# AN ADAPTIVE LEVEL OF DETAIL APPROACH TO NONLINEAR ESTIMATION

*Friedrich Faubel, Dietrich Klakow*

Spoken Language Systems,
Saarland University, D-66123 Saarbrücken, Germany
{friedrich.faubel, dietrich.klakow}@lsv.uni-saarland.de

## ABSTRACT

In this work, we present a general method for approximating nonlinear transformations of Gaussian mixture random variables. It is based on transforming the individual Gaussians with the unscented transform. The level of detail is adapted by iteratively splitting those components of the initial mixture that exhibited a high degree of nonlinearity during transformation. After each splitting operation, the affected components are re-transformed. This procedure gives more accurate results in cases where a Gaussian fit does not well represent the true distribution. Hence, it is of interest in a number of signal processing fields, ranging from nonlinear adaptive filtering to speech feature enhancement. In simulations, the proposed approach achieved a 48-fold reduction of the approximation error, compared to a single unscented transform.

***Index Terms***— Nonlinear estimation, Adaptive estimation, Approximation methods, Gaussian distributions

## 1. INTRODUCTION

Many signal processing tasks require nonlinear transformations of one or several random variables $X_1, \ldots, X_m$. For that, it is of interest to know the distribution of the transformed random variable $Y = f(X_1, \ldots, X_m)$, which can seldom be obtained in analytic fashion. Hence, various approximations have been proposed. The most prominent one is probably the linear Gaussian approximation, which assumes firstly, that the joint distributions of the $X_i$ can be well-approximated by a Gaussian and secondly, that the function $f$ is reasonably linear about the joint mean of the original distributions. Then, "locally" linearizing $f$ about the mean results in a Gaussian approximation of the transformed random variable. Examples for this approach are the extended Kalman filter as well as the vector Taylor series approach for environment-independent speech recognition [1]. Van der Merwe and Wan [2] proposed a different solution to the transformation problem. It consists of drawing samples from the original distribution; individually transforming the samples; and then approximating the transformed distribution by a parametric fit. This is more exact. However, it is also more expensive from a computational point of view.

In this paper, we follow [3] in that we approximate the original distribution of $X_1, \ldots, X_m$ by a mixture of Gaussians. The idea is to choose the variances small enough for the local linearizations to be approximately valid for each of the Gaussians. Then, performing locally linearized transformations of the individual Gaussians yields a more accurate approximation of the distribution of $Y$. We amend this approach by introducing a method that adapts the level of detail of the Gaussian mixture representation to the nonlinearities present in the transformation. By that we mean keeping more Gaussians in regions where the degree of nonlinearity is high and fewer Gaussians in regions where it is low. This is achieved by generalizing the splitting approach, described in [4] for the specific case of a Gaussian mixture filter, to general transformations of Gaussian mixture random variables. Deviating from [4], the measure of nonlinearity is derived in a more well-founded way, based on the coefficient of determination used in linear regression. In addition to that, we show how stacked (augmented) variables can be treated and how the splitting priority of a Gaussian can be determined as geometric interpolation between degree of nonlinearity and mixture weight.

The remaining part of this paper is organized as follows. In the upcoming section we describe the unscented transform and everything related to it. In Section 3, we establish the adaptive level of detail approach, which is finally evaluated in Section 4.

This work was supported by the Federal Republic of Germany through the German Research Foundation (DFG), under the research training network IRTG 715 "Language Technology and Cognitive Systems".

## 2. THE UNSCENTED TRANSFORM

The unscented transform was introduced by Julier and Uhlmann [5] in order to approximate a nonlinear transform $Y = f(X)$ of an $n$-dimensional Gaussian random variable $X$, based on a point-mass representation that, by design, captures the first two moments of the distribution. Let $p_X(x)$ denote the distribution of $X$,

$$p_X(x) = \mathcal{N}(x; \mu_X, \Sigma_X)$$

with mean $\mu_X$ and covariance matrix $\Sigma_X$. Further, let $R^T R$ be the Cholesky decomposition of $\Sigma_X$. Then, denoting the rows of $R$ by $R_i$ and defining $\lambda = n + \kappa$ for an arbitrary $\kappa \in \mathcal{R}$, the distribution of $X$ can be represented by the weighted empirical distribution

$$\tilde{p}_X(x) = \sum_{i=0}^{2n} W_i \delta(x - \mathcal{X}_i) \tag{1}$$

where $\delta$ is the Dirac delta and where the points and weights, $\mathcal{X}_i$ and $W_i$, are given by

$$\begin{array}{rclcrcl} \mathcal{X}_0 & = & \mu_X & & W_0 & = & \kappa/\lambda \\ \mathcal{X}_{2i+1} & = & \mu_X + \sqrt{\lambda} R_i & & W_{2i+1} & = & 1/(2\lambda) \\ \mathcal{X}_{2i+2} & = & \mu_X - \sqrt{\lambda} R_i & & W_{2i+2} & = & 1/(2\lambda) \end{array} \tag{2}$$

$i = 0, \ldots, (n-1)$. Note that $\kappa$ specifies how much weight is placed on the mean, $\mathcal{X}_0$. Setting it to $1/2$ results in a weight of $1/n$ for each of the points. Setting it to $3 - n$ minimizes the error in the fourth moment [6]. Similar as in Monte Carlo methods, the weighted points $\mathcal{X}_i$ can be instantiated through the function $f$, $\mathcal{Y}_i = f(\mathcal{X}_i)$, which then in turn yields a weighted empirical distribution of $Y$:

$$\tilde{p}_Y(y) = \sum_{i=0}^{2n} W_i \delta(y - \mathcal{Y}_i). \tag{3}$$

Consequently, a Gaussian approximation $\hat{p}_Y(y) = \mathcal{N}(Y; \hat{\mu}_Y, \hat{\Sigma}_Y)$ of the transformed distribution can be obtained by estimating the mean and covariance of $Y$ in a maximum likelihood fashion:

$$\hat{\mu}_Y = \sum_{i=0}^{2n} W_i \mathcal{Y}_i, \quad \hat{\Sigma}_Y = \sum_{i=0}^{2n} W_i (\mathcal{Y}_i - \mu_Y)(\mathcal{Y}_i - \mu_Y)^T. \tag{4}$$

That is the unscented transform. It is exact for linear transforms. For nonlinear transforms its mean and covariance estimates are accurate up to the second order term of the Taylor series expansion [6].

### 2.1. Degree of Nonlinearity

In the unscented transform, each triple $\{\mathcal{X}_{2i+1}, \mathcal{X}_0, \mathcal{X}_{2i+2}\}$ forms a set of equidistant points on a line. The same holds for the triples of transformed points, $\{\mathcal{Y}_{2i+1}, \mathcal{Y}_0, \mathcal{Y}_{2i+2}\}$, if the transformation $Y = g(X)$ is linear. That motivates the idea of determining the degree of nonlinearity by fitting to each $\{\mathcal{Y}_{2i+1}, \mathcal{Y}_0, \mathcal{Y}_{2i+2}\}$ a set $\{\mathcal{Z}_{2i+1}, \mathcal{Z}_0, \mathcal{Z}_{2i+2}\}$ of equidistant points on a line and then computing the coefficient of determination, $R^2$, a measure for the goodness of fit used in linear regression. Without loss of generality, let

$$\mathcal{Z}_0 = b, \quad \mathcal{Z}_{2i+1} = b + a_i, \quad \mathcal{Z}_{2i+2} = b - a_i.$$

Minimizing the squared error between the $\mathcal{Z}$ and $\mathcal{Y}$ by taking derivatives with respect to $a_i$ and $b$ and then equating them to zero gives $a_i = \frac{1}{2}(\mathcal{Y}_{2i+1} - \mathcal{Y}_{2i+2})$ and $b = \mathcal{Y}_0$. Hence we have

$$\begin{aligned} \mathcal{Z}_0 &= \mathcal{Y}_0 \\ \mathcal{Z}_{2i+1} &= \mathcal{Y}_0 + \tfrac{1}{2}(\mathcal{Y}_{2i+1} - \mathcal{Y}_{2i+2}) \\ \mathcal{Z}_{2i+2} &= \mathcal{Y}_0 - \tfrac{1}{2}(\mathcal{Y}_{2i+1} - \mathcal{Y}_{2i+2}) \end{aligned} \tag{5}$$

for $i = 1, \ldots, n$. The corresponding coefficients of determination are calculated as:

$$R_i^2 = 1 - \frac{\|\mathcal{Y}_0 - \mathcal{Z}_0\|^2 + \|\mathcal{Y}_{2i+1} - \mathcal{Z}_{2i+1}\|^2 + \|\mathcal{Y}_{2i+2} - \mathcal{Z}_{2i+2}\|^2}{\|\mathcal{Y}_0 - \hat{\mu}_{Y_i}\|^2 + \|\mathcal{Y}_{2i+1} - \hat{\mu}_{Y_i}\|^2 + \|\mathcal{Y}_{2i+2} - \hat{\mu}_{Y_i}\|^2},$$

where $\hat{\mu}_{Y_i}$ is the empirical mean, $\frac{1}{3}(\mathcal{Y}_0 + \mathcal{Y}_{2i+1} + \mathcal{Y}_{2i+2})$. It can be shown that $R_i^2$ is always in the range $[0, 1]$. A value close to 1 indicates a good linear fit, or, correspondingly, a low degree of nonlinearity. Hence, we determine the degree of nonlinearity of each triple $\{\mathcal{Y}_{2i+1}, \mathcal{Y}_0, \mathcal{Y}_{2i+2}\}$ as $\eta_i \triangleq 1 - R_i^2$:

$$\eta_i = \frac{\frac{1}{2}\|\mathcal{Y}_{2i+1} + \mathcal{Y}_{2i+2} - 2\mathcal{Y}_0\|^2}{\|\mathcal{Y}_0 - \hat{\mu}_{Y_i}\|^2 + \|\mathcal{Y}_{2i+1} - \hat{\mu}_{Y_i}\|^2 + \|\mathcal{Y}_{2i+2} - \hat{\mu}_{Y_i}\|^2}. \tag{6}$$

The nominator accounts for the variance or modeling error introduced by the linear fit. The denominator normalizes that term by the variance of the points. Dropping the denominator gives an absolute, i.e. non-normalized measure

$$\eta_i' = \frac{1}{2}\|\mathcal{Y}_{2i+1} + \mathcal{Y}_{2i+2} - 2\mathcal{Y}_0\|^2. \tag{7}$$

Both the normalized and non-normalized measure can easily be extended to the extensive unscented transform [7] or to Gauss-Hermite quadrature [8], simply by considering lines in one variable or coefficient and then averaging over all possible values the other variables or coefficients can assume. The total degree of nonlinearity is obtained by averaging over all the $\eta_i$, $i = 0, \ldots, (n-1)$:

$$\eta \triangleq \frac{1}{n}\sum_{i=0}^{n-1} \eta_i \tag{8}$$

If it is close to zero the transformation is approximately linear and the Gaussian approximation of $p_{\mathbf{Y}}(\mathbf{y})$ is justified. For larger values, the parametric Gaussian fit might not well represent the true distribution.

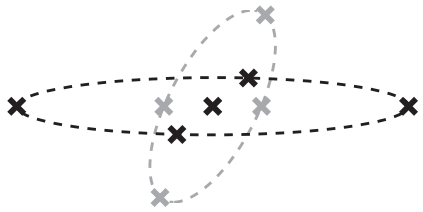


**Fig. 1**. Augmented Unscented Transform. *The picture portrays the points (crosses) used by the augmented unscented transforms along with covariance ellipses (dashed lines). The black crosses indicate the points chosen for $X^{(1)}$, augmented with the mean of $X^{(2)}$. The grey crosses indicate the points chosen for $X^{(2)}$, augmented with the mean of $X^{(1)}$.*

### 2.2. Stacking / Augmentation

Some approaches, such as the unscented Kalman filter [5] or the VTS approach for environment-independent speech recognition [1], necessitate nonlinear transformations $Y = f(X^{(1)}, \ldots, X^{(m)})$ of several Gaussian random variables at a time. This problem can be tackled by stacking, that is by treating the variables as a single, joint Gaussian random variable

$$X = \begin{bmatrix} X^{(1)^T} & \cdots & X^{(m)^T} \end{bmatrix}^T \tag{9}$$

with distribution $p(x) = \mathcal{N}(x, \mu_X, \Sigma_X)$. In Julier and Uhlman's work [5, 6] that is called augmentation as the first variable can be considered to be augmented with the other variables. Figure 1 portrays the effect of augmentation for two 2-dimensional statistically independent, Gaussian random variables $X^{(1)}$ and $X^{(2)}$: each of the 5 points chosen for $X^{(1)}$ is augmented with the mean of $X^{(2)}$; conversely, each of the 5 points chosen for $X^{(2)}$ is augmented with the mean of $X^{(1)}$, resulting in the points on the two ellipses.

In the following, we will show how, for statistically independent random variables, the degree of nonlinearity can be calculated individually for each of the variables. Let $X^{(j)}$ be an $n_j$-dimensional Gaussian random variable with distribution

$$p(x^{(j)}) = \mathcal{N}(x^{(j)}, \mu_{X^{(j)}}, \Sigma_{X^{(j)}}),$$

for $j = 1, \ldots, m$. Moreover, let the $X^{(j)}$ be statistically independent and let $X$ be the stacked variable defined in (9). Then the joint covariance matrix $\Sigma_X$ has block-diagonal form, with the blocks being the covariance matrices $\Sigma_{X^{(j)}}$ of the individual variables. Consequently, the Cholesky factorization $R$ of $\Sigma_X$ has block-diagonal form, with the blocks being the right Cholesky factors $R^{(j)}$ of the individual $\Sigma_{X^{(j)}}$. That means the points considered for $X^{(j)}$,

$$\{\mathcal{X}_{2a_j+1}, \ldots, \mathcal{X}_{2b_j}\}, \text{ with } a_j = \sum_{i=1}^{j-1} n_i, \; b_j = \sum_{i=1}^{j} n_i,$$

differ only in the coordinates $a_j$ to $b_j$. All the other $\mathcal{X}_k$, $k \notin [a_j, b_j]$, have these coordinates fixed to $\mu_{X^{(j)}}$. Hence, the degree of nonlinearity contributed by $X^{(j)}$ can be calculated as

$$\eta^{(j)} = \frac{1}{n}\sum_{i=a_j}^{b_j} \eta_i. \tag{10}$$

## 3. A SEQUENCE OF UNSCENTED TRANSFORMS

The unscented transform approximates the transformation $Y = f(X)$ of a Gaussian random variable by – surprise – a Gaussian random variable. That is perfectly reasonable for linear and approximately linear transforms[1]. In the presence of considerable nonlinearities, however, the Gaussian fit might not well represent the true distribution. Therefore, Alspach and Sorenson [3] proposed to approximate the distribution $p(x) = \mathcal{N}(x, \mu_X, \Sigma_X)$ of $X$ as a mixture

$$p(x) \approx m(x) \triangleq \sum_{k=1}^{K} c_k \underbrace{\mathcal{N}(x, \mu_X^{(k)}, \Sigma_X^{(k)})}_{=p(x|k)} \tag{11}$$

of Gaussian distributions with weights $c_k$, means $\mu_X^{(k)}$ and covariances $\Sigma_X^{(k)}$. Then $X$ can be transformed by transforming the individual mixture components, i.e. the conditional variables $X|k$. If the number of mixture components is sufficiently large the covariances can be chosen small enough for $f$ to be approximately linear for each of the transforms and the distribution of $Y$ is well approximated by

$$p(y) \approx \sum_{k=1}^{K} c_k \mathrm{UT}\{p(x|k), f\}, \tag{12}$$

where $\mathrm{UT}\{p(x|k), f\}$ denotes the unscented transform of the Gaussian random variable $X|k$ with respect to the function $f$. The mixture parameters in (11) are typically chosen as to minimize the mean squared error to the true distribution. In order to do that, the Gaussians are arranged on an equidistant grid with equal covariance matrices. Then the mixture weights are optimized [3].

### 3.1. Adapting the Level of Detail to the Nonlinearities

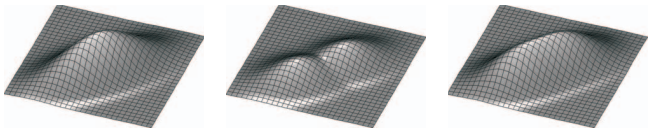
Replacing a Gaussian by an equidistant grid of Gaussians, as in [3], can be regarded as increasing the level of detail in a uniform fashion. In contrast to that, we propose here to adapt the level of detail according to the nonlinearities. By that we mean keeping fewer Gaussians in relatively linear regions, where the transformation is accurate; more Gaussians in nonlinear regions, where the approximation error is higher. That can be achieved by performing the following procedure: At the beginning, the mixture is initialized with a single Gaussian component having the distribution of $X$ – the variable to be transformed. Then, after an initial unscented transform, the mixture, and thereby the transform, is iteratively refined by:

1. splitting that mixture component with the highest degree of nonlinearity into two Gaussians that have half the weight;
2. repeating the unscented transform for the split components and re-evaluating their degrees of nonlinearity.

The iteration is stopped when either the degree of nonlinearity has dropped below a certain threshold or when a certain number of mixture components has been reached. After, the complexity of the Gaussian mixture approximation of $p_Y(y)$ can be reduced by Gaussian mixture reduction techniques [9].

In this work, splitting was performed as described in Appendix B of [4]. For stacked variables (see Section 2.2), only the variable $X^{(j)}$ with the highest degree of nonlinearity $\eta^{(j)}$ was split. The other variables $X^{(l)}$, $l \neq j$ remained unchanged.

[1] A linear transform of a Gaussian random variable always results in a Gaussian random variable.



*(a) original* *(b) split components* *(c) mixture*

**Fig. 2**. Splitting a Gaussian. *The picture to the left shows the original Gaussian distribution; the pictures in the middle and to the right show the maximum and the mixture of the split components, respectively.*

### 3.2. Splitting Priority

Selecting the mixture component to be split based only on its degree of nonlinearity can sometimes result in repeated splits of components whose weights are getting increasingly smaller. That might be suboptimal, as components with a very low weight represent only a small amount of probability mass and thereby do not contribute much to the transformation. Hence, we replace the splitting criterion from the previous section – the component's degree of nonlinearity $\eta^{(k)}$ – by the splitting priority $\rho^{(k)}$, which we define as geometric interpolation between the component's mixture weight and its degree of nonlinearity:

$$\rho^{(k)} \triangleq \mathrm{pow}(c_k, \beta) \cdot \mathrm{pow}(\eta^{(k)}, 1-\beta), \tag{13}$$

where $\beta$ is the interpolation weight and where $\mathrm{pow}(a, b)$ denotes the $b$-th power of $a$, used here for disambiguation from superscript $(k)$.

### 3.3. Transforming Gaussian Mixtures

The adaptive level of detail approach from Section 3.1 can easily be extended to transforming Gaussian mixture random variables, simply by initializing the procedure with a Gaussian mixture instead of a single Gaussian.

## 4. EXPERIMENTS

The proposed method was evaluated by performing a simulation, in which the distribution $p_Y(y)$ of noisy speech $Y$ was to be approximated from the distribution $\mathcal{N}(x; \mu_X, \Sigma_X)$ of clean speech $X$, given the distribution $\mathcal{N}(n; \mu_N, \Sigma_N)$ of noise $N$ as well as a nonlinear interaction function in the speech feature (logarithmic Mel spectra) domain:

$$y = \underbrace{\log(\exp(x) + \exp(n))}_{=f(x,n)}. \tag{14}$$

This transformation is the central point of all speech feature enhancement approaches that estimate the distribution of noise with the expectation maximization (EM) algorithm [1]. Motivated by the fact that frequency bands can be treated independently if the Gaussians are assumed to have diagonal covariance matrices – as is quite common in automatic speech recognition – we simulate the transformation for only one dimension. In the simulation, speech had a Gaussian distribution with mean 5.9 and variance 0.6. Noise had a Gaussian distribution with mean 3.3 and variance 3.0. The smaller variance for speech is sensible, as in [1] noise is modelled as a single Gaussian while speech is modelled as a Gaussian mixture whose components are transformed individually. In order to have a reference to compare to, we generated 10 million samples from the speech and noise distributions, which were then transformed according to (14). The resulting empirical distribution was used to "learn"

**Table 1**. *Approximation error (KLD ·10) for the adaptive level of detail transform under different splitting criteria*

| splitting criterion | number of Gaussians | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 4 | 8 | 16 | 32 |
| weight | 1.898 | 0.777 | 0.404 | 0.306 | 0.383 | 0.568 |
| n-dnl | 1.898 | 0.777 | 0.791 | 0.761 | 0.758 | 0.329 |
| a-dnl | 1.898 | 0.777 | 0.398 | 0.199 | 0.095 | 0.052 |
| spp | 1.898 | 0.777 | 0.255 | 0.173 | 0.066 | 0.039 |

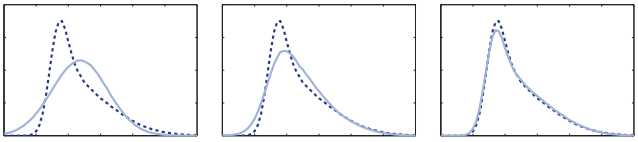
**Table 2**. *Approximation error and computational times in seconds for Monte Carlo transformation + Gaussian mixture recovery*

| | number of samples | | | | |
|---|---|---|---|---|---|
| | 100 | 1K | 10K | 100K | 1M |
| KLD ·10 | 0.423 | 0.078 | 0.059 | 0.056 | 0.044 |
| time in seconds | 0.062 | 0.234 | 1.687 | 15.58 | 156.0 |

a mixture distribution of 10 Gaussian components, whose weights, means and variances were found by performing 50 iterations of k-Means clustering and, thereafter, 50 iterations of the expectation maximization (EM) algorithm [10].

Table 1 shows the Kullback-Leibler divergence (KLD) between the reference and approximations obtained with the adaptive level of detail transform from section 3. The KLD is given for different splitting criteria and in dependency of the number of Gaussians. It was multiplied by a factor of 10 in order to improve readability. In the first row (weight) of the table, components were split based on their mixture weight only. That resulted in a Gaussian mixture approximation of the original distribution, with components that had equally spaced means and equal covariance matrices. For the first couple of iterations, the approximation error (KLD) dropped with the number of Gaussians. After about the 10th iteration, however, it started to raise again. That can be explained by the fact that splitting, though being helpful for overcoming nonlinearities, also changes the original probability density function [4], i.e. it might be detrimental in linear regions. The second and third rows, labeled n-dnl and a-dnl, show approximation errors for the case where the normalized and absolute measures for the degree of nonlinearity were used as splitting criteria. In both cases, the error dropped monotonically with the number of Gaussians. With the absolute measure, however, the error dropped much more sharply – that is, it roughly halved when the number of Gaussians was doubled. With the normalized measure, there seemed to be a tendency towards splitting Gaussians in the same area, resulting in very low weights of few components. For the results shown in the last row (spp) of the table, the splitting priority from Section 3.2 was used as splitting criterion, with interpolation weight $\beta = 0.5$. This consistently reduced the approximation error, in some cases up to 30% compared to the next best criterion.

Table 2 shows results for the Monte Carlo (MC) approach used in the penultimate paragraph in order to obtain the reference distribution. For the results shown in Table 2, we used a lower number of samples and performed only 10 iterations clustering and EM training. As can be seen, the approximation error decreased with the number of samples, though this seemed to abate after approximately 1000 samples. In addition to the KLD, we give computational times. A brief look at the table immediately reveals the major disadvantage of the MC method: its computational expense. Even with just 100 samples, the approximation took 62 milliseconds to complete on a 1.2 GHz Intel Atom CPU, compared to less than 2 milliseconds for the adaptive level of detail transform with 32 Gaussians. At the same time, the approximation error was 10 times lower for the latter.



*(a) ALoDT-1* *(b) ALoDT-4* *(c) ALoDT-16*

**Fig. 3**. *Transformed Distribution. The dashed curve shows the true distribution. The solid curves show approximations obtained with the adative level of detail transform (ALoDT) with 1, 4 and 16 Gaussians, using the splitting priority (spp) from Section 3.2.*

## 5. CONCLUSIONS

We have presented a novel method for approximating nonlinear transformations of Gaussian mixture random variables. It is based on adapting the mixture's level of detail to the nonlinearities present in the transformation. For that, we have derived a measure for the degree of nonlinearity based on the coefficient of determination, and, subsequently refined it into an efficient splitting criterion. In the experimental section, we have successfully applied the proposed method to a particular transformation that is commonly used in speech feature enhancement.

## 6. REFERENCES


[1] P. Moreno, B. Raj, and R. Stern, "A vector Taylor series approach for environment-independent speech recognition," *Proc. ICASSP*, vol. 3068, pp. 733–736, May 1996.

[2] R. van der Merwe and E. Wan, "Gaussian mixture sigma-point particle filters for sequential probabilistic inference in dynamic state-space models," *Proc. ICASSP*, vol. 6, no. IV, pp. 701–704, Apr. 2003.

[3] D. L. Alspach and H. W. Sorenson, "Nonlinear Bayesian estimation using Gaussian sum approximations," *IEEE Trans. Autom. Control*, vol. 17, no. 4, pp. 439–448, Aug. 1972.

[4] F. Faubel, J. McDonough, and D. Klakow, "The split and merge unscented Gaussian mixture filter," *IEEE Signal Process. Lett.*, vol. 16, no. 9, pp. 786–789, Sept. 2009.

[5] S. J. Julier and J. K. Uhlmann, "A new extension of the Kalman filter to nonlinear systems," *Proc. AeroSense*, pp. 182–193, Apr. 1997.

[6] S. J. Julier and J. K. Uhlmann, "Unscented filtering and nonlinear estimation," *Proc. IEEE*, vol. 92, no. 3, pp. 401–422, Mar. 2004.

[7] F. Faubel and D. Klakow, "A transformation-based derivation of the Kalman filter and an extensive unscented transform," *IEEE Workshop on Statistical Signal Processing*, Sept. 2009.

[8] K. Ito and K. Xiong, "Gaussian filters for nonlinear filtering problems," *IEEE Trans. Autom. Control*, vol. 45, no. 5, pp. 910–927, May 2000.

[9] Andrew R. Runnalls, "A Kullback-Leibler approach to Gaussian mixture reduction," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 43, no. 3, pp. 989–999, July 2007.

[10] A. P. Dempster, N. M. Laird., and D. B. Rubin, "Maximum-likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, May 1977.