A TRACK BEFORE DETECT APPROACH FOR SEQUENTIAL BAYESIAN TRACKING OF MULTIPLE SPEECH SOURCES

Pasi Pertilä, Member, IEEE
Tampere University of Technology
Department of Signal Processing
P.O. Box 553, Tampere, FI-33101, Finland
email: pasi.pertila@tut.fi

Matti S. Hämäläinen
Nokia Research Center
P.O. Box 1000, Tampere, FI-33721, Finland
email: matti.s.hamalainen@nokia.com

ABSTRACT
This paper describes a novel multiple acoustic source tracking method based on track before detect paradigm. Multiple particle filters are used to represent the state of all sources. Sources are detected and removed using a likelihood ratio obtained from particle weights. The weights are obtained by evaluating the likelihood of microphone pair phase difference. Tracking performance from recorded data with rich sequences of speech is presented using multiple object tracking metrics. Results show that the proposed method can detect and track multiple temporally overlapping speech sources as well as switching talkers even in weak signal-to-noise ratios.

Index Terms—Acoustic Tracking, Multiple Sources, Particle Filters, Likelihood Ratio, Track Management

1. INTRODUCTION
Acoustic source localization (ASL) has received much attention during past decades. Time difference of arrival (TDOA) methods have been extensively utilized for wide-band sources, such as the human talker. A popular TDOA estimation method is the generalized cross-correlation (GCC) with phase transform (PHAT) termed GCC-PHAT. However, even a slight reverberation can result in false GCC peaks and thus lead to erroneous TDOA values [1]. This corrupts the TDOA based localization performance.

A class of more robust localization schemes is based on maximizing the steered response power (SRP) of source position. For a wide-band source the PHAT weighting has been used successfully applied for ASL in reverberant environments [2]. Sequential information can be included efficiently in the form of Bayesian Monte Carlo based particle filtering (PF) [3], which is suitable for non-linear and non-Gaussian problems, such as the ASL.

In multiple target tracking (MTT) applications the measurement device outputs a thresholded observation and the problem is to associate the observation to a track or background. This is referred as the data association problem. An alternative approach is to use raw sensor data without thresholding. The detection is then based on the posterior information, i.e., the probability density of the state given the measurement history. This approach is termed track before detect (TBD) and is suited for tracking weak targets in noisy environments [4]. The TBD implicitly solves the data association problem and is seen as a natural approach for acoustic source localization, since measurements are difficult to threshold in a general way and the thresholding would present unnecessary false detections and missed observations as pointed out in [5].

In [5] TBD method is applied for multiple acoustic source tracking. The method uses a fixed measurement grid of steered beamformer (SBF) values evaluated at low frequencies [100,400] Hz. Any other SBF is evaluated with a grid at 10 cm distance integrated over frequencies [200,6000] Hz. A probabilistic approach based on likelihood ratios is considered for source detection. A single joint state PF is applied to tracking of multiple sources.

In [6] a method using particle filters based on SRP-PHAT measurement is considered for localization of two simultaneous sources. The method uses measurement exclusion by removing the first source evidence from the GCC-PHAT functions in order to search the second target. A pre-set SRP-PHAT threshold value is used to determine target existence. In [7] multiple target directions are estimated with SBF using data association, measurement exclusion and particle filtering.

A MTT method using TBD for direction estimation is considered in [8] with a mixture particle filter using TDOA measurements. The approach assigns particles to separate targets using a clustering approach. Since resampling is applied on a single particle filter a trajectory reconstruction method is also included.

This paper presents an acoustic MTT method based on particle filtering using TBD paradigm. The particle weights are proposed to be obtained from signal phase differences between microphone pairs, which are modeled as normally distributed. Each source is tracked with a corresponding particle filter. The source posterior density is propagated sequentially in time in the form of weighted particles. The set of particle filters represents the state of all sound sources. Track addition and deletion is based on a likelihood ratio test, and the likelihood ratio is extracted from particle weights [4]. A concept of initialization and tracking filters is introduced for detection, tracking, and removal of acoustic sources. The initialization filter combines in a novel way the i) detection of new targets by distributing a fraction of particles to search space ii) estimation of background level for likelihood ratio tests. This avoids the use of fixed grids and is therefore scalable to large spaces without extra computational cost. In addition, by using a relative and not absolute detection threshold, the method is adaptive to various background noise levels. The performance of the proposed method is evaluated with multiple object tracking (MOT) metrics, presented in [9]. The used data contains speaker switches, different amount of overlapping speech, and was acquired in three signal-to-noise ratio (SNR) conditions. Results show that the proposed method is able to track multiple speakers in a real room environment.

This paper is organized as follows. Section 2 describes the phase difference likelihood measurement function and tracking of multiple sources using PF. Source detection and track management is discussed in Section 3. The MOT metrics are summarized in Section 4 after which the recordings are described in Section 5. Results are presented in Section 6. Section 7 concludes the discussion.
2. SOUND SOURCE LIKELIHOOD MODEL

A sound source at 3D Cartesian coordinates \( s = [s_x, s_y, s_z]^T \) emits a signal \( s(t) \), which is received by microphone \( k \) located at \( m_k \). The signal is convolved with the room impulse response \( (h_{s,m_k}(t)) \) between source and microphone:

\[
x_k(t) = s(t) * h_{s,m_k}(t) + n_k(t),
\]

where "\(*\)" denotes linear convolution, \( t \) is time, and \( n_k(t) \) is i.i.d noise, uncorrelated with the source and other microphone signals.

The phase difference can be measured from the input signals between microphone pair index \( p = \{k, j\} \), where \( p = 1, \ldots, P \) as:

\[
\phi_{i,p}(\omega) = \angle X_k(\omega, t) - \angle X_j(\omega, t),
\]

where \( X_k(\omega, t) \) denotes input signal \( k \) at angular frequency \( \omega \), and operator "\( \angle \)" denotes argument of a complex number. The theoretical value of a source located at \( s \) is obtained from:

\[
x_{s}^{(p)}(\omega) = \omega \cdot \frac{1}{c} \cdot \frac{||s - m_k|| - ||s - m_j||}{||s - m_k - s + m_k||}, \quad \text{for} \quad |\omega| = \text{Euclidean distance}. \quad (3)
\]

A function that normalizes the phase angle difference between values \([-1, 1]\) results the parameterized observation

\[
\vartheta_{(t)}^{(p)} = \sin\left(0.5\left(\phi_{i,p}(\omega) - \phi_{s}^{(p)}(\omega)\right)\right),
\]

where \( \vartheta_{(t)}^{(p)} \) is the basis of phase domain metric. In order to obtain a measurement likelihood, the phase differences are assumed to be normally distributed. A likelihood function for phase difference at frequency \( \omega \) at \( s \) is written as:

\[
p(\vartheta_{(t)}^{(p)}|s, \omega, \sigma_{\vartheta}^{2}) = \frac{1}{\sqrt{2\pi}\sigma_{\vartheta}^{2}} \exp\left(-\frac{1}{2\sigma_{\vartheta}^{2}}(\vartheta_{(t)}^{(p)} - \vartheta_{s}^{(p)})^{2}\right) \quad (5)
\]

where \( \sigma_{\vartheta}^{2} \) is phase difference variance.

Let \( x^t \) represent the state of source \( t \) at time \( t \), consisting of hidden state variables including source position. Phase difference measurements at \( N_f \) frequencies \( \Omega = [\omega_1, \omega_2, \ldots, \omega_{N_f}] \) is written in a vector form \( x_{(t)}^{(p)} = \begin{bmatrix} \vartheta_{(t)}^{(p)} \end{bmatrix} \). The set of measurements at time \( t \) from all \( P \) microphone pairs is denoted as \( Z(t) = \{x_{(t)}^{(p)}|p = 1, \ldots, P\} \). By assuming independence Eq. (5) can be written for phase values in range \( \Omega \) for \( P \) separate microphone pairs:

\[
p(Z(t)|x^t) \propto \exp(-\frac{1}{2} \sum_{p=1}^{P} \frac{1}{\sigma_{\vartheta}^{2}} \begin{bmatrix} x_{(t)}^{(p)} \end{bmatrix}^T \Sigma_{(t,p)}^{-1} \begin{bmatrix} x_{(t)}^{(p)} \end{bmatrix}),
\]

where \( \Sigma_{(t,p)} \) is microphone pair specific covariance matrix of phase error at time \( t \).

2.1. Particle Filter for Multiple Target Tracking

This section reviews the particle filter (PF) based sequential Monte Carlo filtering for multiple target tracking. See [11] for a tutorial on PF.

Let \( X^t \) represent the state at time \( t \), and the state consists of single object states \( X^{t_{(i)}} = \{x_i^t\}_{i=1}^S \). The set of all measurements up to and including time \( t \) is denoted as \( Z^t \). The following non-linear discrete time system is considered

\[
x_{i}^{t+1} = f(x_i^t) + v_t
\]

\[
z_t = g(x_i^t) + w_t,
\]

where \( f(\cdot) \) and \( g(\cdot) \) are the possibly non-linear state transition function and measurement function (3), respectively, and \( v_t \) and \( w_t \) are uncorrelated noise processes.

The new state is predicted using the state transition model

\[
p(x_i^t|Z^{t-1}) = \int p(x_i^{t}|x_i^{t-1})p(x_i^{t-1}|Z^{t-1})dx_i^{t-1}.
\]

A new set of measurements is then used to update the prediction

\[
p(x_i^t|Z^t) \propto p(x_i^t|Z^t)p(x_i^t|Z^{t-1}).
\]

The PF approximates the posterior density with a set of \( N \) particles

\[
p(x_i^t|Z^t) \approx \frac{1}{N} \sum_{i=1}^{N} w_i^{(t)} \delta(x_i^t - x_i^{(i)}),
\]

where \( \delta(\cdot) \) is the Dirac delta function, \( x_i^{(i)} \) is the \( i \)th particle state value, and \( w_i^{(t)} \) is the corresponding particle weight obtained from the likelihood function (6). The joint state probability density function \( p(X^{t} | Z^{t}) \) is then the collection of the \( S \) individual state probability densities \( p(X^{t_{(i)}} | Z^{t}) \approx \{x_i^{(i)}, w_i^{(t)}\}_{i=1}^{S} \). The proposed method uses the sampling importance resampling (SIR) to resample particles in each time step [11]. The state transition (7) is set to Brownian movement: \( x_i^{t+1} = x_i^t + \nu_t \), where the covariance of \( \nu_t \) is \( \Sigma_{\nu}^\nu \), the median coordinate value of a filter's particle locations is taken as the source point estimate \( \hat{a}_i \).

3. DETECTION AND TRACK MANAGEMENT

Two types of particle filter designs are used: an initialization filter and a tracking filter. The initialization filter is used to i) detect new sources and ii) measure background level. On average 30 % of initialization filter's particles are selected to subset \( R \). The particles are denoted \( \{x^{(i)}, w^{(i)}\}_{i \in \mathbb{R}} \) and they are randomly distributed into the state space to search for new targets. The non-random particles belong to subset \( T \). If a new source is found, the initialization filter will converge to the target due to resampling.

The detection is implemented here with the likelihood ratio (LR) test. Given two hypotheses \( H_0: \) "No target present" and \( H_1: \) "Target present", the logarithm of LR is given

\[
\lambda(t) = \log \left(p(Z(t)|H_1)/p(Z(t)|H_0)\right),
\]

and the existence of the target can be agreed if the LR exceeds a threshold value \( \lambda(t) \geq \lambda_t \). It is shown in [4] that the likelihood of source existence can be extracted from the particle filter weights. In [4] the background model \( p(Z(t)|H_0) \) is assumed to be known, and does not need to be estimated. However, in the ASL the background model is generally unknown. In [12] the background likelihood of a video image was proposed to be measured by considering a set of particles evenly distributed in the search space. To avoid numerical instability both hypotheses are extracted from logarithmic particle weights \( \tilde{w}^{(t)} \):

\[
p(Z(t)|H_1) \approx \frac{1}{|R|} \sum_{i \in \mathbb{R}} \tilde{w}^{(t)}(i),
\]

\[
p(Z(t)|H_0) \approx \frac{1}{|T|} \sum_{i \in \mathbb{R}} w^{(t)}(i).
\]

Since LR (12) can fluctuate rapidly between sequential frames a low-pass filter is used [12] with \( \gamma \in [0, 1] \) as a smoothing factor:

\[
\tilde{\lambda}(t) = \gamma \cdot \tilde{\lambda}(t-1) + (1 - \gamma) \cdot \lambda(t).
\]
Fig. 1. Illustration of recording room layout. Microphones are integrated into a plane array of a Y-shaped geometry and the center microphone coordinates are given as array coordinates. Arrays are mounted on a wooden table. Loudspeakers surround the table.

The initialization filter can be used to evaluate the likelihood of $\mathcal{H}_0$ since particles in subset $\mathcal{R}$ sample the background. This avoids a grid based search of new targets and is additionally scalable into large spaces without extra computational cost. The likelihood for $\mathcal{H}_i$ is obtained from the non-random particle subset $\mathcal{T}$ (14). A track is added if the initialization filter’s smoothed LR (15) exceeds a pre-selected threshold value $T_b$. Then, the particle states are copied into a new tracking filter, and the initialization filter starts the search for a new target. A tracking filter’s particle weights are used to evaluate the likelihood $\mathcal{H}_i$, and the initialization filter’s value for likelihood $\mathcal{H}_0$ is used to obtain the LR of the tracking filter. A tracking filter is removed if the smoothed LR falls below a threshold value $T_b$.

Two tracking filters are merged if their corresponding point estimate source positions become closer than $T_m$ apart. The initialization filter avoids existing tracks by de-emphasizing its particle likelihood using generalized logistic function $\left(\alpha \left| \mathbf{s}_i \right| - \beta \right)_{+}^{-1}$, where $\mathbf{s}_i$ is the initialization filter’s $i$th particle position, $\mathbf{s}_i$ ($i > 0$) is the point estimate position of closest tracking filter, $\alpha = 4.2$, and $\beta = 7.2$.

4. MULTIPLE OBJECT TRACKING METRICS

A set of metrics have been proposed for evaluation of MOT performance (9). During a time frame $t$, several sources can be simultaneously active and their joint locations are referred as ground truth ($G_t$). The tracker output is referred as hypothesis which consists of a set of (source) coordinates at each time frame. A hypothesis point corresponds to a ground truth point if their Euclidean distance $(d_{ht})$ is smaller than match threshold ($S_h$), and $t$ denotes match index. The amount of correct unique matches at time $t$ is denoted as $c_t$. A false positive ($F_t$) hypothesis does not correspond to any ground truth point, and a miss ($M_t$) is a ground truth point without a corresponding hypothesis. Correspondence between ground truth and hypothesis points is made using the Hungarian algorithm, see [9] for details. Summary metrics are false positive ratio $F$, miss ratio $M$, and the multiple object tracking accuracy (MOTA) [9]:

$$\bar{F} = \frac{\sum_{t} F_t}{\sum_{t} G_t}, \quad \bar{M} = \frac{\sum_{t} M_t}{\sum_{t} G_t}, \quad \text{MOTA} = 1 - \frac{\sum_{t} (M_t + F_t)}{\sum_{t} G_t}$$

(16)

The multiple object tracking precision (MOTP) is defined as [9]

$$\text{MOTP} = \frac{\sum_{t} d_{ht}(\sum_{t} c_t)^{-1} \text{[mm]}}{\sum_{t} c_t}$$

(17)

5. REAL DATA EVALUATION

To evaluate the performance of the proposed MTT method, a set of real data recordings was performed. Four Genelec 1029A active loudspeakers were used to play back male and female speech samples from the TIMIT database. The data was gathered at 48000 Hz with 32 bits per sample precision. The room dimensions are $4.53 \times 3.96 \times 2.59$ m and the reverberation time ($T_{60}$) is approximately 0.26 s. The room consists of a table, sofas, and small equipment. The microphones were placed on the table in three arrays of Y-shaped geometry. DPA 4060-BM pre-polarized omnidirectional miniature condenser microphones were used with a 48 V phantom feed. Refer to Fig. 1 for details.

The recordings consisted of six sequences. In each sequence the talker gender, active loudspeaker index, and percentage of speech overlap was varied. Sequences 1, 2 and 5 correspond to scenarios where single speakers take turns. Sequences 3 and 4 contain fully overlapping speech from two speakers, and sequence 6 contains partially overlapping speech between two speakers. See Table 1 for details. The A-weighted sound pressure level (SPL) of white noise at 30 cm distance from each speaker was first calibrated to 80 dB. The sequences 1–6 were then played back. A mixer was then used to lower the SPL 21 dB, and then further 11 dB. These correspond to SNR A, B, and C conditions, respectively. The background noise consisted mostly of air conditioning and internal microphone noise. Figure 2 displays recorded sequence 5 in the different SNR levels, where in case A the speech signal dominates the background, in SNR B the signal is already at the level of background, and SNR C is weak.

6. RESULTS

Table 2 contains the used parameter values. Note the separate amount of particles in the initialization and tracking filters. All

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Gender</th>
<th>Overlap</th>
<th>Speakers</th>
<th>Length</th>
<th>Speaker</th>
<th>Turns</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>M</td>
<td>0%</td>
<td>1-4-1</td>
<td>24.1 s</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>M</td>
<td>0%</td>
<td>1-4-1</td>
<td>27.5 s</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>M+F</td>
<td>100%</td>
<td>1-2-1</td>
<td>21.5 s</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>M+F</td>
<td>100%</td>
<td>3-4-1</td>
<td>21.3 s</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>M</td>
<td>0%</td>
<td>1-4-1</td>
<td>19.5 s</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>M+F</td>
<td>33%</td>
<td>1-4-1</td>
<td>15 s</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

1 Add and delete threshold values are set equal for simplicity.
2 Mismatch errors are omitted. A mismatch happens if a correspondence of single track is made with more than one ground truth objects over time.
Fig. 3. Results of multiple target tracking over 40 repetitions. Mean values together with standard deviation (vertical bars, ±σ/2).

scores are calculated from 2D (x,y) coordinate values. Source existence is assumed for the whole speech file and scoring is performed at 21 ms intervals (output rate of tracker). Thus, any embedded silence in the waveform’s end or beginning lowers the overall score. Only microphone pairs inside each array were used and inter-array pairs were excluded. In addition, the phase differences were pre-calculated at accuracy corresponding to half a sample to speed up calculations.

Figure 3 displays the average results of 40 repetitions of the method for each sequence in each SNR. The panels of Fig. 3 from left to right represent the MOTA score, miss percent, false positive ratio (16) with S = 50 cm. The rightmost panel shows the effect of using different scoring thresholds S = 30, 40, . . . , 70 cm. The sequence number is represented on the ordinate, the ordinate is the score value. The MOTP (17) over all sequences was 170 mm, 183 mm, and 214 mm for SNR conditions A, B, and C, respectively.

The miss percent behavior is identical between the three SNR conditions, and is nearly zero for the single speaker sequences 1, 2, and 5 in SNR A and B. This indicates that the system accurately tracks switching talkers. False positive ratio is higher in the SNR case A compared to other SNR levels, which means that tracks are not deleted fast enough or false targets are generated. The detection threshold (Tth) can be used to control the tradeoff between misses and false positives. The rightmost panel of Fig. 3 shows the MOTA score (SNR B) using different range match thresholds. When S is increased, the MOTA for sequence 4 improves most, since inaccurate estimates are accepted due to more loose matching distance. In contrast, the MOTA for sequence 3 is not increased after S = 40 cm, which indicates that this source to sensor geometry is more amenable for localization. It is noted, that better geometries could be utilized over the used ad hoc geometry.

7. CONCLUSIONS

A particle filtering based multiple acoustic source tracking method using track before detect paradigm was presented. The method utilizes phase differences between microphone pairs, and can be used in any frequency range for multiple microphones. Two particle filter designs were used to detect and track sources and to estimate the source existence using a likelihood ratio test. The method was demonstrated to successfully track two switching and overlapping speakers in a real room in various SNR conditions.

8. REFERENCES