. The steps of the IPLS algorithm are summarized in [7].

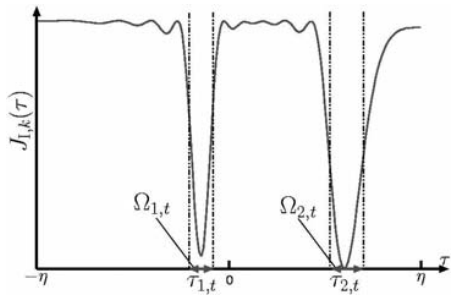

Fig. 3. Evaluation function of IPLS algorithm and those feasible regions in a case that two talkers exist.

IV. Resetting of Feasible Region

Although $J_I(\tau_{j,\theta(t)})$ has the global minimum for every talker direction during the active speech period, it might be that there is no minimum during the silent period or the low speech energy period. Therefore, a new $\Omega_{j,t}$ must be set if an increment of the number of active talkers is detected. Thus, it is required to estimate the number of active talkers and the those rough directions. Then, a low cost realization is required for the high speed tracking and a high accuracy realization is desired for the precise tracking. In [7], the histogram of the DOAs estimated by the instantaneous phase difference between two adjacent microphones are used for the resetting. That is, the number of talkers is estimated by the number of peaks of the histogram, and the talker directions to be estimated correspond to those peaks. Although this method can reset $\Omega_{j,t}$ fast, the estimation accuracy tends to be low because a spatial resolution of the two microphones is not always high.

In the proposed method, the delayed-sum-array (DSA) method is used for the resetting of $\Omega_{j,t}$. The DSA method calculates the power $P_{DS,k}(\theta)$ toward the look-direction θ and the peaks of $P_{DS,k}(\theta)$ indicate the talker directions. $P_{DS,k}(\theta)$ can be calculated by the following equation:

$$P_{DS,k}(\theta) = \boldsymbol{a}^H(\theta)\boldsymbol{X}_t(k). \tag{7}$$

For the resetting of $\Omega_{j,t}$, a summation of $P_{DS,k}(\theta)$ over all the frequencies considered as following:

$$P'_{DS}(\theta) = \sum_k P_{DS,k}(\theta). \tag{8}$$

Moreover, just a small number of microphones are used for the calculation of $P'_{DS}(\theta)$ because of the low computational requirement. The microphone width is spread widely to maintain the estimation accuracy. However, such an array brings $P'_{DS}(\theta)$ false peaks due to the spatial aliasing especially in the high frequency band. Therefore, a peak selection operation is carried out as shown in Fig.4. In the operation, the inner product $\langle \hat{\boldsymbol{Q}}_S(k,t), \boldsymbol{a}_k(\tilde{\tau}_{i,t})\rangle$ between $\hat{\boldsymbol{Q}}_S(k,t)$ and $\boldsymbol{a}_k(\tilde{\tau}_{i,t})$ is

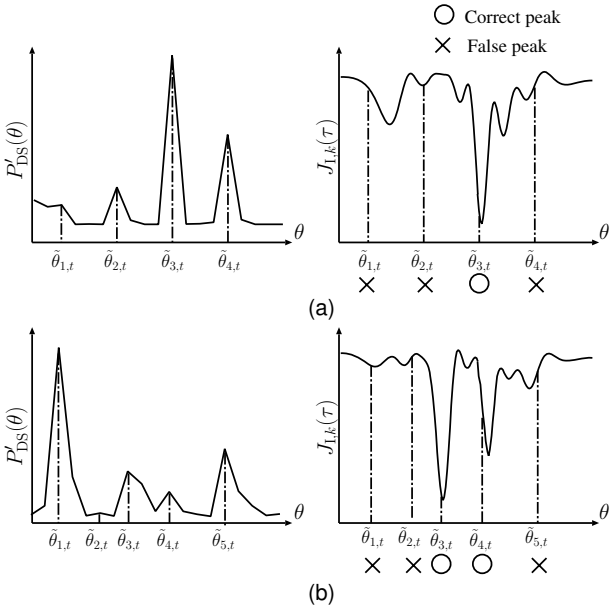

Fig. 4. Directions estimated from peaks of DSA power spectrum and corresponding peaks of IPLS evaluation function.

TABLE I
Experimental conditions.

Condition	Value
f_s	8 [kHz]
N	256
M	16
d	40 [mm]
Frequency band used for the PAST–IPLS	$2 \sim 4$ [kHz]
Initial value of IPLS $\hat{\tau}_{i,0}$	0
The number of microphones for delayed sum array method	4
Microphone width for delayed sum array method	160 [mm]
Frequency band used delayed sum array method	$3 \sim 4$ [kHz]
Width of peak search	10°
C_{th}	0.7

used as a criterion for deciding whether the peak is true or not, where $\tilde{\tau}_{i,t} = d\sin\tilde{\theta}_{i,t}/c$ is the i-th estimate of DSA in which false peaks are involved. Because the inner product takes a large value when $\tilde{\theta}_{i,t}$ corresponds to the true peak, just the correct talker directions can be selected for the resetting of $\Omega_{j,t}$.

V. Experimental Results

Several experiments in an actual room environment were conducted to confirm the efficiency of the proposed method. The noise level was 37.6[dB], and the reverberation time of the room was about 500[ms]. The experimental conditions are listed in Table.I. The threshold value for the decision of the true peak in the power spectrum was set to $C_{th} = 0.7$ from the preliminary experiments. The estimation accuracy was measured by the following equation:

$$\varepsilon = \sqrt{\frac{1}{2}\sum_{i=1}^{2}|\theta_{i,t} - \hat{\theta}_{i,t}|^2}, \tag{9}$$

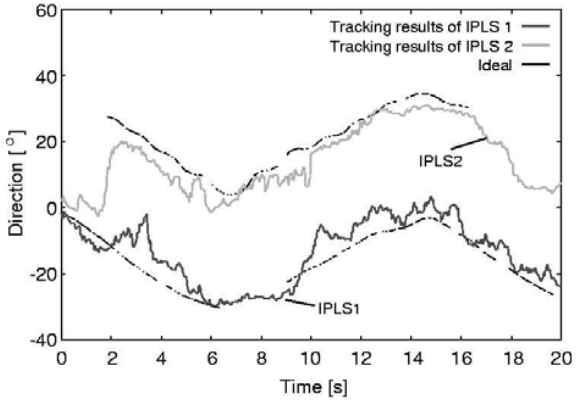

Fig. 5. Tracking results of pattern 1 when the proposed method was used.

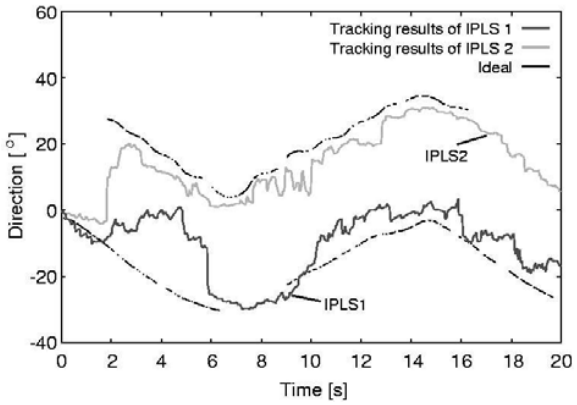

Fig. 6. Tracking results of pattern 1 when the conventional method was used.

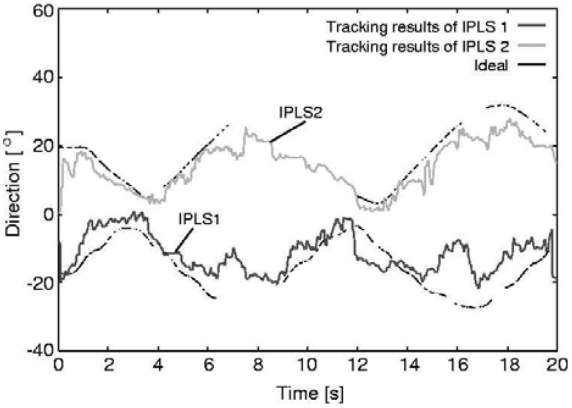

Fig. 7. Tracking results of pattern 2 when the proposed method was used.

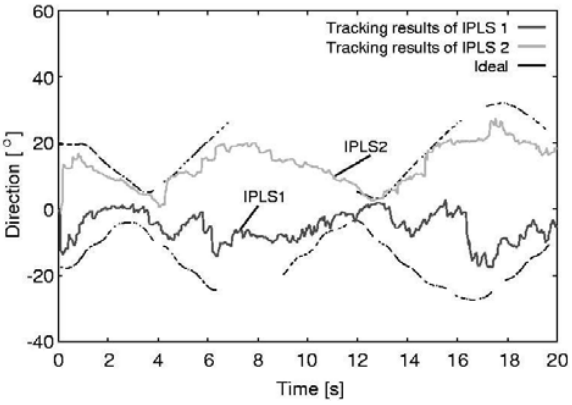

Fig. 8. Tracking results of pattern 2 when the conventional method was used.

where $\hat{\theta}_{i,t}$ is the true value of the talker direction which was measured by the camera array image sensing.

Four moving patterns have been attempted. The tracking results of pattern 1 and pattern 2 using the proposed method and the conventional method [7] are shown in Fig.5-8. The estimation error ε in each of the patterns are shown in Table.II. These results show that the proposed method can estimate better than the conventional method. In addition, the computational time for the array acquisition signals per 1[s] was 0.6[s] in either method.

TABLE II
COMPARISON OF ESTIMATION ERROR BETWEEN THE PROPOSED METHOD AND THE CONVENTIONAL METHOD

pattern	proposed method	conventional method
1	4.91	7.99
2	5.78	9.09
3	5.08	7.00
4	8.24	12.12

VI. Conclusion

In this paper, the moving talker tracking method by the integration of two successive algorithms was proposed. In the method, the delayed-sum-array method was used for estimating the number of active talkers and those rough directions for the resetting of the feasible region. The experimental results in the actual room environment showed the efficiency of the proposed method.

References

[1] M. Brandstein and D. Ward, "Microphone arrays signal processing techniques and applications," Springer, 2001.
[2] M. Fallon and S. Godsill, "Multi target acoustic source tracking using track before detect" in *Proc. IEEE WASPAA*, pp. 102–105, 2007.
[3] D. Tsuji and K. Suyama, "A moving sound source tracking based on two successive algorithms," in *Proc. IEEE ISCAS*, C2L-E5-5, 2009.
[4] R. O. Schmidt, "Multiple emitter location and signal parameter estimation" *IEEE Trans. IEEE AP.*, vol. 34, no. 3, pp. 276–280, 1986.
[5] B. Yang, "Projection approximation subspace tracking," *IEEE Trans. SP.*, vol. 43, no. 1, pp. 95–107, 1995.
[6] K. H. Afkhamie, Z. Luo, and K. M. Wong, "Adaptive linear filtering using interior point optimization techniques," *IEEE Trans. SP.*, vol. 48, no. 6, 2000.
[7] N. Ohwada and K. Suyama, "Multiple sound source tracking method based on subspace tracking," in *Proc. IEEE WASPAA2009*, pp.217-220, 2009.