IDIAP RESEARCH REPORT

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

Youssef Oualil Friedrich Faubel Dietrich Klakow

Idiap-RR-09-2012

MARCH 2012

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

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, Martigny, Switzerland

ABSTRACT

In this work, we address the problem of tracking an acoustic source based on measured time differences of arrival (TDOAs). 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 be estimated reliably; and (2) that the 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 propagating the detection uncertainty. That is achieved by sampling from the GCCs and then integrating the resulting TDOAs 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 *time difference of arrival* (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 multi-path 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] and Gehrig [6] proposed the use of a multiple hypothesis particle filter and a probabilistic data association filter, respectively.

In this work, we continue along the lines of [5, 6] by propagating the uncertainty in the TDOA estimates to the tracking stage. This is achieved by interpreting the GCC as a likelihood function of the TDOAs, similar as originally proposed in [5] and firstly applied in [7] for a steered response power (SRP) [4] approach. Ideally, we would now like to (1) use all possible TDOA combinations from different sensor pairs, weighted with the respective GCC values; and then (2) pass these combinations to a multiple hypothesis Gaussian mixture filter (MH-GMF) [8], as observations. As the Cartesian product of all TDOA combinations is computationally intractable, we proceed by approximating the combined likelihood function over the TDOAs as an empirical distribution of around 100 sampled observations. These samples give a compact point mass representation of the GCCs, which can now reasonably be processed with a MH-GMF. The main innovation is a Monte Carlo scheme which creates the observations by first drawing TDOAs from the individual GCCs and then combining the TDOAs in a "proximately consistent" fashion. The angular error of the resulting filter is 69% lower than that of a UKF [2] and up to 55% lower than that of the particle filter approach from [5]. This result was obtained on a real corpus [9], with a quickly moving human speaker in a meeting room.

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).

In the remaining part of this paper, we proceed by briefly reviewing the MH-GMF from [8], 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

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.
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.
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]:

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

Both these transformations can generally be performed with the fundamental transformation law of probability. A particular 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* (KF).

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 observations 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) [6, 11]. In [8], we have presented an alternative to these approaches. It treats the multiple observation problem by (1) splitting each Kalman filter at time t into K filters; (2) assigning each of the resulting filters to one of the observations; and then (3) updating them according to the conditioning step (the third step) from the previous page. 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:

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

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:

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

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.

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 [8]. 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 [8].

3 Application to Source Localization

The arrival of sound waves at an array of microphones 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 (i, j)-th sensor pair (m_i, m_j) can be calculated as:

$$\tau_{i,j}\left(p[\theta,\phi]\right) = \frac{p[\theta,\phi]^T (m_j - m_i)}{c} \tag{3}$$

where c denotes the speed of sound and where $p[\theta, \phi]$ 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].

As our approach falls into the second category we proceed by first explaining TDOA estimation, in Section 3.1, and then elaborate on how source localization can be performed with a Kalman filter (Section 3.2). Section 3.3 finally presents the proposed multiple observation approach, which integrates these two stages more closely by passing the uncertainty from the detection (i.e. TDOA estimation) to the tracking stage, through use of the MH-GMF from Section 2.

Regarding the above two categories, it is worth mentioning that there is a large variety of other approaches, including multi-channel cross correlation [12], sub-space approaches [12], combinations of the SRP with particle filters [7], and many more.

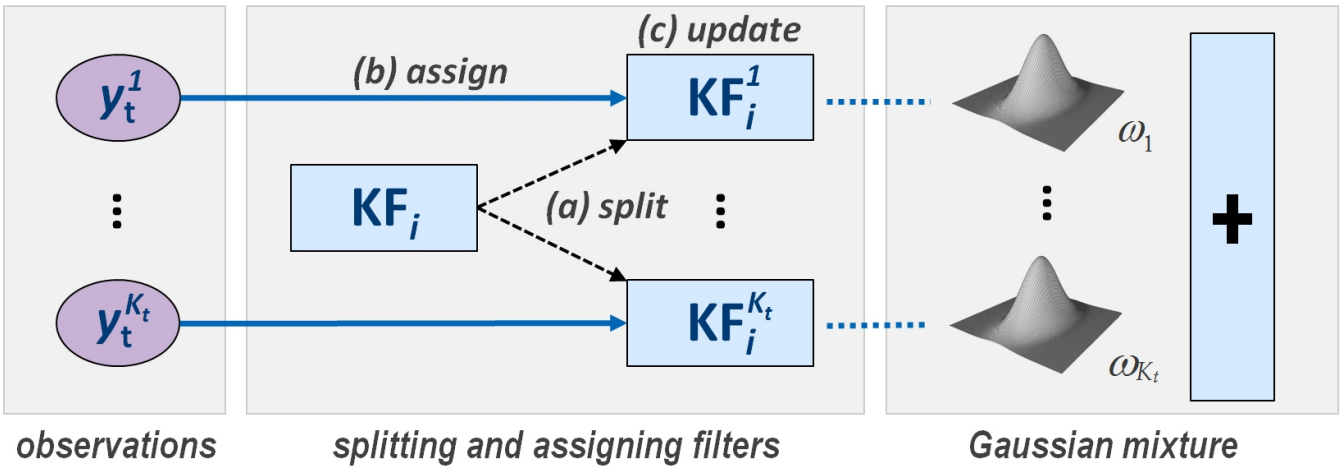

Fig. 1. *Handling multiple observations with a Kalman filter (KF_i).*

3.1 GCC-Based TDOA Estimation

The most popular approach to estimate the TDOA between two microphones m_i and m_j is to use the generalized cross-correlation (GCC) with PHAT weighting [1]. This approach is based on calculating the correlation of the signals $s_i(t)$ and $s_j(t)$, which have been received at the microphones, according to:

$$\mathcal{R}_{i,j}(\tau) = \frac{1}{2\pi} \int_0^{2\pi} \frac{S_i(w) S_j^*(\omega)}{\left| S_i(\omega) S_j^*(w) \right|} e^{j\omega\tau} \mathrm{d}\omega \tag{4}$$

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

$$\widehat{\tau}_{i,j} = \mathrm{argmax}_\tau \, \mathcal{R}_{i,j}(\tau) \tag{5}$$

3.2 Acoustic Source Tracking Based on Estimated TDOAs

Once the TDOA has been estimated for a number of N 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

$$y_t = h\left(\begin{bmatrix} \theta_t \\ \phi_t \end{bmatrix}, \mathbf{w}_t \right) = \begin{bmatrix} \tau_{i_1,j_1}\left(p[\theta_t, \phi_t]\right) + w_{t,1} \\ \vdots \\ \tau_{i_N,j_N}\left(p[\theta_t, \phi_t]\right) + w_{t,N} \end{bmatrix} \tag{7}$$

as a measurement model. In this equation, $\tau_{i_n,j_n}\left(p[\theta_t, \phi_t]\right)$ denotes the predicted TDOA of the n-th microphone pair (i_n, j_n), 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, as the calculation of the predicted TDOAs according to (3) involves evaluating sines and cosines for the direction of arrival $p[\theta_t, \phi_t]$. Hence, we use an unscented Kalman filter (UKF), as originally proposed in [2].

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,

$$y_t = \left[\hat{\tau}_1, \ldots, \hat{\tau}_N \right] \text{ with } \hat{\tau}_n = \hat{\tau}_{i_n,j_n}.$$

The error is assumed to follow a Gaussian distribution [3, 2]. 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. In order to achieve this, let us first consider the Cartesian product of all possible TDOAs from N different microphone pairs (m_{i_n}, m_{j_n}):

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

with $y^k = \left[\tau_1^k, \ldots, \tau_N^k\right]$ and with $\tau_{\max}$ denoting the maximum TDOA. Note that the y here do not have a subscript t as they are theoretical combinations, which are independent of time. Then, interpreting the GCC as a likelihood function (as done in [7] for the SRP) and further assuming that the 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 $R_{i_n,j_n}(\tau_n^k)$:

$$\gamma_t^k = \prod_{n=1}^{N} \tilde{R}_{i_n,j_n}(\tau_n^k) \;\text{ with }\; \tilde{R}_{i_n,j_n}(\tau) \triangleq \frac{R_{i,j}(\tau)}{\sum_{\tau'} R_{i,j}(\tau')} \tag{9}$$

where the division by $\sum_{\tau'=-\tau_{\max}}^{\tau_{\max}} R_{i,j}(\tau')$ 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 $K = (2\tau_{\max}+1)^N$ different combinations, this approach has to be dismissed as intractable. Hence, we reduce the number of observations by approximating (10) through sampling.

Sampling from the GCCs: In order to obtain a set $\{y_t^1, \ldots, y_t^{K'}\}$ of $K' << K$ samples from (10), we first draw K' TDOAs from each normalized GCC $\tilde{R}_{i_n,j_n}$ (interpreted as a pdf); and then combine the resulting τ_n^k to 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 $1/K'$.

Justification for Sampling: Random 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 (see Monte Carlo methods in general).

Voice Activity Detection and Gating: As there is no point in tracking an inactive speaker, we use a voice activity detector [13] for suppressing observations during silence frames. As a further precaution against outliers, the above sampling scheme is extended through the integration of gating [11]. This is achieved by (1) merging all the predicted observation densities of the Kalman filters to a single Gaussian $p(y_t|y_{1:t-1}) = \mathcal{N}(y_t; \mu, \Sigma)$; (2) defining a gating area $\mathcal{G}_n \triangleq \left\{\tau_n | (\tau_n - \mu_n)^2/\Sigma_{n,n} \leq T\right\}$ for each sensor pair (i_n, j_n); and then (3) sampling the TDOAs τ_n^k from the "gated" pdf

$$\bar{R}_{i_n,j_n}(\tau_n) = \frac{R_{i_n,j_n}(\tau_n) \cdot I_{\mathcal{G}_n}(\tau_n)}{\sum_{\tau'=-\tau_{\max}}^{\tau_{\max}} R_{i_n,j_n}(\tau') \cdot I_{\mathcal{G}_n}(\tau_n)}.$$

In these equations, T 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.

Proximate Consistency: The above sampling scheme consists of drawing the τ_n^k independently from the GCCs of different microphone pairs. This gave good results in practice. But it can also create inconsistent observations. By that we mean observations y_t^k for which the chosen combination $y_t^k = \left[\tau_1^k, \ldots, \tau_N^k\right]$ of TDOAs does not correspond to a physically possible location. In order to tackle this problem, the filter's predicted observation likelihood $p(y_t^k|y_{1:t})$ can be used as an approximate measure for 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 $\{\tau_n^1, \ldots, \tau_n^{K'}\}$ that $\tau_n^{k_n}$ with the highest projected observation likelihood $p(\tau_n^{k_n}|y_{1:t-1})$; (2) combines these samples to an observation $y_t^k = [\tau_1^{k_1}, \ldots, \tau_N^{k_N}]$; (3) removes the $\tau_n^{k_n}$ from the respective sample sets and (4) repeats this procedure until all samples are combined.

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 [9]. In this corpus, real human speakers were recorded in a normal meeting room (approximately 30m^2 in size) with a 20cm 8-channel circular microphone array. The real mouth position is known with an error of $\leq$ 1.2cm [9]. Table 1 shows the results for two different sequences of this corpus: the highly non-stationary sequence "seq11-1p-0100", in which a single speaker is quickly moving in the room;

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 [11, 3], 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.*

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; and they can be looked at under http://www.glat.info/ma/av16.3/session09/seq11-1p-0100 and http://www.glat.info/ma/av16.3/session09/seq02-1p-0000. 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.

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 AV16.3 task is to get the elevation right (the authors of [12] even claim it is too hard to estimate and, hence, do not report any numbers). But, having a closer look at Table 1, we find that it is exactly here where our method shows its true strength. In order to compare the above experiments to other results that have been reported in the literature, we also give numbers under the same conditions (same corpus, sequence and evaluation scheme) which have been used in [12]. Here, we get an RMSE of 1.41°for the azimuth in comparison to 1.66°for the SRP [12] – when matching the anomaly rate (AR) of 35.27% from [12] for reasons of comparability. At a matched AR of 30.43%, we get an RMSE of 1.60°in comparison to 1.84°for the MCCC [12]. Without matching, the AR of the MH-GMF is 15.54%. The discrepancy to Table 1 can be explained by the fact that the authors of [12] discard DOAs with an error $\geq 5°$ as "anomalies" and exclude them from the RMSE.

Now having a look at the real-time factors in Table 1, we find that all the considered 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 100 UKFs in parallel. This indicates that most of the computation time is spend in the generalized cross correlation (GCC). In particular note that the GCCs were all 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 100 for the MH-GMF. The particle filters were using 100 particles (a larger number did not improve the results).

5 Conclusions

We have presented a new multiple hypothesis Gaussian mixture filter for acoustic source localization and tracking. It reduces the problem of erroneous TDOA estimates by propagating the uncertainty of the TDOAs 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] T. Gehrig, U. Klee, J. McDonough, S. Ikbal, M. Wölfel, and C. Fügen, "Tracking and beamforming for multiple simultaneous speakers with probabilistic data association filters," in *Proc. Interspeech*, Sept. 2006.

[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] 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.

[9] 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.

[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. 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.

[13] 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.