Hindawi Publishing Corporation
EURASIP Journal on Audio, Speech, and Music Processing
Volume 2008, Article ID 278185, 14 pages
doi:10.1155/2008/278185

Research Article

Measurement Combination for Acoustic Source Localization in a Room Environment

Pasi Pertilä, Teemu Korhonen, and Ari Visa

Department of Signal Processing, Tampere University of Technology, P.O. Box 553, 33101 Tampere, Finland

Correspondence should be addressed to Pasi Pertilä, pasi.pertila@tut.fi

Received 31 October 2007; Revised 4 February 2008; Accepted 23 March 2008

Recommended by Woon-Seng Gan

The behavior of time delay estimation (TDE) is well understood and therefore attractive to apply in acoustic source localization (ASL). A time delay between microphones maps into a hyperbola. Furthermore, the likelihoods for different time delays are mapped into a set of weighted nonoverlapping hyperbolae in the spatial domain. Combining TDE functions from several microphone pairs results in a spatial likelihood function (SLF) which is a combination of sets of weighted hyperbolae. Traditionally, the maximum SLF point is considered as the source location but is corrupted by reverberation and noise. Particle filters utilize past source information to improve localization performance in such environments. However, uncertainty exists on how to combine the TDE functions. Results from simulated dialogues in various conditions favor TDE combination using intersection-based methods over union. The real-data dialogue results agree with the simulations, showing a 45% RMSE reduction when choosing the intersection over union of TDE functions.

1. INTRODUCTION

Passive acoustic source localization (ASL) methods are attractive for surveillance applications, which are a constant topic of interest. Another popular application is human interaction analysis in *smart rooms* with multimodal sensors. Automating the perception of human activities is a popular research topic also approached from the aspect of localization. Large databases of smart room recordings are available for system evaluations and development [1]. A typical ASL system consists of several spatially separated microphones. The ASL output is either source direction or location in two- or three-dimensional space, which is achieved by utilizing received signal phase information [2] and/or amplitude [3], and possibly sequential information through tracking [4].

Traditional localization methods maximize a spatial likelihood function (SLF) [5] to locate the source. Localization methods can be divided according to the way the spatial likelihood is formed at each time step. The steered beamforming approach sums delayed microphone signals and calculates the output power for a hypothetical location. It is therefore a direct localization method, since microphone signals are directly applied to build the SLF.

Time delay estimation (TDE) is widely studied and well understood and therefore attractive to apply in the source localization problem. The behavior of correlation-based TDE methods has been studied theoretically [6] also in reverberant enclosures [7, 8]. Other TDE approaches include determining adaptively the transfer function between microphone channels [9], or the impulse responses between the source and receivers [10]. For more discussion on TDE methods, see [11].

TDE-based localization methods first transform microphone pair signals into a time delay likelihood function. These pairwise likelihood functions are then combined to construct the spatial likelihood function. It is therefore a two-step localization approach in comparison to the direct approach. The TDE function provides a likelihood for any time delay value. For this purpose, the correlation-based TDE methods are directly applicable. A hypothetical source position maps into a time delay between a microphone pair. Since the TDE function assigns a likelihood for the time delay, the likelihood for the hypothetical source position is obtained. From a geometrical aspect, time delay is inverse-mapped

as a hyperbola in 3D space. Therefore, the TDE function corresponds to a set of weighted nonoverlapping hyperbolae in the spatial domain. The source location can be solved by utilizing spatially separated microphone pairs, that is, combining pairwise TDE functions to construct a spatial likelihood function (SLF). The combination method varies. Summation is used in [12–14], multiplication is used in [15, 16], and the determinant, used originally to determine the time delay from multiple microphones in [17], can also be applied for TDE function combination in localization. The traditional localization methods consider the maximum point of the most recent SLF as the source location estimate. However, in a reverberant and noisy environment, the SLF can have peaks outside the source position. Even a moderate increase in the reverberation time may cause dominant noise peaks [7], leading to the failure of the traditional localization approach [15]. Recently, particle filtering (PF)-based sound source localization systems have been presented [13, 15, 16, 18]. This scheme uses information also from the past time frames to estimate the current source location. The key idea is that spatially inconsistent dominant noise peaks in the current SLF do not necessarily corrupt the location estimate. This scheme has been shown to extend the conditions in which an ASL system is usable in terms of signal to noise ratio (SNR) and reverberation time (T60) compared to the traditional approach [15].

As noted, several ways of combination TDE functions have been used in the past, and some uncertainty exists about a suitable method for building the SLF for sequential 3D source localization. To address this issue, this work introduces a generalized framework for combining TDE functions in TDE-based localization using particle filtering. Geometrically, the summation of TDE functions represents the union of pairwise spatial likelihoods, that is, union of the sets of weighted hyperbolae. Such SLF does have the maximum value at the correct location but also includes the unnecessary tails of the hyperbolae. Taking the intersection of the sets reduces the unnecessary tails of the hyperbolae, that is, acknowledges that the time delay is eventually related only to a single point in space and not to the entire set of points it gets mapped into (hyperbola). TDE combination schemes are compared using a simulated dialogue. The simulation reverberation time (T60) ranges from 0 to 0.9 second, and the SNR ranges from -10 to $+30$ dB. Also real-data from a dialogue session is examined in detail.

The rest of this article is organized as follows: Section 2 discusses the signal model and TDE functions along with signal parameters that affect TDE. Section 3 proposes a general framework for combining the TDE functions to build the SLF. Section 4 categorizes localization methods based on the TDE combination operation they apply and discusses how the combination affects the SLF shape. Iterative localization methods are briefly discussed. Particle filtering theory is reviewed in Section 5 for sequential SLF estimation and localization. In Section 6, simulations and real-data measurements are described. Selected localization methods are compared in Section 7. Finally, Sections 8 and 9 conclude the discussion.

2. SIGNAL MODEL AND TDE FUNCTION

The sound signal emitted from a source is propagated into the receiving microphone. The received signal is a convolution of source signal and an impulse response. The impulse response encompasses the measurement equipment response, room geometry, materials as well as the propagation delay from a source $\mathbf{r}^n$ to a microphone $\mathbf{m}_i$ and reverberation effects. The ith microphone signal is a superposition of convoluted source signals [14, 15]:

$$x_i(t) = \sum_{n=1}^{N} s_n(t) * h_{i,n}(t) + w_i(t), \tag{1}$$

where $i \in [1, \ldots, M]$, and $s_n(t)$ is the signal emitted by the nth source, $n \in [1, \ldots, N]$, $w_i(t)$ is assumed here to be independent and identically distributed noise, t represents discrete time index, $h_{i,n}(t)$ is the impulse response, and $*$ denotes convolution. The propagation time from a source point $\mathbf{r}^n$ to microphone i is

$$\tau_{i,\mathbf{r}^n} = ||\mathbf{r}^n - \mathbf{m}_i|| \cdot c^{-1}, \tag{2}$$

where c is the speed of sound, and $\|\cdot\|$ is the Euclidean norm. Figure 1(a) illustrates propagation delay from source to microphones, using a 2D simplification.

A wavefront emitted from point $\mathbf{r}$ arrives at spatially separated microphones i, j according to their corresponding distance from point $\mathbf{r}$. This time difference of arrival (TDOA) value between the pair $p = \{i, j\}$ in samples is [14]

$$\Delta\tau_{p,\mathbf{r}} = \lceil (||\mathbf{r} - \mathbf{m}_i|| - ||\mathbf{r} - \mathbf{m}_j||) \cdot f_s \cdot c^{-1} \rfloor, \tag{3}$$

where f_s is the sampling frequency, and $\lceil \cdot \rfloor$ denotes rounding. Conversely, a delay between microphone pair $\Delta\tau_{p,\mathbf{r}}$ defines a set of 3D locations $\mathcal{H}_{p,\mathbf{r}}$ forming a hyperbolic surface that includes the unique location $\mathbf{r}$. The geometry is illustrated in Figure 1(b), where hyperbolae related to TDOA values $-30, -20, \ldots, 30$ are illustrated.

In this work, a TDE function between microphone pair p is defined $\mathcal{R}_p(\tau_p) \in [0, 1]$, where the delay can have values:

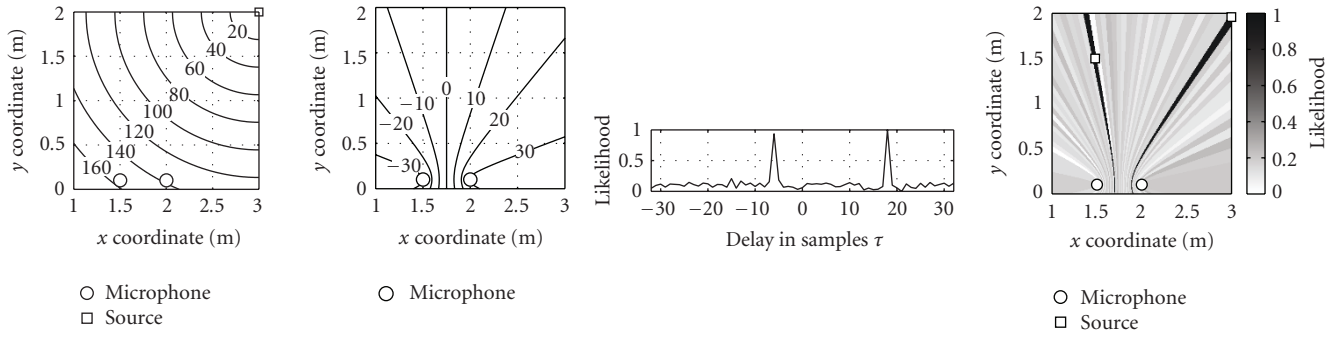
$$\tau_p \in [-\tau_{\max}, \tau_{\max}], \quad \tau_p \in \mathbb{Z}, \tag{4}$$

$$\tau_{\max} = \lceil ||\mathbf{m}_j - \mathbf{m}_i|| \cdot f_s \cdot c^{-1} \rfloor. \tag{5}$$

The unit of delay is one sample. TDE functions include the generalized cross correlation (GCC) [19] which is defined for a frame of microphone pair p data:

$$\mathcal{R}_p^{\mathrm{GCC}}(\tau_p) = \mathcal{F}^{-1}\{W_p(k) X_i(k) X_j(k)^*\}, \tag{6}$$

where $X_j(k)^*$ is a complex conjugate transpose of the DFT of the jth microphone signal, k is discrete frequency, $\mathcal{F}^{-1}\{\cdot\}$ denotes inverse DFT, and $W_p(k)$ is a weighting function, see [19]. Phase transform (PHAT) weighting $W_p(k) = |X_i(k)X_j(k)^*|^{-1}$ causes sharper peaks in the TDE function compared to the nonweighted GCC and is used by several TDE-based localization methods, including the steered response power using phase transform (SRP-PHAT) [14].



(a) Propagation delay from source 1

(b) TDOA values and corresponding hyperbolae

(c) TDE function values, $\mathcal{R}_p(\tau_p)$

(d) Spatial likelihood function (SLF) for a microphone pair

FIGURE 1: Source localization geometry is presented. The sampling frequency is 22050 Hz, the speed of sound is 343 m/s, the source signal is colored noise, and SNR is +24 dB. The sources are located at $\mathbf{r}^1 = \langle 3, 2\rangle$ and $\mathbf{r}^2 = \langle 1.5, 1.5\rangle$ or at TDOA values $\Delta\tau_1 = 18$ and $\Delta\tau_2 = -6$. In panel (a), the propagation time from source at $\mathbf{r}^1$ is different for the two microphones (values given in samples). This difference is the TDOA value of the source. Panel (b) illustrates how different TDOA values are mapped into hyperbolae. In panel (c), the two peaks at locations $\tau_p = 18$ and $\tau_p = -6$ in the TDE function correspond to the source locations $\mathbf{r}^1$ and $\mathbf{r}^2$, respectively. Panel (d) displays the TDE function values from panel (c) mapped into a microphone pairwise spatial likelihood function (SLF).

An example of TDE function is displayed in Figure 1(c). Other weighting schemes include the Roth, Scot, Eckart, the Hannan-Thomson (maximum likelihood) [19], and the Hassab-Boucher methods [20].

Other applicable TDE functions include the modified average magnitude difference function (MAMDF) [21]. Recently, time frequency histograms have been proposed to increase TDE robustness against noise [22]. For a more detailed discussion on TDE refer to [11]. The evaluation of different TDE methods and GCC weighting methods is, however, outside the scope of this work. Hereafter, the PHAT-weighted GCC is utilized as the TDE weighting function since it is the optimal weighting function for a TDOA estimator in a reverberant environment [8].

The correlation-based TDOA is defined as the peak location of the GCC-based TDE function [19]. Three distinct SNR ranges (high, low, and the transition range in between) in TDOA estimation accuracy have been identified in a nonreverberant environment [6]. In the high SNR range, the TDOA variance attains the Cramer-Rao lower bound (CRLB) [6]. In the low SNR range, the TDE function is dominated by noise, and the peak location is noninformative. In the transition range, the TDE peak becomes ambiguous and is not necessary related to the correct TDOA value. TDOA estimators fail rapidly when the SNR drops into this transition SNR range [6]. According to the modified Ziv-Zakai lower bound, this behavior depends on time-bandwidth product, bandwidth to center frequency ratio, and SNR [6]. In addition, the CRLB depends on the center frequency.

In a reverberant environment the correlation-based TDOA performance is known to rapidly decay when the reverberation time (T60) increases [7]. The CRLB of the correlation-based TDOA estimator in the reverberant case is derived in [8] where PHAT weighting is shown to be optimal. In that model, the signal to noise and reverberation ratio (SNRR) and signal frequency band affect the achievable minimum variance. The SNRR is a function of the acoustic reflection coefficient, noise variance, microphone distance from the source, and the room surface area.

3. FRAMEWORK FOR BUILDING THE SPATIAL LIKELIHOOD FUNCTION

Selecting a spatial coordinate $\mathbf{r}$ assigns a microphone pair p with a TDOA value $\Delta\tau_{p,\mathbf{r}}$ as defined in (3). The TDE function (6) indexed with this value, that is, $\mathcal{R}_p(\Delta\tau_{p,\mathbf{r}})$, represents the likelihood of the source existing at the locations that are specified by the TDOA value, that is, hyperboloid $\mathcal{H}_{p,\mathbf{r}}$. The pairwise SLF can be written as

$$P(\mathcal{R}_p \mid \mathbf{r}) = \mathcal{R}_p(\Delta\tau_{p,\mathbf{r}}) \in [0,1], \tag{7}$$

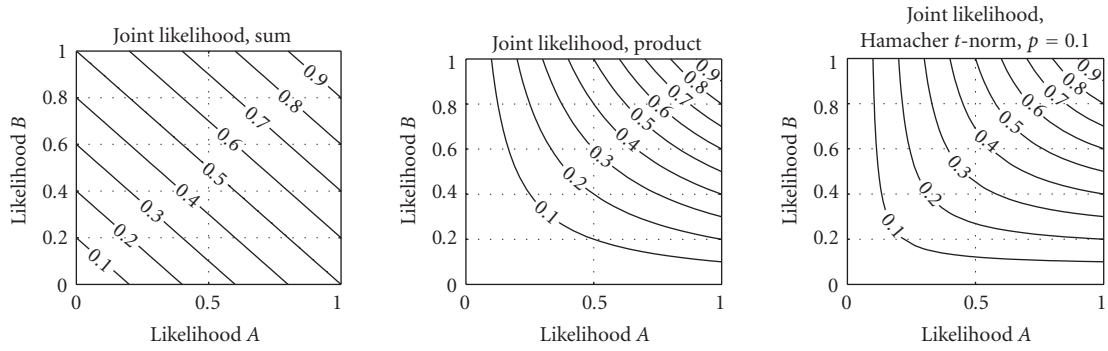
where $P(\cdot \mid \cdot)$ represents conditional likelihood, normalized between $[0,1]$. Figure 1(d) displays the pairwise SLF of the TDE measurement displayed in Figure 1(c). Equation (7) can be interpreted as a likelihood of a source having location $\mathbf{r}$ given the measurement $\mathcal{R}_p$.

The pairwise SLF consists of weighted nonoverlapping hyperbolic objects and therefore has no unique maximum. A practical solution to reduce the ambiguity of the maximum point is to utilize several microphone pairs. The combination operator used to perform fusion between these pairwise SLFs influences the shape of the resulting SLF. Everything else except the source position of each of the hyperboloid's shape is nuisance.

A binary operator combining two likelihoods can be defined as

$$\otimes : [0,1] \times [0,1] \longrightarrow [0,1]. \tag{8}$$

Among such operators, ones that are commutative, monotonic, associative, and bounded between $[0,1]$ are of interest



(a) Sum $0.5(A + B)$ (b) Product AB (c) Hamacher t-norm

FIGURE 2: Three common likelihood combination operators, normalized sum (s-norm), product (t-norm), and Hamacher t-norm are illustrated along their resulting likelihoods. The contour lines represent constant values of output likelihood.

here. For likelihoods A, B, C, D, these rules are written as

$$A \otimes B = B \otimes A, \tag{9}$$

$$A \otimes B \leq C \otimes D, \quad \text{if } A \leq C \text{ and } B \leq D, \tag{10}$$

$$A \otimes (B \otimes C) = (A \otimes B) \otimes C. \tag{11}$$

Such operations include t-norm and s-norm. s-norm operations between two sets represent the union of sets and have the property $A \otimes 0 = A$. The most common s-norm operation is summation. Other well- known s-norm operations include the Euclidean distance and maximum value.

A t-norm operation represents the intersection of sets and satisfies the property $A \otimes 1 = A$. Multiplication is the most common such operation. Other t-norm operations include the minimum value and Hamacher t-norm [23] which is a parameterized norm and is written for two values A and B:

$$h(A, B, \gamma) = \frac{AB}{\gamma + (1 - \gamma)(A + B - AB)}, \tag{12}$$

where $\gamma > 0$ is a parameter. Note that the multiplication is a special case of (12) when $\gamma = 1$.

Figure 2 illustrates the combination of two likelihood values, A and B. The likelihood values are displayed on the axes. The leftmost image represents summation, the middle represents product and the rightmost is Hamacher t-norm ($\gamma = 0.1$). The contour lines represent the joint likelihood. The summation is the only s-norm here. In general, the t-norm is large only if all likelihoods are large. Similarly, the s-norm can be large even if some likelihood values are small.

The combination of pairwise SLFs can be written: (using $\otimes$ with prefix notation.)

$$P(\boldsymbol{\mathcal{R}} \mid \mathbf{r}) = \bigotimes_{p \in \Omega} \mathcal{R}_p(\Delta\tau_{p,\mathbf{r}}), \tag{13}$$

where each microphone pair p belongs to a microphone pair group Ω, and $\boldsymbol{\mathcal{R}}$ represents all the TDE functions of the group. There exists $\binom{M}{2}$ unique microphone pairs in the set of all pairs. Sometimes partitioning the set of microphones into groups or *arrays* before pairing is justified. The signal coherence between two microphones decreases as microphone distance increases [24] which favors partitioning the microphones into groups with low sensor distance. Also, the complexity of calculating all pairwise TDE function values is $\mathcal{O}(M^2)$, which is lower for partitioned arrays. Selecting too small sensor separation may lead to over-quantization of the possible TDOA values where only a few delay values exist, see (5).

4. TDE-BASED LOCALIZATION METHODS

Several TDE-based combination schemes exist in the ASL literature. The most common method is the summation. This section presents four distinct operations in the generalized framework.

4.1. Summation operator in TDE-based localization

The method in [12] sums GCC values, which is equivalent to the steered beamformer. The method in [13] sums precedence-weighted GCC values (for direction estimation). SRP-PHAT method sums PHAT-weighted GCC values [14]. All these methods use the summation operation which fulfills the requirements (9)–(11). Using (13), the SRP-PHAT is written as

$$P_{\text{SRP-PHAT}}(\boldsymbol{\mathcal{R}} \mid \mathbf{r}) = \sum_{p \in \Omega} \mathcal{R}_p^{\text{GCC-PHAT}}(\Delta\tau_{p,\mathbf{r}}). \tag{14}$$

Every high value of the pairwise SLF is present in the resulting SLF since the sum represents a union of values. In a multiple source situation with more than two sensors, this approach generates high probability regions outside actual source positions, that is, ghosts. See Figure 3(a) for illustration, where ghosts appear, for example, at x, y coordinates $\langle 3.1, 1.2\rangle$ and $\langle 2.6, 1.3\rangle$.

4.2. Multiplication operator in TDE-based localization

In [15, 16], product was used as the likelihood combination operator which is a probabilistic approach. (In [15] negative

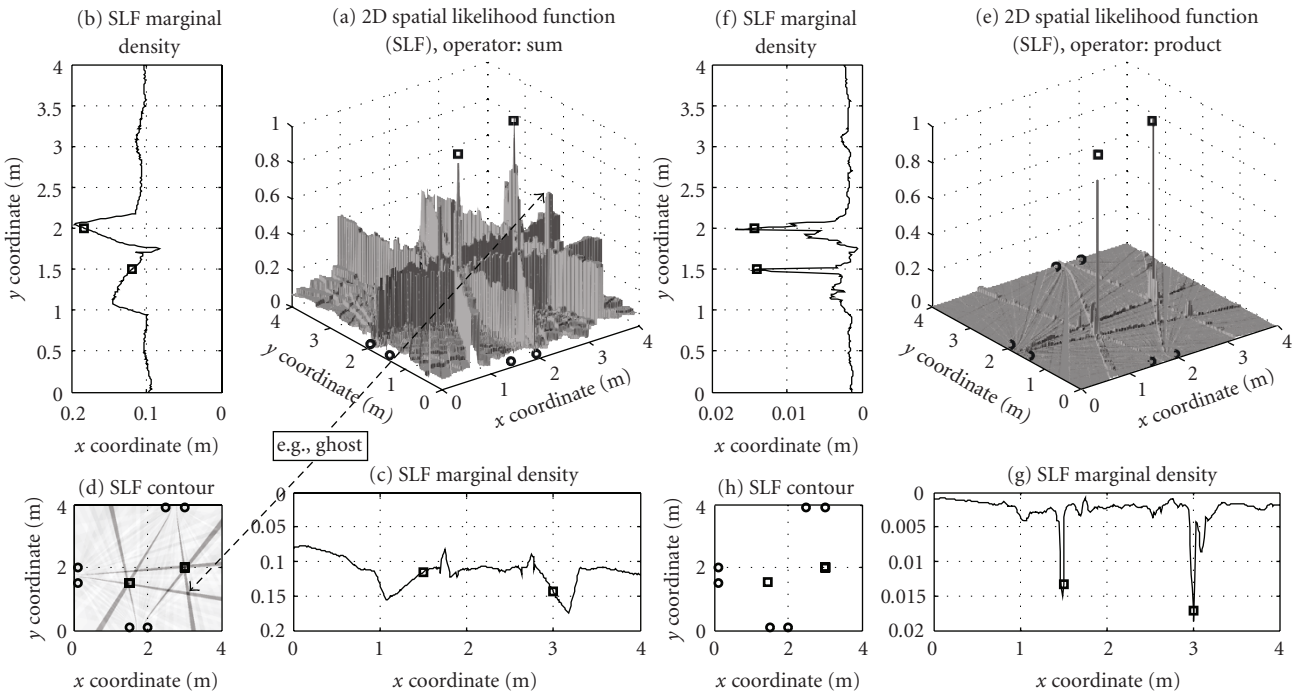


FIGURE 3: A two-source example scenario with three microphone pairs is illustrated. The source coordinates are $\mathbf{r}^1 = \langle 3, 2\rangle$ and $\mathbf{r}^2 = \langle 1.5, 1.5\rangle$. Two combination operators *sum* and *product* are used to produce two separate spatial likelihood functions (SLFs). The SLF contours are presented in panels (d) and (h). Circle and square represent microphone and source locations, respectively. Panels (a) and (e) illustrate the resulting 2D SLF, produced with the sum and product operations, respectively. The marginal distributions of the SLFs are presented in panels (b) and (c) for the sum, and (f) and (g) for the product. The panel (a) distribution has *ghosts* which are the result of summed observations, see example ghost at $\langle 3.1, 1.2\rangle$. Also, the marginal distributions are not informative. In the panel (e), SLF has sharp peaks which are in the presence of the actual sound sources. The marginal distributions carry source position information, though this is not guaranteed in general.

GCC values are clipped and the resulting positive values are raised to power q) If the likelihoods are independent, the intersection of sets equals their product. The method, termed here multi-PHAT, multiplies the pairwise PHAT-weighted GCC values together in contrast to summation. The multi-PHAT fulfills (9)–(11) and is written using (13)

$$P_{\text{multi-PHAT}}(\boldsymbol{\mathcal{R}} \mid \mathbf{r}) = \prod_{p\in\Omega} \mathcal{R}_p^{\text{GCC-PHAT}}(\Delta\tau_{p,\mathbf{r}}). \tag{15}$$

This approach outputs the common high likelihood areas of the measurements, and so the unnecessary peaks of the SLF are somewhat reduced. The ghosts experienced in the SRP-PHAT method are eliminated in theory by the intersection-based combination approach. This is illustrated in Figure 3(b). The SLF has two distinct peaks that correspond to the true source locations.

4.3. Hamacher t-norm in TDE-based localization

Several other methods that have the properties (9)–(11) can be used to combine likelihoods. These methods include parameterized t-norms and s-norms [23]. Here, the Hamacher t-norm (12) is chosen because it is relatively close to the product and represents the intersection of sets. The Hamacher t-norm is defined as a dual norm, since it operates on two inputs.

The parameter $\gamma > 0$ in the Hamacher t-norm (12) defines how the norm behaves. For example, $h(0.5, 0.2, 0.1) \approx 0.16$ whereas their product equals $0.2\cdot 0.5 = 0.1$, and $h(0.5, 0.2, 15) \approx 0.085$. Figures 2(b) and 2(c) represent the multiplication and Hamacher t-norm ($\gamma = 0.1$). The Hamacher t-norm-based TDE localization method is written using (13):

$$\begin{aligned} &P_{\text{Hamacher-PHAT}}(\boldsymbol{\mathcal{R}} \mid \mathbf{r}, \gamma) \\ &\quad = h(\ldots h(\mathcal{R}_1(\Delta\tau_{\mathbf{r}}), \mathcal{R}_2(\Delta\tau_{\mathbf{r}}), \gamma), \ldots, \mathcal{R}_J(\Delta\tau_{\mathbf{r}}), \gamma), \end{aligned} \tag{16}$$

where $\mathcal{R}_J(\Delta\tau_{\mathbf{r}})$ is abbreviated notation of $\mathcal{R}_J^{\text{GCC}-\text{PHAT}}(\Delta\tau_{J,\mathbf{r}})$, that is, the PHAT-weighted GCC value from the Jth microphone pair for location $\mathbf{r}$, where J is the total number of pairs, and $h(\cdot, \cdot, \gamma)$ is the Hamacher t-norm (12). Since the norm is commutative, the TDE measurements can be combined in an arbitrary order. Any positive γ value can be chosen, but values $\gamma < 1$ were empirically found to produce good results.

Note that multi-PHAT is a special case of Hamacher-PHAT when $\gamma = 1$.

4.4. Other combination methods in TDE-based localization

Recently, a spatial correlation-based method for TDOA estimation has been proposed [17], termed the multichannel cross correlation coefficient (MCCC) method. It combines cross correlation values for TDOA estimation and is considered here for localization. The correlation matrix from a M microphone array is here written:

$$\widetilde{\mathbf{R}} = \begin{bmatrix} \mathcal{R}_{1,1}(\Delta\tau_{\mathbf{r}}) & \mathcal{R}_{1,2}(\Delta\tau_{\mathbf{r}}) & \dots & \mathcal{R}_{1,M}(\Delta\tau_{\mathbf{r}}) \\ \mathcal{R}_{2,1}(\Delta\tau_{\mathbf{r}}) & \mathcal{R}_{2,2}(\Delta\tau_{\mathbf{r}}) & \dots & \mathcal{R}_{2,M}(\Delta\tau_{\mathbf{r}}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{R}_{M,1}(\Delta\tau_{\mathbf{r}}) & \mathcal{R}_{M,2}(\Delta\tau_{\mathbf{r}}) & \dots & \mathcal{R}_{M,M}(\Delta\tau_{\mathbf{r}}) \end{bmatrix}, \quad (17)$$

where $\mathcal{R}_{i,j}(\Delta\tau_{\mathbf{r}})$ equals $\mathcal{R}_p^{\text{GCC-PHAT}}(\Delta\tau_{p,\mathbf{r}})$. In [17], the matrix (17) is used for TDOA estimation, but here it is interpreted as a function of source position using (13)

$$P_{\text{MCCC}}(\boldsymbol{\mathcal{R}} \mid \mathbf{r}) = 1 - \det\widetilde{\mathbf{R}}. \quad (18)$$

The spatial likelihood of, for example, a three microphone array is

$$\begin{aligned} P_{\text{MCCC}}(\boldsymbol{\mathcal{R}} \mid \mathbf{r}) &= 1 - \det\widetilde{\mathbf{R}}_{3\times 3} \\ &= \mathcal{R}_{1,2}(\Delta\tau_{\mathbf{r}})^2 + \mathcal{R}_{1,3}(\Delta\tau_{\mathbf{r}})^2 + \mathcal{R}_{2,3}(\Delta\tau_{\mathbf{r}})^2 \\ &\quad - 2\mathcal{R}_{1,2}(\Delta\tau_{\mathbf{r}})\mathcal{R}_{1,3}(\Delta\tau_{\mathbf{r}})\mathcal{R}_{2,3}(\Delta\tau_{\mathbf{r}}). \end{aligned} \quad (19)$$

The MCCC method is argued to remove the effect of a channel that does not correlate with the other channels [17]. This method does not satisfy the monotonicity assumption (10). Also, the associativity (11) does not follow in arrays larger than three microphones.

4.5. Summary of the TDE combination methods

Four different TDE combination schemes were discussed, and existing localization methods were categorized accordingly. Figure 3 displays the difference between the intersection and the union of TDE function in localization. The SLF produced with the Hamacher t-norm differs slightly from the multiplication approach and is not illustrated. Also, the SLF produced with the MCCC is relatively close to the summation, as seen later in Figure 10. The intersection results in the source location information. The union contains the same information as the intersection but also other regions, such as the tails of the hyperbolae. This extra information does not help localization. In fact, likelihood mass outside true source position increases the estimator variance. However, this extra likelihood mass can be considered in other applications, for example, to determine the speaker's head orientation [25].

```
1 X_t = SIR {X_{t-1}, R_t};
2 for j = 1 to N_j do
3     r_t^j ~ P(r_t | r_{t-1}^j);
4     Calculate w_t^j = P(R_t | r_t^j);
5 end
6 Normalize weights, w_t^{1:N_j} / Σ_{j=1}^{N_j} w_t^j;
7 X_t = RESAMPLE {X_t};
```

Algorithm 1: SIR algorithm for particle filtering [30].

4.6. Iterative methods for TDE-based source location estimation

A straightforward but computationally expensive approach for source localization is to exhaustively find the maximum value of the SLF. The SRP-PHAT is perhaps the most common way of building the SLF so a lot of algorithms, including the following ones, have been developed to reduce the computational burden. A stochastic [26] and a deterministic [27] ways of reducing the number of SLF evaluations have been presented. These methods iteratively reduce the search volume that contains the maximum point until the volume is small enough. In [28], the fact that a time delay is inverse-mapped into multiple spatial coordinates was utilized to reduce the number of SLF grid evaluations by considering only the neighborhood of the n highest TDE function values. In [29], the SLF is maximized initially at low frequencies that correspond to large spatial blocks. The maximum-valued SLF block is selected and further divided into smaller blocks by increasing the frequency range. The process is repeated until a desired accuracy is reached.

5. SEQUENTIAL SPATIAL LIKELIHOOD ESTIMATION

In the Bayesian framework, the SLF represents the noisy measurement distribution $P(\boldsymbol{\mathcal{R}}_t \mid \mathbf{r}_t)$ at time frame t, where $\boldsymbol{\mathcal{R}}_t$ represents measurement and $\mathbf{r}_t$ state. In the previous section, several means of building the measurement distribution were discussed. The next step is to estimate the source position using the posterior distribution $P(\mathbf{r}_{0:t} \mid \boldsymbol{\mathcal{R}}_{1:t})$. The subindices emphasize that the distribution includes all the previous measurements and state information, unlike the iterative methods discussed above. The state $\mathbf{r}_0$ represents a priori information. The first measurement is available at time frame $t = 1$.

It is possible to estimate the posterior distribution in a recursive manner [4]. This can be done in two steps, termed prediction and update. The prediction of the state distribution is calculated by convolving the posterior distribution with a transition distribution $P(\mathbf{r}_t \mid \mathbf{r}_{t-1})$ written as

$$P(\mathbf{r}_t \mid \boldsymbol{\mathcal{R}}_{1:t-1}) = \int P(\mathbf{r}_t \mid \mathbf{r}_{t-1}) P(\mathbf{r}_{t-1} \mid \boldsymbol{\mathcal{R}}_{1:t-1}) d\mathbf{r}_{t-1}. \quad (20)$$

The new SLF, that is, $P(\boldsymbol{\mathcal{R}}_t \mid \mathbf{r}_t)$ is used to correct the prediction distribution:

$$P(\mathbf{r}_t \mid \boldsymbol{\mathcal{R}}_{1:t}) = \frac{P(\boldsymbol{\mathcal{R}}_t \mid \mathbf{r}_t) P(\mathbf{r}_t \mid \boldsymbol{\mathcal{R}}_{1:t-1})}{\int P(\boldsymbol{\mathcal{R}}_t \mid \mathbf{r}_t) P(\mathbf{r}_t \mid \boldsymbol{\mathcal{R}}_{1:t-1}) d\mathbf{r}_t}, \quad (21)$$

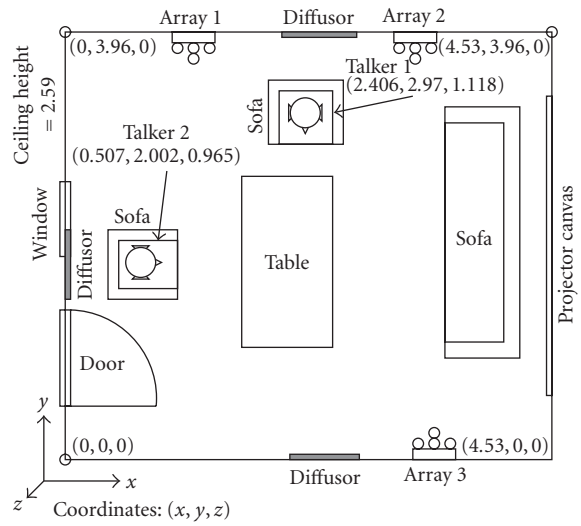


FIGURE 4: A diagram of the meeting room. The room contains furniture, a projector canvas, and three diffusors. Three microphone arrays are located on the walls. Talker positions are given [m], and they are identical in the simulations and in the real-data experiments.

where the nominator is a normalizing constant. For each time frame t, the two steps (20) and (21) are repeated.

In this work, a particle filtering method is used to numerically estimate the integrals involved [4, 30]. For a tutorial on PF methods, refer to [30]. PF approximates the posterior density with a set of N_j weighted random samples $\mathcal{X}_t = \{\mathbf{r}_t^j, w_t^j\}_{j=1}^{N_j}$ for each frame t. The approximate posterior density is written as

$$P(\mathbf{r}_{0:t} \mid \boldsymbol{\mathcal{R}}_{1:t}) \approx \sum_{j=1}^{N_j} w_t^j \delta(\mathbf{r}_{0:t} - \mathbf{r}_{0:t}^j), \tag{22}$$

where the scalar weights $w_t^{1,\ldots,N_j}$ sum to unity, and δ is the Dirac's delta function.

In this work, the particles $\mathbf{r}_t^{1,\ldots,N_j}$ are 3D points in space. The specific PF method used is the sampling importance resampling (SIR), described in Algorithm 1. The algorithm propagates the particles according to the motion model which is here selected as a dual-Gaussian distribution (Brownian motion). Both distributions are centered on the current estimate with standard deviations of σ and 4σ, (see Algorithm 1 Line 3). The new weights are calculated from the SLF on Line 4.

The resampling is applied to avoid the degeneracy problem, where all but one particle have insignificant weight. In the resampling step, particles of low weight are replaced with particles of higher weight. In addition, a percentage of the particles are randomly distributed inside the room to notice events like the change of the active speaker. After estimating the posterior distribution, a point estimate is selected to represent the source position. Point estimation methods include the maximum a posteriori (MAP), the conditional mean (CM), and the median particle. If the SLF is multimodal, CM will be in the center of the mass and thus not necessarily near any source. In contrast, MAP and median will be inside a mode. Due to the large number of particles, the median is less likely to oscillate between different modes than MAP. In SIR, the MAP would be the maximum weighted particle from the SLF and thus prone to spurious peaks. Also, the MAP cannot be taken after the resampling step since the weights are effectively equal. Therefore, the median is selected as the source state estimate:

$$\hat{\mathbf{r}}_t = \text{median}\left\{\mathbf{r}_t^1, \mathbf{r}_t^2, \ldots, \mathbf{r}_t^{N_j}\right\}. \tag{23}$$

6. SIMULATION AND RECORDING SETUP

A dialogue situation between talkers is analyzed. The localization methods already discussed are compared using simulations and real-data measurements performed in a room environment. The simulation is used to analyze how the different TDE combination methods affect the estimation performance when noise and reverberation are added. The real-data measurements are used to verify the performance difference.

The meeting room dimensions are $4.53 \times 3.96 \times 2.59$ m. The room layout and talker locations are illustrated in Figure 4. The room contains three identical microphone arrays. Each array consists of four microphones, and their coordinates are given in Table 1. The real room is additionally equipped with furniture and other small objects.

6.1. Real-data measurements

The measured reverberation time T60 of the meeting room is 0.25 seconds, obtained with the maximum-length sequence (MLS) technique [31] using the array microphones and a loudspeaker. A sampling rate of 44.1 kHz is used, with 24 bits per sample, stored in linear PCM format. The array microphones are Sennheiser MKE 2-P-C electret condenser microphones with a 48 V phantom feed.

Table 1: Microphone geometry for the arrays is given for each microphone (mm). The coordinate system is the same used in Figure 4.

Array 1				Array 2				Array 3			
Mic	x	y	z	Mic	x	y	z	Mic	x	y	z
1	1029	3816	1690	5	3127	3816	1715	9	3714	141	1630
2	1405	3818	1690	6	3507	3813	1715	10	3335	144	1630
3	1215	3819	2088	7	3312	3814	2112	11	3527	140	2030
4	1215	3684	1898	8	3312	3684	1940	12	3517	270	1835

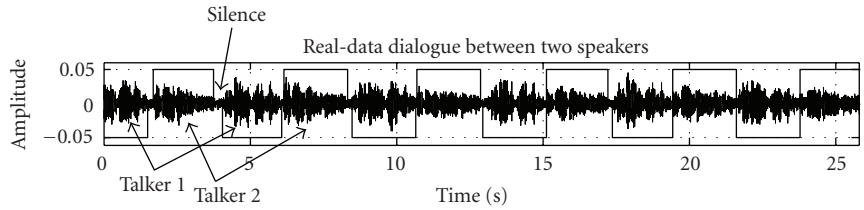


Figure 5: The real-data dialogue signal is plotted from one microphone. The signal is annotated into "talker 1", "talker 2", and "silence" segments. The annotation is also illustrated. The talkers repeated their own sentence.

A 26 second dialogue between human talkers was recorded. The talkers uttered a predefined Finnish sentence and repeated the sentence in turns for six times. The SNR is estimated to be at least 16 dB in each microphone. The recording signal was manually annotated into three different classes "talker 1", "talker 2", and "silence". Figure 5 displays the signal and its annotation. The reference position is measured from the talker's lips and contains some errors due to unintentional movement of the talker and the practical nature of the measurement.

6.2. Simulations

The meeting room is simulated using the image method [32]. The method estimates the impulse response $h_{i,n}(t)$ between the source n and receiving microphone i. The resulting microphone signal is calculated using (1). The reverberation time (T60) of the room is varied by changing the reflection coefficient of the walls β_w, and the ceiling and floor $\beta_{c,f}$ which are related by $\beta_{c,f} = \sqrt{\beta_w}$. The coefficient determines the amount of sound energy reflected from a surface. Recordings with 10 different T60 values between 0 and 0.9 second are simulated with SNR ranging from −10 dB to +30 dB in 0.8 dB steps for each T60 value. The simulation signals consisted of 4 seconds of recorded babble. The active talker switches from talker 1 to talker 2 at time 2.0 seconds. The total number of recordings is 510. The T60 values are [0, 0.094, 0.107, 0.203, 0.298, 0.410, 0.512, 0.623, 0.743, 0.880]. These are median values of channel T60 values calculated from the impulse response using Schroeder integration [33].

7. LOCALIZATION SYSTEM FRAMEWORK

The utilized localization system is based on the ASL framework discussed in this work. Microphone pairwise TDE functions are calculated inside each array with GCC-PHAT [19]. Pairwise GCC values are normalized between [0,1] by first subtracting the minimum value and dividing by the largest such GCC value of the array. A Hamming windowed frame of size 1024 samples is utilized (23.2 milliseconds) with no overlapping between sequential frames. The microphones are grouped into three arrays, and each array contains four microphones, see Table 1. Six unique pairs inside each array are utilized. Microphone pairs between the arrays are not included in order to lessen the computational complexity. The TDE function values are combined with the following schemes, which are considered for ASL:

(1) SRP-PHAT + PF: PHAT-weighted GCC values are summed to form the SLF (14), and SIR-PF algorithm is applied.

(2) Multi-PHAT + PF: PHAT-weighted GCC values are multiplied together to form the SLF (15), and SIR-PF algorithm is applied.

(3) Hamacher-PHAT + PF: PHAT-weighted GCC values are combined pairwise using the Hamacher t-norm (16), with parameter value $\gamma = 0.75$. The SIR-PF algorithm is then applied.

(4) MCCC + PF: PHAT-weighted GCC values are formed into a matrix (17), and the determinant operator is used to combine the pairwise array TDE functions (18). Multiplication is used to combine the resulting three array likelihoods together. In the simulation, multiplication produced better results than using the determinant operator for the array likelihoods. The SIR-PF algorithm is also applied.

The particle filtering algorithm discussed in Section 5 (SIR-PF) is used with 5000 particles. The systematic resampling was applied due to its favorable resampling quality and low computational complexity [34]. The particles are confined to room dimensions and in the real-data analysis also

between heights of 0.5–1.5 m to reduce the effects of ventilation noise. The 5000 particles have a Brownian motion model, with empirically chosen standard deviation σ values 0.05 and 0.01 m for the simulations and real-data experiments, respectively. The Brownian motion model was selected since the talkers are somewhat stationary. Different dynamic models could be applied if the talkers move [35].The particles are uniformly distributed inside the room at the beginning of each run, that is, the a priori spatial likelihood function is uniform.

7.1. Estimator performance

The errors are measured in terms of root mean square (RMS) values of the 3D distance between the point estimate $\hat{\mathbf{r}}_t$ and reference position $\mathbf{r}_t$. The RMS error of an estimator is defined as

$$\text{RMSE}\{\text{method}\} = \frac{1}{T}\sqrt{\left(\sum_{t=1}^{T}||\hat{\mathbf{r}}_t - \mathbf{r}_t||^2\right)}, \qquad (24)$$

where t is the frame index, and T represents the number of frames.

In the real-data analysis, the time frames annotated as "silence" are omitted. 0.3 second of data is omitted from the beginning of the simulation and after the speaker change to reduce the effects of particle filter convergence on the RMS error. Omitting of nonspeech frames could be performed automatically with a voice activity detector (VAD), see for example [36].

7.2. Results for simulations

Results for the simulations using the four discussed ASL methods are given in Figures 6 and 7, for talker locations 1 and 2, respectively. The subfigures (a) to (d) represent the RMS error contours for each of the four methods. The x-axis displays the SNR of the recording, and y-axis displays the reverberation time (T60) value of the recording. A large RMS error value indicates that the method does not produce meaningful results.

For all methods, talker location 1 results in better ASL performance, than location 2. The results of location 1 are examined in detail.

The multi- and Hamacher-PHAT (intersection) methods clearly exhibit better performance. At +14 dB SNR, the intersection methods have RMSE $\leq$ 20 cm when reverberation time T60 $\leq$ 0.4 second. In contrast, the SRP- and MCCC-PHAT attain the same error with T60 $\leq$ 0.2 second.

The results for talker location 2 are similar, except that there exists a systematic increase in RMS error. The decrease in performance is mainly caused by the slower convergence of the particle filter. At the start of the simulation, talker 1 becomes active and all of the particles are scattered randomly inside the room, according to the a priori distribution. When talker 2 becomes active and talker 1 silent, most of the particles are still at talker 1 location, and only a percent of the particles are scattered in the room. Therefore, the particle filter is more likely to converge faster to talker 1 than to talker 2, which is seen in the systematic increase of RMSE.

Evident in larger area of RMS error contour below 0.2 m multi- and Hamacher-PHAT increase the performance both in noisy and reverberant environments compared to SRP- and MCCC-PHAT.

7.3. Results for real-data measurements

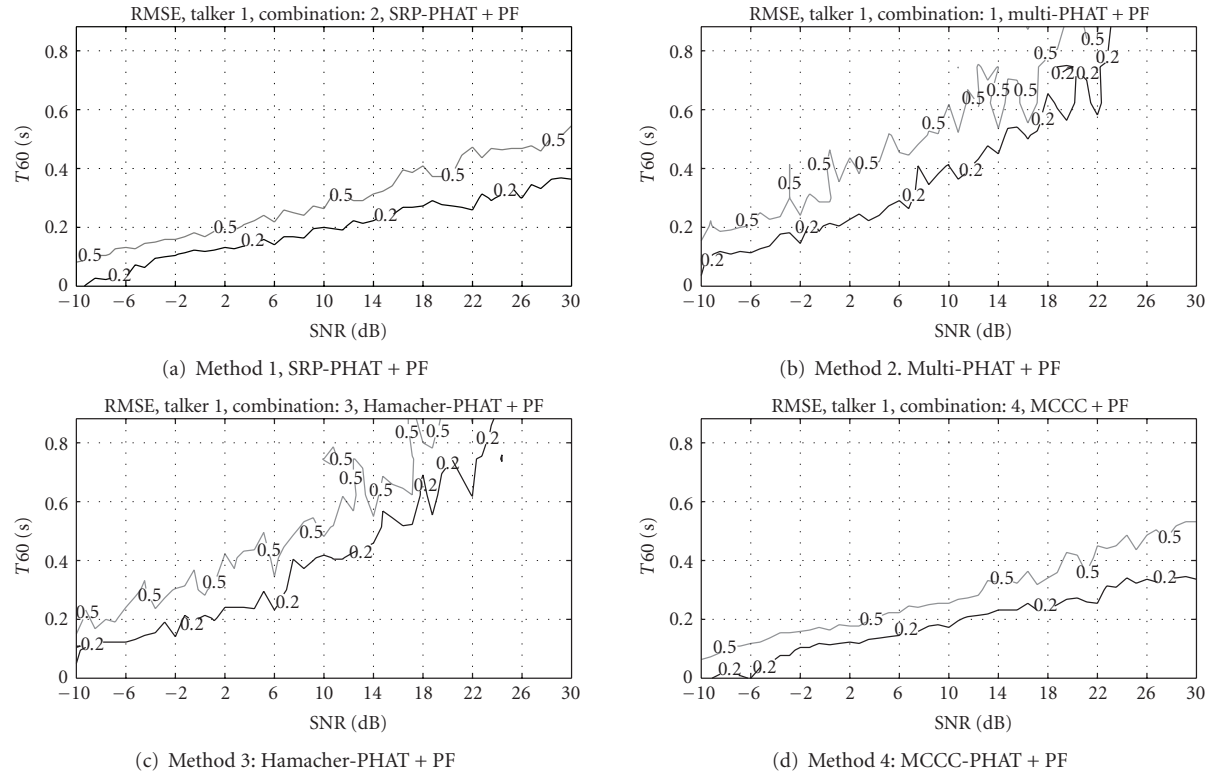
Since the location estimation process utilizes a stochastic method (PF), the calculations are repeated 500 times and then averaged. The averaged results are displayed for the four methods in Figure 8. The location estimates are plotted with a continuous line, and the active talker is marked with a dashed line. All methods converge to both speakers. The SRP-PHAT and MCCC-PHAT behave smoothly. The multi-PHAT and Hamacher-PHAT adapt to the switch of the active speaker more rapidly than other methods and also exhibit rapid movement of the estimator compared to the SRP- and MCCC-PHAT methods.

The RMS errors of the real-data segment are SRP-PHAT: 0.31 m, MCCC-PHAT: 0.29 m, Hamacher-PHAT: 0.14 m, and multi-PHAT: 0.14 m. The performance in the real-data scenario is further illustrated in Figure 9. The percentage of estimates outside a sphere centered at the ground truth location of both talkers is examined. The sphere radius is used as a threshold value to determine if an estimate is an outlier. The Hamacher-PHAT outperforms the others methods. SRP-PHAT has 80.6% of estimates inside the 25 cm error threshold, the MCCC-PHAT has 81.8%, the Hamacher-PHAT has 93.1%, and the multi-PHAT has 92.4%.

The results agree with the simulations. The reason for the performance difference can be further examined by looking at the SLF shape. For this analysis, the SLFs are evaluated with a uniform grid of 5 cm density over the whole room area at three different elevations (0.95, 1.05, and 1.15 m). The marginal SLF is generated by integrating SLFs over the z-dimension and time. The normalized marginal spatial likelihood functions are displayed in Figure 10. In the RMSE sense (24), the likelihood mass is centered around the true position $\mathbf{r}$ in all cases. However, Hamacher- and multi-PHAT likelihood distributions have greater peakiness with more likelihood mass concentrated around the talker. The SRP-PHAT and MCCC-PHAT have a large evenly distributed likelihood mass, that is, large variance. Note that only a single talker was active at a time, and the marginal SLFs are multimodal due to integration over the whole recording time.

8. DISCUSSION

The simulations use the image method which simplifies the acoustic behavior of the room and source. The simulations neglect that the reflection coefficient is a function of the incident angle and frequency, and that the air itself absorbs sound [37]. The effect of the latter becomes more significant in large enclosures. The human talker is acoustically modeled as a point source. This simplification is valid for the simulations, since the data is generated using this assumption. In the real-data scenario, the sound does not originate from a



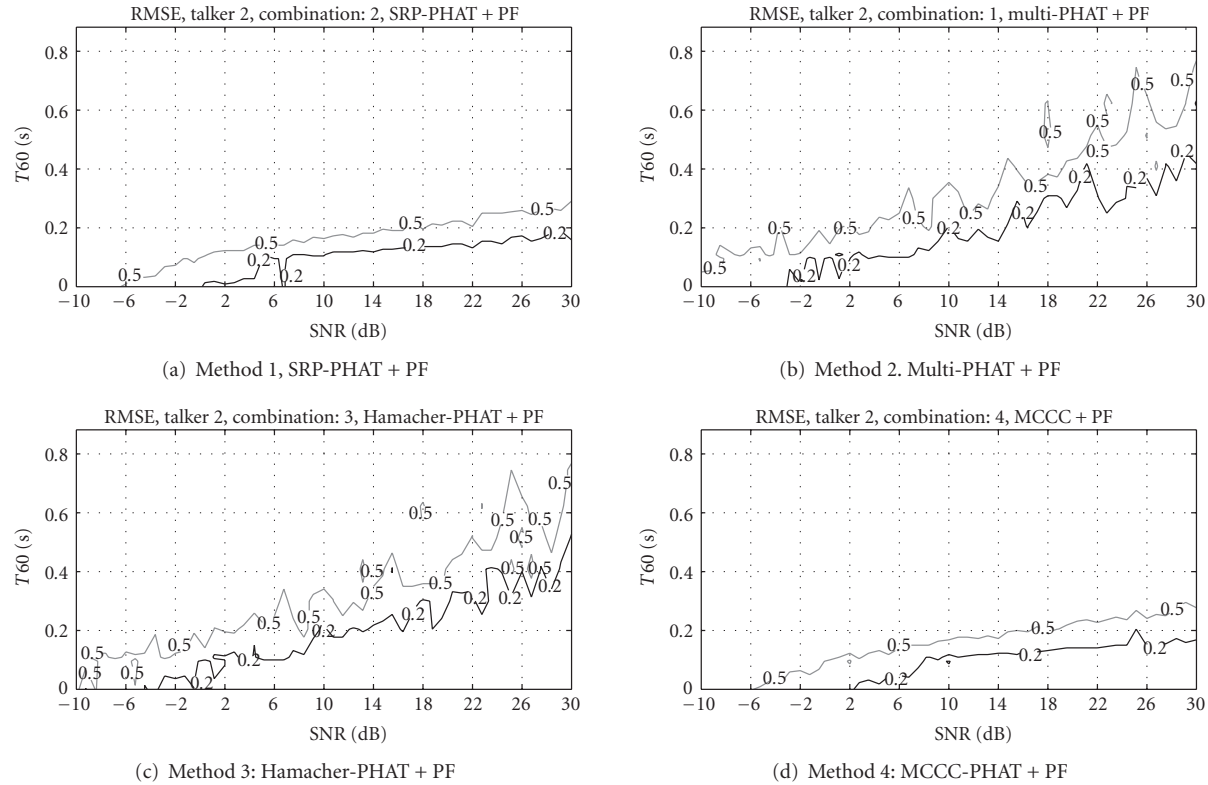
(a) Method 1, SRP-PHAT + PF

(b) Method 2. Multi-PHAT + PF

(c) Method 3: Hamacher-PHAT + PF

(d) Method 4: MCCC-PHAT + PF

Figure 6: The figure presents simulation results for talker location 1. The four ASL methods used are described in Section 7. The RMS error is defined in Section 7.1. The signals SNR values range from −10 to 30 dB, with reverberation time T60 between 0 and 0.9 second, see Section 6. The contour lines represent RMS error values at steps [0.2, 0.5] m.



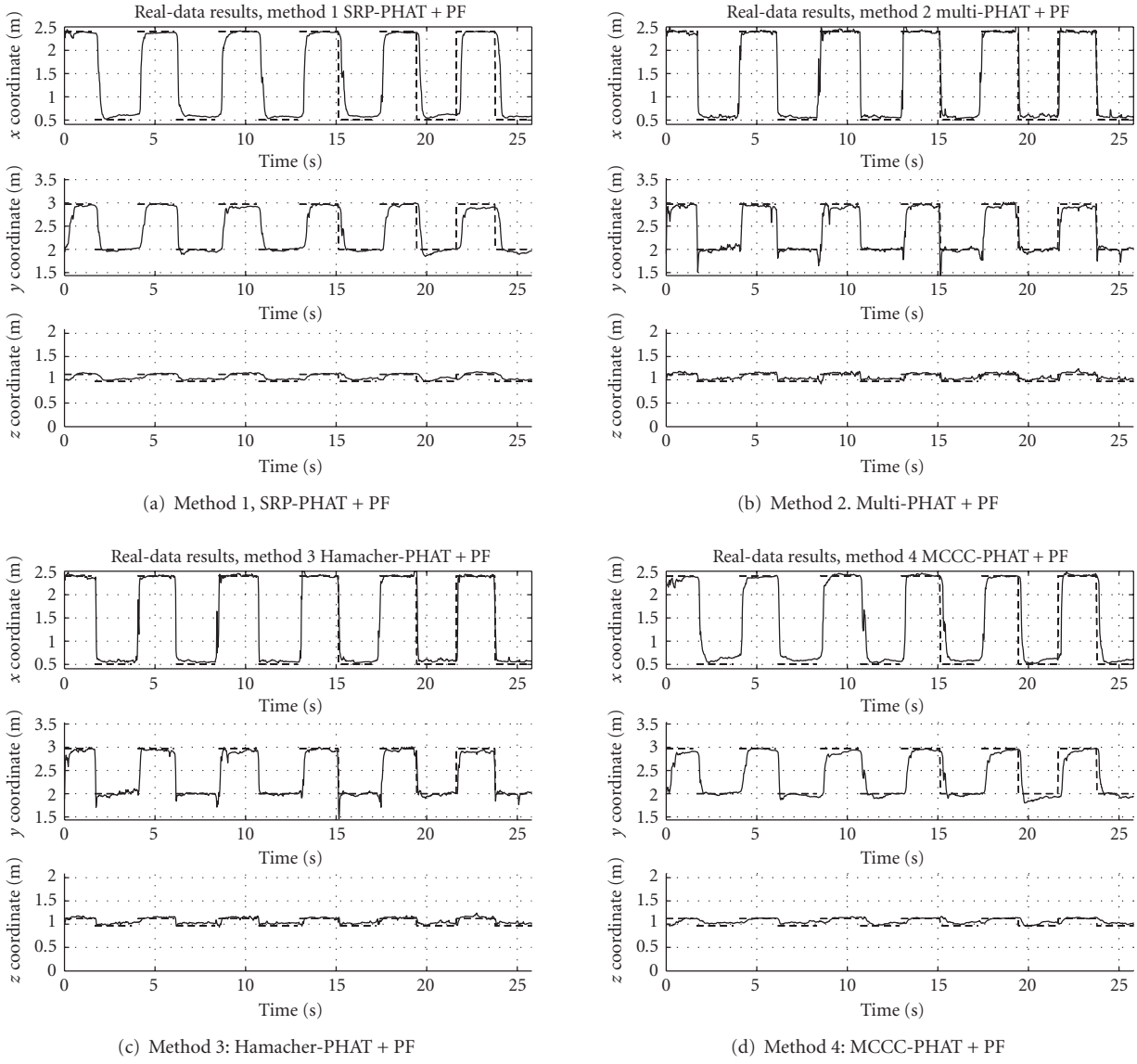
(a) Method 1, SRP-PHAT + PF

(b) Method 2. Multi-PHAT + PF

(c) Method 3: Hamacher-PHAT + PF

(d) Method 4: MCCC-PHAT + PF

Figure 7: The figure presents simulation results for talker location 2. The four ASL methods used are described in Section 7. The RMS error is defined in Section 7.1. The signals SNR values range from −10 to 30 dB, with reverberation time T60 between 0 and 0.9 second, see Section 6. The contour lines represent RMS error values at steps [0.2, 0.5] m.



(a) Method 1, SRP-PHAT + PF

(b) Method 2. Multi-PHAT + PF

(c) Method 3: Hamacher-PHAT + PF

(d) Method 4: MCCC-PHAT + PF

FIGURE 8: Real-data results averaged over 500 runs using the four methods described in Section 7 are plotted. The reference is also plotted with a dashed line. Refer to Figure 4 for room geometry. The x-axis in each picture represents time in seconds. The y-axis displays the corresponding x, y, z coordinates of the result.

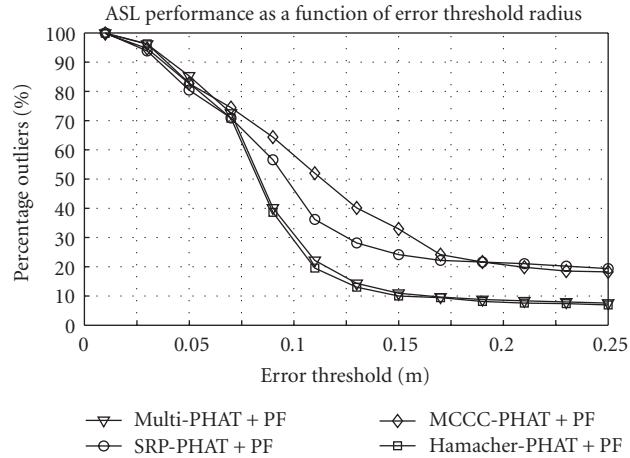


FIGURE 9: The figure displays the percentage of the estimates (y-axis) falling outside of a sphere centered at the active speaker. The sphere radius is plotted on the x-axis (threshold value).

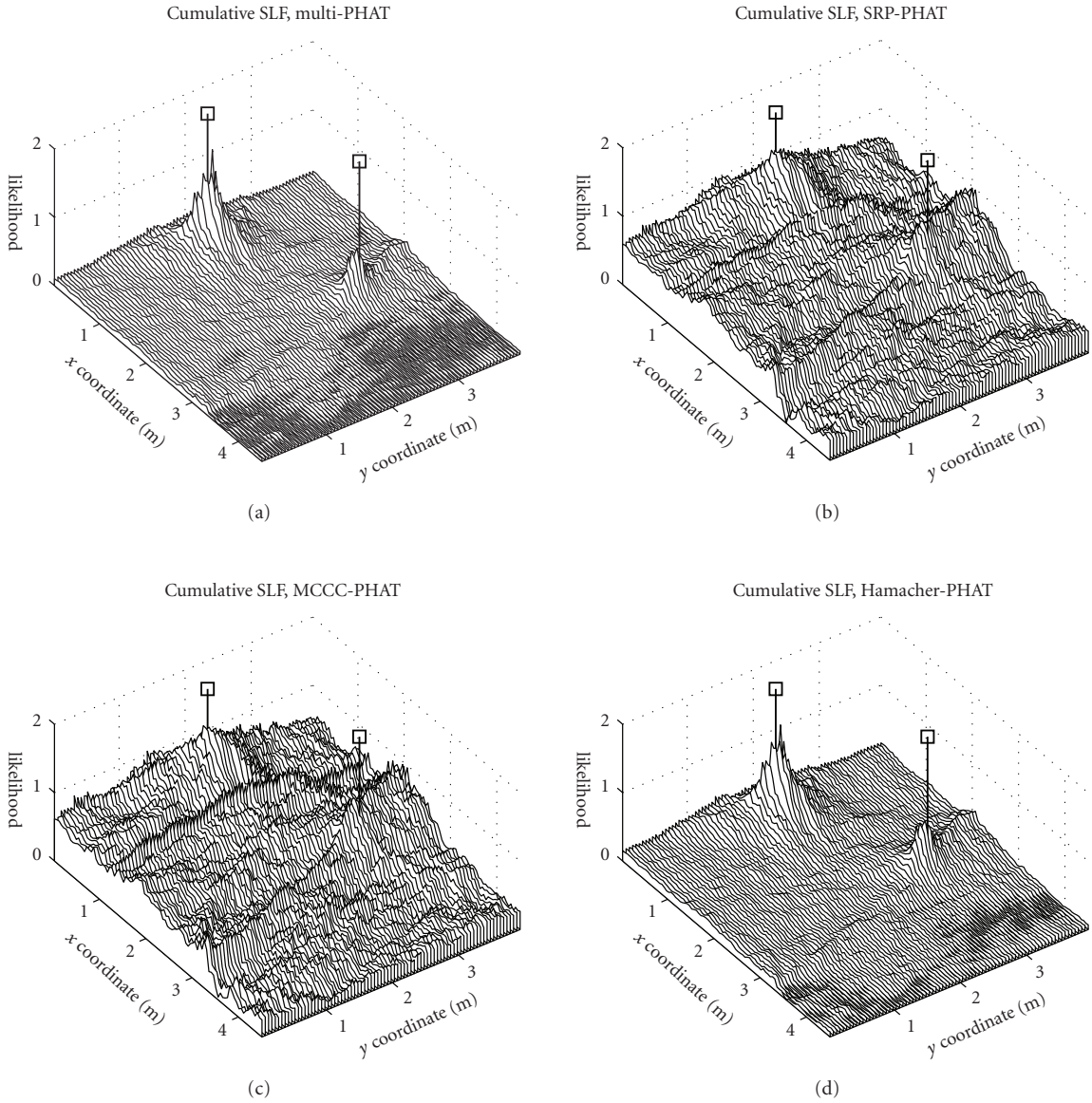


Figure 10: The marginal spatial likelihood functions from real-data recording are presented. The talker locations are marked with a square symbol ("□"). The z-axis is the marginalized spatial likelihood over the whole conversation. In the RMSE sense (24), the likelihood mass is centered around the true position $\mathbf{r}$ in all cases.

single point in space, but rather from the whole mouth area of the speaker. Human speech is also directive, and the directivity increases at higher frequencies [37].

Due to the above facts, the simulation results presented here are not absolute performance values and can change when the system is applied in a real environment. However, the same exact simulation data was applied when comparing the methods. The results, therefore, give information about the relative performance of the methods under the simulation assumptions.

The methods were tested on a real recorded dialogue. All the methods were capable of determining the location of the sound source with varying accuracy. It is likely that the manual annotation and reference measurements contain some errors that affect the reported performance. The only difference between the methods was the way the spatial likelihood function was constructed from the pairwise microphone TDE functions. Since the intersection-based TDE combination methods have better variance, they offer more evidence for the sound source and therefore their convergence is also faster.

9. CONCLUSION

This article discusses a class of acoustic source localization (ASL) methods based on a two-step approach where first the measurement data is transformed using a time delay estimation (TDE) function and then combined to produce

the spatial likelihood function (SLF). The SLF is used in a sequential Bayesian framework to obtain the source position estimate.

A general framework for combining the TDE functions to construct the SLF was presented. Combining the TDE functions using a union operation distributes more likelihood mass outside the source position compared to the intersection of TDE functions. The variance of the spatial likelihood distribution that is constructed with the intersection is thus lower. The particle filter converged faster with a low variance spatial likelihood function than a large variance likelihood function. This is evident in the simulation and real-data results.

Four different schemes to build the SLF from PHAT-weighted GCC values are implemented, specifically: multiplication, Hamacher t-norm (generalized multiplication), summation, and a determinant-based combination. The first two methods represent intersection, the summation represents union, and the determinant falls out of the presented TDE function categorization. In the experiments, the intersection methods gave the best results under different SNR and reverberation conditions using a particle filter. The location RMS error was reduced by 45% by preferring the intersection over the union when constructing the SLF.

ACKNOWLEDGMENTS

The authors wish to thank Dr. Eric Lehmann, for providing a simulation tool for the image method simulations, Sakari Tervo (M.S.) for assistance, Mikko Parviainen (M.S.), and the anonymous reviewers for their comments and suggestions.

REFERENCES

[1] R. Stiefelhagen and J. Garofolo, "Eval-ware: multimodal interaction," *IEEE Signal Processing Magazine*, vol. 24, no. 2, pp. 154–155, 2007.

[2] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 3, pp. 288–292, 1997.

[3] X. Sheng and Y.-H. Hu, "Maximum likelihood multiple-source localization using acoustic energy measurements with wireless sensor networks," *IEEE Transactions on Signal Processing*, vol. 53, no. 1, pp. 44–53, 2005.

[4] A. Doucet, N. de Freitas, and N. Gordon, Eds., *Sequential Monte Carlo Methods in Practice*, Statistics for Engineering and Information Science, Springer, New York, NY, USA, 2001.

[5] P. Aarabi, "The fusion of distributed microphone arrays for sound localization," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 4, pp. 338–347, 2003.

[6] A. Weiss and E. Weinstein, "Fundamental limitations in passive time delay estimation—part 1: narrow-band systems," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 31, no. 2, pp. 472–486, 1983.

[7] B. Champagne, S. Bédard, and A. Stéphenne, "Performance of time-delay estimation in the presence of room reverberation," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 2, pp. 148–152, 1996.

[8] T. Gustafsson, B. D. Rao, and M. Trivedi, "Source localization in reverberant environments: modeling and statistical analysis," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 791–803, 2003.

[9] F. Reed, P. Feintuch, and N. Bershad, "Time delay estimation using the LMS adaptive filter—static behavior," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 3, pp. 561–571, 1981.

[10] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *Journal of the Acoustical Society of America*, vol. 107, no. 1, pp. 384–391, 2000.

[11] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: an overview," *EURASIP Journal on Applied Signal Processing*, vol. 2006, Article ID 26503, 19 pages, 2006.

[12] J. C. Chen, R. E. Hudson, and K. Yao, "A maximum-likelihood parametric approach to source localizations," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '01)*, vol. 5, pp. 3013–3016, Salt Lake City, Utah, USA, May 2001.

[13] J.-M. Valin, F. Michaud, and J. Rouat, "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering," *Robotics and Autonomous Systems*, vol. 55, no. 3, pp. 216–228, 2007.

[14] J. DiBiase, H. F. Silverman, and M. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays: Signal Processing Techniques and Applications*, pp. 157–180, chapter 8, Springer, Berlin, Germany, 2001.

[15] E. A. Lehmann, "Particle filtering methods for acoustic source localisation and tracking," Ph.D. dissertation, Australian National University, Canberra, Australia, July 2004.

[16] T. Korhonen and P. Pertilä, "TUT acoustic source tracking system 2007," in *Proceedings of the 2nd Annual International Evaluation Workshop on Classification of Events, Activities and Relationships (Clear '07)*, R. Stiefelhagen, R. Bowers, and J. Fiscus, Eds., Baltimore, Md, USA, May 2007.

[17] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation using spatial correlation techniques," in *Proceedings of the 8th International Workshop Acoustic Echo and Noise Control (IWAENC '03)*, pp. 207–210, Kyoto, Japan, September 2003.

[18] D. B. Ward, E. A. Lehmann, and R. C. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 826–836, 2003.

[19] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[20] J. Hassab and R. Boucher, "Performance of the generalized cross correlator in the presence of a strong spectral peak in the signal," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 3, pp. 549–555, 1981.

[21] J. Chen, J. Benesty, and Y. Huang, "Performance of GCC- and AMDF-based time-delay estimation in practical reverberant environments," *EURASIP Journal on Applied Signal Processing*, vol. 2005, no. 1, pp. 25–36, 2005.

[22] P. Aarabi and S. Mavandadi, "Robust sound localization using conditional time-frequency histograms," *Information Fusion*, vol. 4, no. 2, pp. 111–122, 2003.

[23] J.-S. Jang, C.-T. Sun, and E. Mizutani, *Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence*, chapter 2, Prentice-Hall, Upper Saddle River, NJ, USA, 1997.

[24] J. N. Ash and R. L. Moses, "Acoustic time delay estimation and sensor network self-localization: experimental results," *Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 841–850, 2005.

[25] A. Brutti, "Distributed microphone networks for sound source localization in smart rooms," Ph.D. dissertation, DIT - University of Trento, Trento, Italy, 2007.

[26] H. Do and H. F. Silverman, "A fast microphone array SRP-PHAT source location implementation using coarse-to-fine region contraction (CFRC)," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA '07)*, pp. 295–298, New Paltz, NY, USA, October 2007.

[27] H. Do, H. F. Silverman, and Y. Yu, "A real-time SRP-PHAT source location implementation using stochastic region contraction(SRC) on a large-aperture microphone array," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '07)*, vol. 1, pp. 121–124, Honolulu, Hawaii, USA, April 2007.

[28] J. Dmochowski, J. Benesty, and S. Affes, "A generalized steered response power method for computationally viable source localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2510–2526, 2007.

[29] D. N. Zotkin and R. Duraiswami, "Accelerated speech source localization via a hierarchical search of steered response power," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 499–508, 2004.

[30] S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Transactions on Signal Processing*, vol. 50, no. 2, pp. 174–188, 2002.

[31] D. Rife and J. Vanderkooy, "Transfer-function measurement with maximum-length sequences," *Journal of the Audio Engineering Society*, vol. 37, no. 6, pp. 419–444, 1989.

[32] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[33] M. R. Schroeder, "New method of measuring reverberation time," *Journal of the Acoustical Society of America*, vol. 37, no. 3, pp. 409–412, 1965.

[34] J. Hol, T. Schön, and F. Gustafsson, "On resampling algorithms for particle filters," in *Proceedings of the Nonlinear Statistical Signal Processing Workshop*, pp. 79–82, Cambridge, UK, September 2006.

[35] E. A. Lehmann, A. M. Johansson, and S. Nordholm, "Modeling of motion dynamics and its influence on the performance of a particle filter for acoustic speaker tracking," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA '07)*, pp. 98–101, New Paltz, NY, USA, October 2007.

[36] E. A. Lehmann and A. M. Johansson, "Particle filter with integrated voice activity detection for acoustic source tracking," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, Article ID 50870, 11 pages, 2007.

[37] L. Beranek, *Acoustics*, American Institute of Physics, New York, NY, USA, 1986.