A MULTIPLE HYPOTHESIS GAUSSIAN MIXTURE FILTER FOR ACOUSTIC SOURCE LOCALIZATION AND TRACKING

Youssef Oualil[1,2], *Friedrich Faubel*[1] *and Dietrich Klakow*[1]

[1] Spoken Language Systems, Saarland University, Saarbrücken, Germany
[2] Idiap Research Institute, CH-1920 Martigny, Switzerland

ABSTRACT

In this work, we address the problem of tracking an acoustic source based on measured time differences of arrival (TDOA). The classical solution to this problem consists in using a detector, which estimates the TDOA for each microphone pair, and then applying a tracking algorithm, which integrates the "measured" TDOAs in time. Such a two-stage approach presumes 1) that TDOAs can reliably be estimated; and 2) that errors in detection behave in a well-defined fashion. The presence of noise and reverberation, however, causes larger errors in the TDOA estimates and, thereby, ultimately lowers the tracking performance. We propose to counteract this effect by considering a multiple hypothesis filter, which propagates the TDOA estimation uncertainty to the tracking stage. That is achieved by considering multiple TDOA estimates and then integrating the resulting TDOA observations in the framework of a Gaussian mixture filter. Experimental results show that the proposed filter has a significantly lower angular error than a multiple hypothesis particle filter.

Index Terms— Direction of arrival estimation, Microphone Arrays, Monte Carlo methods, Kalman filters

1. INTRODUCTION

The problem of TDOA-based source localization can be formulated as a two-stage approach, which consists in first estimating the TDOA that has been introduced at each sensor pair; and then triangulating the source position by integrating the estimated TDOAs in a consistent fashion. While the former is typically performed with the generalized cross correlation (GCC) [1], the latter can elegantly be achieved with a Kalman filter (KF) [2, 3]. Unfortunately, the performance of this approach degrades in the presence of noise and multipath effects, especially under room acoustical conditions where early reflections and reverberation corrupt the GCCs through smearing as well as through the introduction of secondary peaks [4, 5]. This in turn affects the Kalman filter which assumes the error to be a stationary Gaussian process whereas the TDOA error in a multi-path environment is rather time-varying and multimodal. In an attempt to mitigate this problem, Vermaak [5] proposed to use a multiple hypothesis particle filter. This approach has been further improved in [6], where an extended particle filter has been introduced.

In this work, we continue along the lines of [5, 6] by proposing a new multiple hypothesis Gaussian mixture filter (MH-GMF), which propagates the uncertainty in the TDOA estimates to the tracking stage. Contrary to previous multiple hypothesis filters, our approach treats each observation individually, by running a bank of Unscented Kalman filters (UKF). In doing so, the proposed approach incorporates the individual information introduced by each hypothesis. The main problem consists in obtaining the observations. Ideally, we would like to use all possible TDOA combinations from different sensor pairs, weighted with their respective GCC values. As this Cartesian product is computationally intractable, we propose to reduce the number of combinations by first drawing TDOA candidates from the individual GCCs and then combining these TDOAs in a "proximately consistent" fashion. In doing so, we statistically focus on TDOA combinations with high likelihood. In fact, the proposed approach interprets the normalized GCC as a probability density function (pdf) of the TDOA, similar as originally proposed in [5] and firstly applied in [7] for a steered response power (SRP) [4] approach, and then approximates the joint pdf of the TDOA (from all microphone pairs) by an empirical distribution. The angular error of the resulting filter is 69% lower than that of a UKF [2] and up to 35% lower than that of the particle filter approach from [5]. This result was obtained on a real corpus [8], with a quickly moving human speaker in a meeting room.

In the remaining part of this paper, we proceed by briefly reviewing the MH-GMF from [9], in Section 2. This is followed by an explanation of how the MH-GMF can be applied to source localization, in Section 3, as well as a presentation of experimental results, in Section 4.

2. MULTIPLE HYPOTHESIS FILTER

The problem of tracking a time-varying system state x_t based on a sequence $y_{1:t} = \{y_1, \ldots, y_t\}$ of corresponding observations is usually formulated as a Bayesian estimation problem in which

- Step 1: A process model $x_t = f(x_{t-1}, v_t)$ is used to construct a prior $p(x_t|y_{1:t-1})$ for the state estimation problem at time t.
- Step 2: The joint predictive distribution $p(x_t, y_t|y_{1:t-1})$ of state and observation is constructed according to a measurement model $y_t = h(x_t, w_t)$ with measurement noise w_t.
- Step 3: The posterior distribution $p(x_t|y_{1:t})$ is obtained by conditioning the joint predictive density $p(x_t, y_t|y_{1:t-1})$ on the realized (actually measured) observation $Y_t = y_t$.

The first step is accomplished by transforming the joint random variable of the last state X_{t-1} and process noise V_t according to f: $X_t = f(X_{t-1}, V_t)$. In step 2, the joint distribution of X_t and Y_t is constructed by transforming (X_t, W_t) according to the augmented measurement function $\tilde{h}$ [10]:

$$\begin{bmatrix} X_t \\ Y_t \end{bmatrix} = \tilde{h}\left(\begin{bmatrix} X_t \\ W_t \end{bmatrix}\right) \quad \text{with} \quad \tilde{h}\left(\begin{bmatrix} x_t \\ w_t \end{bmatrix}\right) \triangleq \begin{bmatrix} x_t \\ h(x_t, w_t) \end{bmatrix}.$$

This work was supported by the European Union through the Marie-Curie Initial Training Network (ITN) SCALE (Speech Communication with Adaptive LEarning, FP7 grant agreement number 213850); and by the Federal Republic of Germany, through the Cluster of Excellence for Multimodal Computing and Interaction (MMCI).

Both these transformations can generally be performed with the fundamental transformation law of probability. A particularly simple case, however, occurs if f, h are linear and V_t, W_t are Gaussian. In this case, all the involved random variables remain Gaussian at all times and the posterior can be obtained as a conditional Gaussian distribution [10]. This analytical closed form solution is generally known as the *Kalman filter*.

2.1. Handling Multiple Observations

The Kalman filter was designed to receive a single observation y_t at time t. In many applied tracking scenarios, however, there are several (K) potential observation candidates $y_t = \{y_t^1, \ldots, y_t^K\}$ available, some of which may be due to the object of interest, some of which may be due to clutter (noise, reverberation). This problem is typically treated by taking the single most likely observation or by combining multiple observations in a weighted sum, as it is done in the *probabilistic data association filter* (PDAF) [11]. In [9], we have presented an alternative to these approaches. It treats the multiple observation problem by a) splitting each Kalman filter at time t into K filters; b) assigning each of the resulting filters to one of the observations; and then c) updating them according to the conditioning step (Step 3 in Section 2), as illustrated in Figure 1. In order to integrate the K resulting conditional distributions $p(x_t|y_{1:t-1}, y_t^k)$ in one posterior, $p(x_t|y_{1:t})$ can be written as a marginal distribution of $p(x_t, k|y_{1:t})$, which, when further expanded under use of $p(x_t, k|y_{1:t}) = p(x_t|k, y_{1:t})p(k|y_{1:t})$, gives:

$$p(x_t|y_{1:t}) = \sum_{k=1}^{K} \underbrace{p(x_t|y_t^k, y_{1:t-1})p(k|y_{1:t})}_{=p(x_t,k|y_{1:t})}. \qquad (1)$$

This is a Gaussian mixture distribution in which the indiviudal posteriors $p(x_t|y_t^k, y_{1:t-1}) = p(x_t|k, y_{1:t})$ constitute Gaussian distributions and in which the $p(k|y_{1:t})$ constitute the corresponding weights. The latter can be obtained with Bayes rule:

$$p(k|y_{1:t}) = \frac{p(y_t|k, y_{1:t-1})\gamma_t^k}{\sum_{k'=1}^{K} p(y_t|k', y_{1:t-1})\gamma_t^{k'}} \qquad (2)$$

where $\gamma_t^k = p(k|t)$ denotes the prior observation probability, which accounts for the confidence or certainty that we put into the k-th observation (similar as motivated in [5]). The $p(y_t|k, y_{1:t-1}) = p(y_t^k|y_{1:t-1})$ are observation likelihoods, which can be evaluated by marginalizing the joint predictive distribution $p(x_t, y_t|y_{1:t-1})$ from step two of the Kalman filter with respect to x_t.

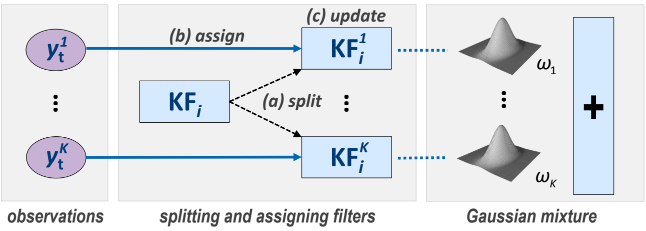

Fig. 1. *Handling multiple observations with a Kalman filter (KF_i).*

2.2. Integration into the Gaussian Mixture Filter Framework

After treating the multiple observation problem as proposed above, we have a Gaussian mixture filtering density. This can be handled by maintaining a bank of Kalman filters which are operating in parallel [9]. As each of the filters is split into K filters at each time t, the number of Gaussian components in general grows exponentially in time. Hence, we reduce the number of mixture components after each iteration by merging Gaussians successively in pairs [9].

Contrary to other tracking frameworks, the MH-GMF treats each observation independently, and assigns weights reflecting the "importance" of the observations in the updated Gaussian mixture. In doing so, this filter allows us to propagate the observations uncertainty to the tracking stage, as well as incorporating the individual information introduced by each observation. In the following, we propose to apply this filter to the acoustic source tracking problem, as we propose a sampling scheme, which captures the uncertainty of the TDOA estimates and propagates it to the tracking stage. We proceed by elaborating on how source localization can be performed with a single KF in Section 3.1 and Section 3.2. Section 3.3 finally presents the multiple observation estimation approach, and how it is integrated into the MH-GMF from Section 2.

3. MH-GMF APPLIED TO SOURCE LOCALIZATION

The arrival of sound waves at a microphone array introduces time differences between the individual sensor pairs. This happens in dependence of the angle of arrival – that is, the azimuth θ and elevation ϕ – as well as the positions m_i, $i = 1, \ldots, M$ of the microphones. Under the far field assumption, in which the distance of the source from the microphones is neglected, the TDOA at the n-th sensor pair $n = \{m_i, m_h\}$ with $i \neq h$, can be calculated as:

$$\tau_n\left(d[\theta, \phi]\right) = \frac{d[\theta, \phi]^T(m_i - m_h)}{c} \qquad (3)$$

where c denotes the speed of sound and where $d[\theta, \phi]$ denotes the direction of arrival $\big[\cos(\phi)\sin(\theta),\ \cos(\phi)\cos(\theta),\ \sin(\phi)\big]^T$. Source localization approaches may use these time differences by either

(a) constructing a spatial filter (beamformer), which scans all possible source locations, and then taking that position where the signal energy is maximized [4].

(b) using a two stage approach, which consists in first estimating the TDOAs of all considered microphone pairs and then inferring the most likely source position [2, 3].

3.1. GCC-Based TDOA Estimation

The most popular approach to estimate the TDOA of a microphone pair $n = \{m_i, m_h\}$ is the generalized cross-correlation with Phase Transform (PHAT) weighting [1]. This approach is based on calculating the correlation of the signals $s_i(t)$ and $s_h(t)$, which have been received at the microphones, according to:

$$\mathcal{R}_n(\tau) = \frac{1}{2\pi}\int_0^{2\pi} \frac{S_i(\omega)S_h^*(\omega)}{|S_i(\omega)S_h^*(\omega)|} e^{j\omega\tau} \mathrm{d}\omega \qquad (4)$$

where $S_i(\omega)$ and $S_h(\omega)$ denote the short-time Fourier transforms of $s_i(t)$ and $s_h(t)$, respectively, and where $\mathcal{R}_n$ is their weighted cross correlation. Subsequently, the most "likely" TDOA is extracted as:

$$\widehat{\tau}_n = \mathrm{argmax}_\tau\, \mathcal{R}_n(\tau) \qquad (5)$$

3.2. Acoustic Source Tracking Based on Estimated TDOAs

Once the TDOA has been estimated for a number of $N \leq \binom{M}{2}$ microphone pairs, source localization can be performed with a Kalman filter, as described in [2, 3]. In order to do this, we use the following process model for tracking the azimuth θ and elevation ϕ of the source:

$$\begin{bmatrix}\theta_t\\ \phi_t\end{bmatrix} = f\left(\begin{bmatrix}\theta_{t-1}\\ \phi_{t-1}\end{bmatrix}, v_t\right) = \begin{bmatrix}\theta_{t-1} + v_{t,\theta}\\ \phi_{t-1} + v_{t,\phi}\end{bmatrix} \tag{6}$$

where $v_{t,\theta}$ and $v_{t,\phi}$ denote zero-mean Gaussian process noise with a variance of σ_θ^2 and σ_ϕ^2, respectively. Similar to the approaches taken in [2, 3, 5], we use

$$y_t = h\left(\begin{bmatrix}\theta_t\\ \phi_t\end{bmatrix}, \mathbf{w}_t\right) = \begin{bmatrix}\tau_1\left(d[\theta_t, \phi_t]\right) + w_{t,1}\\ \vdots\\ \tau_N\left(d[\theta_t, \phi_t]\right) + w_{t,N}\end{bmatrix} \tag{7}$$

as a measurement model. In this equation, $\tau_n\left(d[\theta_t, \phi_t]\right)$ denotes the predicted TDOA of the n-th microphone pair, with $n = 1, \ldots, N$, whereas $w_{t,n}$ is zero-mean Gaussian measurement noise with a variance of σ_W^2. This measurement model is nonlinear. Hence the use of an extension of KF is required, we propose to use the UKF, similar as it was originally proposed in [2], but as a single observation filter.

3.3. Applying The Multiple Hypothesis Gaussian Mixture Filter

In the Kalman filtering approach from [2, 3], the most likely TDOA is determined individually for each microphone pair. These individual TDOA estimates are subsequently combined to form a joint measurement $y_t = [\hat{\tau}_1, \ldots, \hat{\tau}_N]$. The error is assumed to follow a Gaussian distribution [2, 3]. This assumption may be true under ideal conditions. In practice, however, the errors in the GCCs (i.e. measurement errors) can be expected to have a multimodal distribution, due to reflections, reverberation and background noise [5]. Hence, we here propose to

1. consider a larger number of observation candidates (hypotheses) y_t^k with associated confidence weights γ_t^k.
2. process these weighted observations with the multiple hypothesis Gaussian mixture filter (MH-GMF) from Section 2, with the KFs being replaced by UKFs.

The aim of this procedure is to propagate the uncertainty from the detection (TDOA estimation) to the tracking stage, by choosing the weighted observation candidates in such a fashion that they capture the observation uncertainty in the GCCs. Let us first consider the observation space $\mathcal{Y}$ (does not depend on time), which can be approximated by the Cartesian product of all possible TDOAs from N different microphone pairs :

$$\mathcal{Y} = \left\{y^1, \ldots, y^K\right\} \triangleq \bigtimes_{n=1}^{N} \left\{-\tau_n^{\max}, \ldots, \tau_n^{\max}\right\} \tag{8}$$

where $y^k = \left[\tau_1^k, \ldots, \tau_N^k\right]$ with $k = 1, \ldots, K$. $\tau_n^{\max}$ denotes the maximum TDOA of microphone pair n and K is the cardinality of $\mathcal{Y}$. Then, interpreting the GCC as a likelihood function (as done in [7] for the SRP) and further assuming that errors in the GCCs are statistically independent [5], the confidence or prior observation likelihood of a particular combination y^k can be calculated as the product of the individual GCC values $R_n(\tau_n^k)$:

$$\gamma_t^k = \prod_{n=1}^{N} \tilde{R}_n(\tau_n^k) \;\text{ with }\; \tilde{R}_n(\tau) \triangleq \frac{R_n(\tau)}{\sum_{\tau'} R_n(\tau')} \tag{9}$$

where the division by $\sum_{\tau'} R_n(\tau')$ normalizes the total probability to 1. This gives us the following observation density:

$$p_{\text{measured}}(y_t) = \sum_{k=1}^{K} \gamma_t^k \delta\left(y_t - y^k\right) \tag{10}$$

where the y^k and γ_t^k are given by (8) and (9), respectively. As a next step, we could now pass this density to the multiple hypothesis filter from Section 2. But, considering the fact that the Cartesian product results in $K = \prod_{n=1}^{N}\left(2\tau_n^{\max} + 1\right)$ different combinations, this approach has to be dismissed as intractable. Hence, we reduce the number of observations by approximating (10) through a sampling scheme, which samples observations from high likelihood regions of the observation space.

Sampling from the GCCs: In order to obtain a set $\{y_t^1, \ldots, y_t^{K'}\}$ of $K' << K$ observations from (10), we first draw K' TDOA from each normalized GCC $\tilde{R}_n$ (through multinomial sampling) and then combine the resulting τ_n^k (from different sensor pairs) to form K' observations $y_t^k = \left[\tau_1^k, \ldots, \tau_N^k\right]$. As a result of sampling, the weights γ_t^k all need to be set to $1/K'$. In particular, note that the use of sampling ensures that we draw more TDOAs from regions of high likelihood (GCC peaks) and less TDOAs from regions of low likelihood (GCC valleys). So, we statistically focus on combinations y_t^k where the observation probability is high.

Proximate Consistency: The main drawback of the above sampling technique is that the TDOAs τ_n^k of different microphone pairs are independently drawn from the GCCs. This may lead to inconsistent observations, i.e. TDOA combinations $\left[\tau_1^k, \ldots, \tau_N^k\right]$ which do not correspond to a physically possible location. In order to alleviate this problem, the filter's predicted observation likelihood $p(y_t^k|y_{1:t})$ can be used as an approximate measure of consistency. This motivates the idea of combining the independently drawn τ_n^k in such a fashion that the total observation likelihood is maximized. In this work, we use a greedy approach which 1) selects from each sampled set $\left\{\tau_n^1, \ldots, \tau_n^{K'}\right\}$ the $\tau_n^{k_n}$ with the highest projected observation likelihood $p(\tau_n^{k_n}|y_{1:t-1})$; 2) combines these samples to form the observation $y_t^k = [\tau_1^{k_1}, \ldots, \tau_N^{k_N}]$, and finally, 3) removes the $\tau_n^{k_n}$ from the respective sample sets. This procedure is repeated until all samples are combined.

Voice Activity Detection and Gating: As there is no point in tracking an inactive speaker, we use a voice activity detector [12] for suppressing observations during silence frames. As a further precaution against outliers, we extend the above sampling scheme through the integration of gating [11]. This is achieved by 1) merging all the predicted observation densities of the Gaussian mixture filter into a single Gaussian $p(y_t|y_{1:t-1}) = \mathcal{N}(y_t; \mu, \Sigma)$; 2) defining a gating area $\mathcal{G}_n \triangleq \left\{\tau_n|(\tau_n - \mu_n)^2/\Sigma_{n,n} \leq \psi\right\}$ for each sensor pair n; and then 3) sampling the TDOAs τ_n^k from the "gated" pdf

$$\bar{R}_n(\tau_n) = \frac{R_n(\tau_n) \cdot I_{\mathcal{G}_n}(\tau_n)}{\sum_{\tau'=-\tau_{\max}}^{\tau_{\max}} R_n(\tau') \cdot I_{\mathcal{G}_n}(\tau_n)} \tag{11}$$

In these equations, ψ denotes the gating threshold and $I_{\mathcal{G}_n}(\tau_n)$ denotes the indicator function, which is 1 if $\tau_n \in \mathcal{G}_n$ and 0 otherwise.

4. EXPERIMENTS AND RESULTS

In order to evaluate the performance of the proposed algorithm, we performed a set of tracking experiments on the AV16.3 corpus [8]. This corpus consists of real human speakers, which were recorded in a smart meeting room (approximately 30m^2 in size), using a 20cm 8-channel circular microphone array. The real mouth position is known within an error $\leq$ 5cm [8]. In this work, we perform experiments on two different sequences: The highly non-stationary se-

Sequence “seq11-1p-0100” / quickly moving

tracking algorithm	root mean square error			real-time factor
	azimuth	elevation	DOA	
UKF	5.56°	15.98°	16.92°	0.336
PF	4.80°	10.33°	11.40°	0.374
UKF + Gating	4.17°	7.12°	8.24°	0.329
MH-PF	3.72°	5.94°	7.00°	0.582
MH-GMF	2.85°	4.25°	5.11°	0.664

Sequence “seq02-1p-0000” / more stationary

tracking algorithm	root mean square error			real-time factor
	azimuth	elevation	DOA	
UKF	8.66°	19.28°	21.14°	0.410
PF	7.54°	19.57°	20.98°	0.432
UKF + Gating	2.71°	8.14°	8.58°	0.329
MH-PF	3.99°	6.44°	7.58°	0.680
MH-GMF	2.71°	4.07°	4.89°	0.793

Table 1. *Average root mean square error (RMSE) in azimuth, elevation and direction of arrival (DOA), with respect to the center of the array. Results are shown under use of the unscented Kalman filter (UKF) [2], a standard sequential importance resampling (SIR) particle filter (PF), the UKF with gating [3, 11], the particle filter (MH-PF) from [5] and the proposed multiple hypothesis Gaussian mixture filter (MH-GMF) from Section 3.3. The last column shows the real-time factor, i.e. the processing time divided by the duration of the input.*

quence “seq11-1p-0100” in which a single speaker is quickly moving in the room; and the relatively stationary sequence “seq02-1p-0000” in which a speaker is moving through 16 predefined locations while uttering one sentence at each of the positions. These sequences are 32 and 185 seconds in length, respectively. The average distance of the speaker from the array is 1.18 and 1.53 meters, with a minimum of 0.57 and a maximum of 2.40. In the experiments which are reported below, all the GCC functions were calculated under use of PHAT [1] weighting. The window length was 1024 samples (64ms). GCC interpolation did not improve the results. The number of used observations (K') was 20 for the MH-GMF. The particle filters were using 100 particles (a larger number did not improve the results).

The results in Table 1 show that the proposed multiple hypothesis Gaussian mixture filter performs significantly better than any of the other methods. Its angular error (DOA) is 69% and 79% lower than that of the UKF [2]; 38% and 43% lower than that of the UKF with Gating [11]; and still 27% and 35% lower than that of the MH-PF from [5]. Regarding these results, it should be noted that the main problem of the small aperture microphone arrays, with a planar geometry, consists in obtaining good estimates of the elevation. But, having a closer look at Table 1, we can see that it is exactly where our method shows its true strength. Regarding the anomaly rate (AR) [13], which represents the percentage of estimates with an error $\geq 5^\circ$, we obtained an AR of 4.79% and 8.34% for the quickly moving sequence “seq11-1p-0100” as well as the more stationary sequence “seq02-1p-0000”. On the sequence “seq01-1p-0000”, we obtained an AR of 8.50%, in comparison to an AR of 30.43% which the authors of [13] reported for multi-channel cross correlation (MCCC) based localization. This result shows that speaker tracking algorithms can significantly improve the RMSE, as they smooths out the erroneous estimates.

In terms of real-time implementation, the time factors in Table 1, show that all methods run faster than real-time on a standard Intel i7-2600K CPU clocked at 3.4GHz. The plain UKF is roughly 2 times faster than the proposed MH-GMF; and that although the latter runs more than 20 UKFs in parallel. This indicates that most of the computation time is spent in the generalized cross correlation.

5. CONCLUSIONS

We have presented a new multiple hypothesis Gaussian mixture filter for acoustic source localization and tracking. It counteracts the problem of TDOA estimation error by propagating the estimation uncertainty rather than passing a point estimate. This approach is justified in room acoustical environments where the presence of reverberation and noise smears and changes the GCC function. We plan to extend the proposed MH-GMF to multiple speaker tracking.

6. REFERENCES

[1] C. H. Knapp and G. C. Carter, “The generalized correlation method for estimation of time delay,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, 1976.

[2] S. Gannot and T. G. Dvorkind, “Microphone array speaker localizers using spatial-temporal inforamtion,” *EURASIP Journal on Applied Signal Processing*, Article 59625, 2006.

[3] U. Klee, T. Gehrig, and J. McDonough, “Kalman filters for time delay of arrival-based source localization,” *EURASIP Journal on Applied Signal Processing*, Article 12378, 2006.

[4] J. H. DiBiase, *A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays*, Ph.D. thesis, Brown University, 2000.

[5] J. Vermaak and A. Blake, “Nonlinear filtering for speaker tracking in noisy and reverberant environments,” in *Proc. ICASSP*, May 2001, vol. 5, pp. 3021–3024.

[6] A. Levy, S. Gannot, and E. A. P. Habets, “Multiple-hypothesis extended particle filter for acoustic source localization in reverberant environments,” *IEEE Trans. Acoust., Speech, Language Process.*, vol. 19, no. 6, pp. 1540–1555, 2011.

[7] D. B. Ward and R. C. Williamson, “Particle filter beamforming for acoustic source localization in a reverberant environment,” in *Proc. ICASSP*, May 2002, vol. 2, pp. 1777–1780.

[8] G. Lathoud, J.-M. Odobez, and D. Gatica-Perez, “AV16.3: An audio-visual corpus for speaker localization and tracking,” in *Proc. MLMI 04 Workshop*, May 2006, pp. 182–195.

[9] F. Faubel, M. Georges, B. Fu, and D. Klakow, “Robust Gaussian mixture filter based mouth tracking in a real environment,” in *Proc. Visual Computing Research Conference (IVCI, Saarbrucken)*, Dec. 2009.

[10] F. Faubel and D. Klakow, “A transformation-based derivation of the Kalman filter and an extensive unscented transform,” in *Proc. SSP*, Sept. 2009, pp. 161–164.

[11] Y. Bar-Shalom and T. E. Fortmann, *Tracking and Data Association*, Academic Press, 1988.

[12] J. Sohn, N. S. Kim, and W. Sung, “A statistical model-based voice activity detection,” *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, 1999.

[13] J. Dmochowski, J. Benesty, and S. Affes, “The generalization of narrowband localization methods to broadband environments via parametrization of the spatial correlation matrix,” in *Proc. EUSIPCO*, Sept. 2007, pp. 763–767.