EXPERIMENTAL PERFORMANCE ASSESSMENT OF A PARTICLE FILTER WITH VOICE ACTIVITY DATA FUSION FOR ACOUSTIC SPEAKER TRACKING

Eric A. Lehmann and Anders M. Johansson

Western Australian Telecommunications Research Institute
35 Stirling Highway, Perth WA 6009, Australia
E-mail: Eric.Lehmann@watri.org.au, ajh@watri.org.au

ABSTRACT

The problem of acoustic source localization and tracking (ASLT) in reverberant environments by means of a microphone array constitutes a challenging task from many viewpoints. One of the main issues when considering real-world situations involving human speakers is the presence of silence gaps in the speech, which can easily send the tracking algorithm off-track, even in practical environments with low to moderate noise and reverberation levels. This work is concerned with an implementation of the ASLT algorithm proposed in [1], which circumvents this problem by integrating measurements from a voice activity detector (VAD) within the tracking algorithm framework. The tracking performance of this method is tested experimentally using audio data recorded in a real reverberant room. To this purpose, we describe a quick and efficient way of determining the ground-truth speaker location versus time, an operation that is not always easy to perform. The experimental results confirm the improved robustness of the method presented in [1] (compared to a previously proposed non-VAD ASLT algorithm) when tracking sources emitting real-world speech signals, which typically involve significant silence gaps between utterances.

1. INTRODUCTION

The concept of speaker tracking using an array of acoustic sensors has become an increasingly important field of research over the last few years [2–5]. Typical applications such as teleconferencing, automated multi-media capture, smart meeting rooms and lecture theaters, etc., are fast becoming an engineering reality. This in turns requires the development of increasingly sophisticated algorithms to deal efficiently with problems related to background noise and acoustic reverberation during the speech acquisition process.

One of the major difficulties in a practical implementation of ASLT for speech-based applications lies in the nonstationary character of typical speech signals, with potentially significant silence periods existing between separate utterances. During such silence gaps, currently available ASLT methods will usually keep updating the source location estimates as if the speaker was still active. The algorithm is therefore likely to momentarily lose track of the true source position since the updates are then based solely on disturbance sources such as reverberation and background noise, whose influence might be quite significant in practice. Consequently, existing works on speaker tracking implicitly rely on the fact that silence periods in the speech signal remain relatively short [2–5].

The work presented in [1] deals with this specific issue by fusing VAD observations within the statistical framework of a sequential Monte Carlo algorithm (particle filter, PF). Simulation results of this algorithm, denoted PF-VAD, are provided in [1] on the basis of synthetic audio data generated with the image method [6]. These simulations show that the newly proposed ASLT algorithm has the potential to drastically outperform more basic PF implementations that do not integrate VAD data, such as those presented in [3]. Whereas the image method is useful for an initial test of ASLT algorithms, it is only through experiments using real-world audio data that the real performance of these methods can be gauged.

This paper focuses on a tracking performance assessment of the PF-VAD algorithm presented in [1] using real audio data recorded in a reverberant and noisy environment. The next section briefly reviews the basics of the PF approach. A summary of the PF-VAD developments in [1] is then given in Sections 3 and 4. Section 5 finally presents the results from experimental algorithm tests, followed by a discussion of these results in Section 6.

2. BAYESIAN FILTERING FOR TARGET TRACKING

Consider an array of M acoustic sensors distributed at known locations in a reverberant environment. Assuming a single sound source, the problem consists in estimating the location of this "target" based on the signals $f_m(t)$, $m \in \{1, \dots, M\}$, provided by the array. It is further assumed that the sensor signals are sampled in time, and subsequently decomposed into a series of successive frames $k = 1, 2, \dots$, of equal length L before being processed.

2.1) **State-Space Filtering.** Let $\mathbf{X}_k$ represent the state variable for time frame k, corresponding to the position and velocity of the target in the state space: $\mathbf{X}_k = [x_k \;\; y_k \;\; \dot{x}_k \;\; \dot{y}_k]^{\mathsf{T}}$. At any time step, each microphone in the array delivers a frame of audio signal which can be processed using some localization technique, such as steered beamforming (SBF). Let $\mathbf{Y}_k$ denote the observation variable (measurement), which here typically corresponds to the localization information resulting from the SBF processing of the audio signals. Using a Bayesian filtering approach and assuming Markovian dynamics, this system can be globally represented as follows:

$$\mathbf{X}_k = g(\mathbf{X}_{k-1}, \mathbf{u}_k)\,, \tag{1a}$$

$$\mathbf{Y}_k = h(\mathbf{X}_k, \mathbf{v}_k)\,, \tag{1b}$$

where $g(\cdot)$ and $h(\cdot)$ are possibly nonlinear functions, and $\mathbf{u}_k$ and $\mathbf{v}_k$ are possibly non-Gaussian noise variables. Ultimately, one would like to compute the so-called posterior probability density function (PDF) $p(\mathbf{X}_k|\mathbf{Y}_{1:k})$, where $\mathbf{Y}_{1:k} = \{\mathbf{Y}_1, \dots, \mathbf{Y}_k\}$ represents the concatenation of all measurements up to time k. This density contains all the statistical information available regarding the current

This work was supported by National ICT Australia (NICTA) and the Australian Research Council (ARC) under grant no. DP0451111.

condition of the state variable $\mathbf{X}_k$, and an estimate $\widehat{\mathbf{X}}_k$ of the state then follows, for instance, as the mean or the mode of $p(\mathbf{X}_k|\mathbf{Y}_{1:k})$.

2.2) **Sequential Monte Carlo Approach.** Particle filtering (PF) is an approximation technique that solves the above Bayesian filtering problem by representing the posterior density as a set of N samples of the state space $\mathbf{X}_k^{(n)}$ (particles) with associated weights $w_k^{(n)}$, $n \in \{1, \ldots, N\}$, see, e.g., [7]. The so-called bootstrap algorithm [8] is an attractive PF variant due to its simplicity and low computational demands. Assuming that the set of particles and weights $\{(\mathbf{X}_{k-1}^{(n)}, w_{k-1}^{(n)})\}_{n=1}^N$ is a discrete representation of the posterior density at time $k-1$, $p(\mathbf{X}_{k-1}|\mathbf{Y}_{1:k-1})$, and given the observation $\mathbf{Y}_k$ obtained at the current time k, the bootstrap PF algorithm forms a new set of particles and weights $\{(\mathbf{X}_k^{(n)}, w_k^{(n)})\}_{n=1}^N$, which is an approximate representation of the current posterior $p(\mathbf{X}_k|\mathbf{Y}_{1:k})$. An estimate $\hat{\boldsymbol{\ell}}_k$ of the source position for the current time step k can then be computed according to

$$\hat{\boldsymbol{\ell}}_k = \mathbb{E}\{\boldsymbol{\ell}_k\} \approx \sum\nolimits_{n=1}^{N} w_k^{(n)} \boldsymbol{\ell}_k^{(n)},$$

where $\boldsymbol{\ell}_k^{(n)} = [x_k^{(n)} \;\; y_k^{(n)}]^{\mathrm{T}}$ corresponds to the location information in the n-th particle vector. A second output from the PF algorithm is a measure of the confidence level in the PF estimates, which can be obtained by computing the standard deviation of the particle set:

$$\varsigma_k = \sqrt{\sum\nolimits_{n=1}^{N} w_k^{(n)} \big\| \boldsymbol{\ell}_k^{(n)} - \hat{\boldsymbol{\ell}}_k \big\|^2},$$

where $\|\cdot\|$ denotes the Euclidean norm. The parameter ς_k provides a direct assessment of how reliable the PF considers its current source position estimate to be.

3. PF FOR ACOUSTIC SOURCE TRACKING

The bootstrap PF algorithm requires the definition of two important concepts [8]: the source dynamics, through the transition function $g(\cdot)$, and the so-called likelihood function $p(\mathbf{Y}_k|\mathbf{X}_k^{(n)})$, $n \in \{1, \ldots, N\}$.

3.1) **Target Dynamics.** In order to remain consistent with previous ASLT literature [3, 4], a Langevin process is used to model the dynamics equation (1a). This process is typically used to characterize various types of stochastic motion, and it has proved to be a good choice for speaker tracking. With this model, the source motion in each of the Cartesian coordinates is assumed to be an independent first-order Markov process.

3.2) **Likelihood Function.** The SBF principle is used here as a basis for the derivation of the likelihood function. With $F_m(\omega) = \mathcal{F}\{f_m(t)\}$ the Fourier transform of the signal data from the m-th sensor, the output $\mathcal{P}(\boldsymbol{\ell})$ of a delay-and-sum beamformer steered to the location $\boldsymbol{\ell} = [x \;\; y]^{\mathrm{T}}$ is given as

$$\mathcal{P}(\boldsymbol{\ell}) = \int_{\Omega} \Big| \sum\nolimits_{m=1}^{M} W_m(\omega)\, F_m(\omega)\, \mathrm{e}^{j\omega \|\boldsymbol{\ell} - \boldsymbol{\ell}_m\|/c} \Big|^2 \,\mathrm{d}\omega\,, \quad (2)$$

where $c = 343\,\mathrm{m/s}$ is the speed of sound, $\boldsymbol{\ell}_m = [x_m \;\; y_m]^{\mathrm{T}}$ is the known position of the m-th microphone, and Ω corresponds to the frequency range of interest, typically defined for speech processing applications as $\Omega = \{\omega \,|\, 2\pi \cdot 300\,\mathrm{Hz} \leqslant \omega \leqslant 2\pi \cdot 3000\,\mathrm{Hz}\}$. The frequency weighting term $W_m(\cdot)$ is computed according to the PHAT (phase transform) weighting, i.e., $W_m(\omega) = 1/|F_m(\omega)|$.

In the PF-VAD implementation, an approach based on the concept of a "pseudo-likelihood" is adopted, as introduced previously in [3]. This concept relies on the idea that the SBF output $\mathcal{P}(\cdot)$ itself can be used as a measure of likelihood. For the n-th particle, the likelihood PDF is hence defined as

$$p(\mathbf{Y}|\mathbf{X}^{(n)}) = q_0 \cdot \mathcal{U}(\boldsymbol{\ell}_k^{(n)}) + \gamma\,(1 - q_0) \cdot \big[\mathcal{P}(\boldsymbol{\ell}_k^{(n)})\big]^r, \quad (3)$$

where $\mathcal{U}(\cdot)$ is the uniform PDF defined over the considered room boundaries, q_0 is the prior probability that an SBF measurement might originate from clutter, and the nonlinear exponent r is used to help shape the SBF output to make it more amenable to source tracking [3]. The parameter γ is a normalization constant ensuring that the two PDFs in the mixture likelihood definition of (3) are properly scaled with respect to each other [1].

4. FUSION OF VAD MEASUREMENTS

4.1) **Voice Activity Detection.** The voice activity detector (VAD) employed in [1] relies on an estimate of the instantaneous signal-to-noise ratio (SNR) in the current signal frame. It assumes that the data recorded at the microphones is an additive combination of the clean speech signal and noise.

The scheme works on the basis of the average noise power spectral density, which is estimated during nonspeech periods. The estimated noise level, which is assumed to vary slowly in relation to the speech power, is then used during periods of speech activity to estimate the SNR from the observed signal. The assumption is that the speaker is active when the frequency-averaged SNR level is higher than a given threshold, which is set in such a way as to minimize the occurrence of false alarms. The specific application considered here also makes it possible to reduce the variance of the signal power estimates by averaging over multiple microphones.

4.2) **VAD Fusion.** The output of the VAD can be linked to the probability q_0 in (3) in an obvious manner. The probability $1 - q_0$ corresponds to the likelihood of the acoustic source being active (non-clutter SBF measurement), an estimate of which is delivered by the VAD. Therefore, instead of setting the variable q_0 to a constant value in the design of the algorithm as done in [3, 4], the following time-varying definition of q_0 is used: $q_0(k) = 1 - \alpha(k)$, with $\alpha(k) \in [0, 1]$ the soft-decision output from the VAD algorithm (where 1 denotes speech and 0 nonspeech). In the current implementation, $\alpha(k)$ corresponds to the estimated speech signal level, derived from the SNR and noise power estimates delivered by the VAD.

The generic PF algorithm resulting from the developments presented so far is denoted PF-VAD.

5. EXPERIMENTAL RESULTS

This section presents the tracking results obtained with algorithm PF-VAD for real audio data. The various parameters of the PF-VAD implementation were optimized empirically and defined as $N = 50$ and $r = 2$. The audio signals were sampled with a frequency $F_{\mathrm{s}} = 16\,\mathrm{kHz}$ and decomposed into frames of $L = 512$ samples each. For comparison purposes, the performance assessment presented in this section also includes results from algorithm SBF-PL, a sound source tracking scheme previously proposed in [3]. The SBF-PL method relies on a particle filtering approach similar to that presented here, but does not include any VAD measurements.

5.1) **Performance Assessment Parameters.** The PF estimation error for the current frame is $\varepsilon_k = \|\boldsymbol{\ell}_{\mathrm{S},k} - \hat{\boldsymbol{\ell}}_k\|$, where $\boldsymbol{\ell}_{\mathrm{S},k}$ is the ground-truth source position at time k. In order to assess the overall performance of the algorithm under test over a given sample of audio data, the average error is simply computed as $\bar{\varepsilon} = \big(\sum_{k=1}^{K} \varepsilon_k\big)/K$, with K representing the total number of frames in the considered audio sample.

Due to the partially random nature of PF implementations, statistical averaging over a large number D of algorithm runs is used in the results presentation. A parameter of particular interest to ASLT is the percentage of these runs for which the tracking algorithm completely loses track of the target during the simulation, typically due to significant silence gaps in the speech. For each simulation run $d \in \{1, \ldots, D\}$, a track loss parameter is thus defined as

$$\xi_d = \begin{cases} 1 & \text{if } \big(\sum_{k=K-k^*}^{K} \varepsilon_{k,d}\big)/(k^* - 1) > \delta\,, \\ 0 & \text{otherwise}\,, \end{cases}$$

where $k^* = \lceil 0.5 \cdot F_\mathrm{s}/L \rceil$. The parameter ξ_d effectively checks if the average estimation error over the last $0.5\,\mathrm{s}$ of audio data is smaller than some threshold, set here to $\delta = 0.1\,\mathrm{m}$, i.e., whether the algorithm is still correctly tracking the target at the end of the simulation run. The global track loss percentage (TLP) $\bar{\xi}$ for a given audio sample is then defined as $\bar{\xi} = \big(\sum_{d=1}^{D} \xi_d\big)/D$.

5.2) **Microphone Array Setup.** An array of $M = 8$ omnidirectional microphones was set up at a constant height of $1.51\,\mathrm{m}$ in a room with dimensions $3.5\,\mathrm{m} \times 3.1\,\mathrm{m} \times 2.2\,\mathrm{m}$, in a square fashion with one sensor pair centered on each side of the square (distance of $0.8\,\mathrm{m}$ between the sensors in each pair). The area spanned by the array was $2.52\,\mathrm{m} \times 2.52\,\mathrm{m}$.

Two different types of environment were considered. In the first one, the walls of the enclosure were fully padded with sound-absorbing panels (no padding on the floor and ceiling), leading to a practical reverberation time $T_{60} \approx 0.27\,\mathrm{s}$ (frequency-averaged up to $24\,\mathrm{kHz}$). In the second environment, parts of the padding were removed, increasing the reverberation to $T_{60} \approx 0.34\,\mathrm{s}$. In each environment, two different signal sources were used. The first one was a male speaker, moving randomly across the room while uttering a series of sentences separated by silence gaps. The second source was a loudspeaker emitting a female speech signal (also containing silence gaps). The loudspeaker was carried randomly across the room during the recordings in order to simulate a mobile source. For each speaker–environment combination, a total of five recordings were taken, each corresponding to a different trajectory and source signal.

In both environments, background noise was recorded separately (i.e., with no speech source present) by means of a series of loudspeakers emitting either average white Gaussian noise (AWGN) or babble noise. The recordings of these noise signals were then combined additively to the speech signals with varying SNR levels in order to generate the input data to the tracking algorithm. This way of splitting the noise and speech recordings specifically allowed the measurement of ground-truth source location data according to the method described below.

5.3) **Source Position Measurements.** In order to achieve an accurate assessment of the tracking performance, it is necessary to obtain ground-truth measurements of the real speaker trajectory during the recordings. A few methods have been proposed and used in previous literature works to this purpose, typically based either on some sort of mechanical system [9], or using the location estimates obtained from a different measurement modality, such as visual tracking [10].

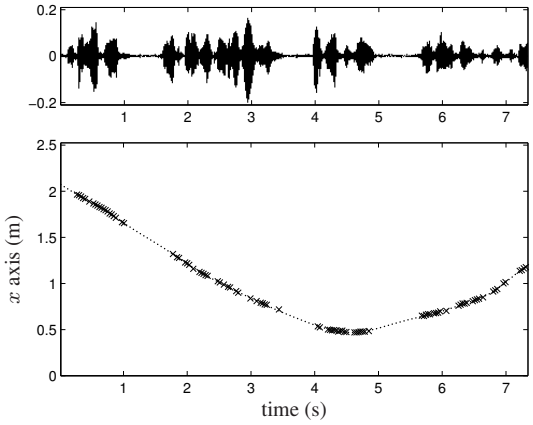

Fig. 1. Ground-truth source position from microphone array data. *Top plot*: example of recorded sensor signal. *Bottom plot*: high-accuracy SBF localization data ($\times$) and polynomial trajectory interpolation (dotted line), for the x coordinate.

These approaches however generate considerable difficulties related to, e.g., synchronization of different data streams (audio and video), the inability to use real human speakers in some cases, and most of all, substantial hardware and software setup costs.

In the present work, a different approach is used: the ground-truth location data is extracted directly from the recorded audio data by means of a high-accuracy beamformer scanning the considered enclosure. The microphone signals are split into successive frames of $32\,\mathrm{ms}$ of data, with a 50% overlapping factor. Each frame is processed using the SBF formula in (2), at first computed on a relatively coarse grid across the entire search space. A coarse estimate of the current source position is obtained as the location maximizing this SBF output, and this estimate is then refined by considering a high-resolution grid (uniform $1\,\mathrm{mm}$ spacing between grid points) centered around the region of interest, i.e., around the coarse location estimate. An approximate knowledge of the overall source path across the room, combined with the use of some voice activity detection scheme, allows the easy discrimination of outliers and yields a series of two-dimensional location data points vs. time, as shown in Figure 1 for the x coordinate (results for the y dimension are similar). This plot presents the ground-truth SBF measurements for a sample of audio data recorded with the male speaker in the non-padded room setup ($T_{60} \approx 0.34\,\mathrm{s}$). Finally, a polynomial approximation can be fitted to the SBF localization data in order to obtain an estimate of the true source trajectory over the entire audio sample length.

The approach described here has the main advantage of producing source position data that is automatically synchronized to the audio signals, and requires no additional hardware setup costs. It also allows any kind of sound source to be used along any arbitrary trajectory, and it has proved to work particularly well for the controlled environment considered in this work. It is estimated that the ground-truth location data generated with this method is accurate to within a couple of centimeters of the true source trajectory.

5.4) **Experimental Results.** Figure 2 shows a typical tracking example for PF-VAD, obtained for a male speaker recording in the padded environment ($T_{60} \approx 0.27\,\mathrm{s}$) with $15\,\mathrm{dB}$ SNR (white noise). This plot clearly demonstrates how the particle set spatially expands during nonspeech periods (increasing standard deviation ς_k), allowing the algorithm to keep track of the silent target and successfully

		$T_{60} \approx 0.27$ s		$T_{60} \approx 0.34$ s	
		male	female	male	female
AWGN 5 dB	error $\bar{\varepsilon}$	0.488	0.558	0.860	0.869
	TLP $\bar{\xi}$	73.8	93.5	100.0	100.0
AWGN 10 dB	error $\bar{\varepsilon}$	0.071	0.114	0.280	0.724
	TLP $\bar{\xi}$	1.9	23.0	66.3	100.0
AWGN 15 dB	error $\bar{\varepsilon}$	0.047	0.065	0.143	0.358
	TLP $\bar{\xi}$	1.9	14.5	35.6	91.5
babble 5 dB	error $\bar{\varepsilon}$	0.049	0.099	0.178	0.374
	TLP $\bar{\xi}$	3.8	24.5	39.4	93.5
babble 10 dB	error $\bar{\varepsilon}$	0.044	0.072	0.106	0.332
	TLP $\bar{\xi}$	1.9	17.5	20.6	92.0
babble 15 dB	error $\bar{\varepsilon}$	0.047	0.062	0.092	0.329
	TLP $\bar{\xi}$	4.4	12.0	17.5	92.0

Table 1. Tracking performance results for PF-VAD algorithm: average estimation error $\bar{\varepsilon}$ (m) and track loss percentage $\bar{\xi}$ (%).

		$T_{60} \approx 0.27$ s		$T_{60} \approx 0.34$ s	
		male	female	male	female
AWGN 5 dB	error $\bar{\varepsilon}$	0.646	0.466	0.765	0.831
	TLP $\bar{\xi}$	85.6	90.0	97.5	95.5
AWGN 10 dB	error $\bar{\varepsilon}$	0.560	0.441	0.726	0.796
	TLP $\bar{\xi}$	78.1	83.0	92.5	95.0
AWGN 15 dB	error $\bar{\varepsilon}$	0.380	0.349	0.662	0.785
	TLP $\bar{\xi}$	66.9	71.5	86.9	93.0
babble 5 dB	error $\bar{\varepsilon}$	0.618	0.553	0.902	0.662
	TLP $\bar{\xi}$	83.1	90.5	98.1	96.0
babble 10 dB	error $\bar{\varepsilon}$	0.455	0.450	0.787	0.650
	TLP $\bar{\xi}$	75.0	87.0	92.5	96.5
babble 15 dB	error $\bar{\varepsilon}$	0.300	0.424	0.764	0.633
	TLP $\bar{\xi}$	62.5	81.0	93.1	96.5

Table 2. Tracking performance results for SBF-PL algorithm: average estimation error $\bar{\varepsilon}$ (m) and track loss percentage $\bar{\xi}$ (%).

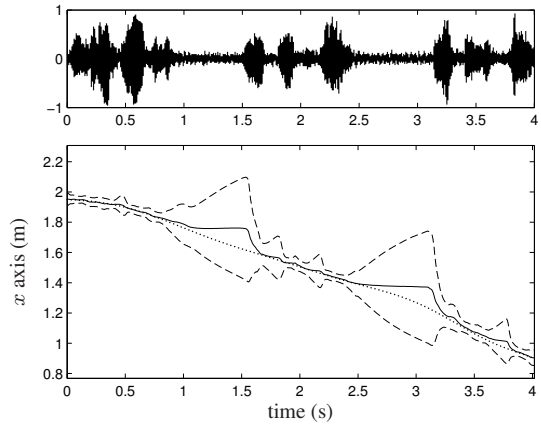

Fig. 2. Tracking result example for PF-VAD. *Top plot*: example of recorded sensor signal. *Bottom plot*: source trajectory (dotted line), position estimates $\hat{\boldsymbol{\ell}}_k$ (solid line), and standard deviation ς_k (dashed line), in x dimension (results for the y coordinate are similar).

resume an accurate tracking when the source becomes active again.

The results presented in Tables 1 and 2 show the average tracking performance obtained with the real-audio recordings for PF-VAD and SBF-PL, respectively. Each result in these tables represents an average computed over 250 values, corresponding to 50 algorithm runs carried out for each of the five audio samples recorded in a given speaker–environment configuration. A comparison between these tables shows that PF-VAD results are consistently better than those obtained for SBF-PL.

6. DISCUSSION

Some of the considered tracking scenarios in the above experiments still prove too difficult to deal with for both tested algorithms, yielding large errors and TLP factors. Apart from the reverberation and noise levels, the degree of tracking difficulty is typically determined by the specific target trajectory (sharp turns), the frequency content of the source signal (male vs. female), and the type of acoustic source utilized (human vs. loudspeaker). The results presented in this work however demonstrate the superiority of PF-VAD compared to a similar PF implementation that does not integrate VAD data. The VAD-based algorithm presented in [1] is more efficient at avoiding track losses during significant speech inactivity periods, and is therefore better suited for practical ASLT implementations.

7. REFERENCES

[1] E. Lehmann and A. Johansson, "Particle filter with integrated voice activity detection for acoustic source tracking," to appear in *EURASIP J. Applied Signal Processing*.

[2] T. Dvorkind and S. Gannot, "Speaker localization exploiting spatial-temporal information," in *Proc. IWAENC*, Kyoto, Japan, Sept. 2003, pp. 295–298.

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

[4] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," in *Proc. IEEE ICASSP*, vol. 5, Salt Lake City, USA, May 2001, pp. 3021–3024.

[5] I. Potamitis, H. Chen, and G. Tremoulis, "Tracking of multiple moving speakers with multiple microphone arrays," *IEEE Trans. Speech Audio Processing*, vol. 12, no. 5, pp. 520–529, 2004.

[6] J. Allen and D. Berkeley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

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

[8] N. Gordon, D. Salmond, and A. Smith, "Novel approach to nonlinear/non-Gaussian Bayesian state estimation," *IEE Proc. F*, vol. 140, no. 2, pp. 107–113, Apr. 1993.

[9] S. Spors, R. Rabenstein, and N. Strobel, "A multi-sensor object localization system," in *Workshop on Vision, Modeling and Visualization*, vol. 5, Stuttgart, Germany, Nov. 2001, pp. 19–26.

[10] M. Krindis, G. Stamou, H. Teutsch, S. Spors, N. Nikolaidis, R. Rabenstein, and I. Pitas, "An audio-visual database for evaluating person tracking algorithms," in *Proc. IEEE ICASSP*, vol. 2, Philadelphia, USA, Mar. 2005, pp. 237–240.