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. Second, 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.

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:
\[
X_{t,m}(k) = \sum_{j=1}^{2} s_{j,t}(k) e^{-j\omega_k(m-1)\tau_{j,\theta(t)}} + \Gamma_m(k),
\]
(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.
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 \( Q_S(k, t) \). 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 \( Q_S(k, t) \) can be calculated by minimizing the following evaluation function:

\[
J_P(Q_S(k, t)) = \sum_{i=1}^{t} \alpha^{t-i} ||X_i(k) - Q_S(k, t) \hat{Q}_S^H(k, i-1) X_i(k) ||^2,
\]

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

### B. ICLS algorithm

In the MUSIC spectrum function, the peak-search involves enormous computational costs because a large number of complex multiplications must be performed. The ICLS 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:

\[
J_I(\tau_j, \hat{\theta}(t)) = a^H(\tau_j, \hat{\theta}(t)) U_N(\tau_j, k) U_N^H(\tau_j, k) a(\tau_j, \hat{\theta}(t))
\]

(\( j = 1, 2 \))

where \( U_N(\tau_j, k) U_N^H(\tau_j, k) = I - q_j(k, t) q_j^H(k, t) \), \( I \) is the identity matrix, \( J_I(\tau_j, \hat{\theta}(t)) \) is used to evaluate the orthogonality. Although \( J_I(\tau_j, \hat{\theta}(t)) \) are non-convex functions, the ICLS algorithm can be used to track each global minimum by setting the appropriate feasible region in \( J_I(\tau_j, \hat{\theta}(t)) \). Then, the constrained minimized problem is considered to be as follows:

\[
\begin{aligned}
\min & J_I(\tau_j, \hat{\theta}(t)), \\
\text{sub. to} & J_I(\tau_j, \hat{\theta}(t)) \leq \zeta_{j,t}, \tau_j^2 \hat{\theta}(t) \leq \eta^2,
\end{aligned}
\]

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

\[
\Omega_{j,t} = \{ \tau_j, \hat{\theta}(t) \in \mathbb{R} | J_I(\tau_j, \hat{\theta}(t)) \leq \zeta_{j,t}, \tau_j^2 \hat{\theta}(t) \leq \eta^2 \}.
\]

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

\[
\phi(\tau_j, \hat{\theta}(t)) = - \log(\zeta_{j,t} - J_I(\tau_j, \hat{\theta}(t))) - \log(\eta^2 - \tau_j^2 \hat{\theta}(t)),
\]

(5)

\( \phi(\tau_j, \hat{\theta}(t)) \) diverges to \( \infty \) 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 ICLS is to estimate the \( \tau_j, \hat{\theta}(t) \) that minimizes \( \phi(\tau_j, \hat{\theta}(t)) \). For this purpose, the initial value \( \tau_j, \hat{\theta}(t-1) \) needs to be set appropriately for setting the appropriate feasible region. Moreover, \( \Omega_{j,t} \) can be adjusted according to the following:

\[
\zeta_{j,t} = J_I(\tau_j, \hat{\theta}(t-1)) + \beta \eta \sqrt{2} |\nabla J_I(\tau_j, \hat{\theta}(t-1))|,
\]

(6)
where $\beta > 0$ is a step-size parameter. The constraint $J_I(\tau_{j,t}(t)) \leq \kappa_{j,t}$ is varied in proportion to $|\nabla J_I(\tau_{j,t}(t-1))|$, and $\Omega_{j,t}$ is modified with time. The analytic center of $\phi(\tau_{j,t}(t-1))$ is updated by a single Newton step. The steps of the IPLS algorithm are summarized in [7].

IV. Resetting of Feasible Region

Although $J_I(\tau_{j,t}(t))$ 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 $\theta$ and the peaks of $P_{DS,k}(\theta)$ indicate the talker directions. $P_{DS,k}(\theta)$ can be calculated by the following equation:

$$P_{DS,k}(\theta) = a^H(\theta)x_t(k).$$  (7)

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

$$P'_{DS}(\theta) = \sum_k P_{DS,k}(\theta).$$  (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 $(Q_S(k,t), a_k(\tilde{\theta}_{i,t}))$ between $Q_S(k,t)$ and $a_k(\tilde{\theta}_{i,t})$ is used as a criterion for deciding whether the peak is true or not, where $\tilde{\theta}_{i,t} = d \sin \theta_{i,t}/c$ 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[msec]. 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:

$$\varepsilon = \sqrt{\frac{1}{2} \sum_{i=1}^{2} |\hat{\theta}_{i,t} - \theta_{i,t}|^2},$$  (9)
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 $\varepsilon$ 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>
<thead>
<tr>
<th>pattern</th>
<th>proposed method</th>
<th>conventional method</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.31</td>
<td>7.99</td>
</tr>
<tr>
<td>2</td>
<td>5.78</td>
<td>9.09</td>
</tr>
<tr>
<td>3</td>
<td>5.08</td>
<td>7.00</td>
</tr>
<tr>
<td>4</td>
<td>8.24</td>
<td>12.12</td>
</tr>
</tbody>
</table>

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