A versatile Gaussian splitting approach to non-linear state estimation and its application to noise-robust ASR

Volker Leutnant, Alexander Krueger, Reinhold Haeb-Umbach
Department of Communications Engineering, University of Paderborn, Germany
{leutnant,krueger,haeb}@nt.uni-paderborn.de

Abstract

In this work, a splitting and weighting scheme that allows for splitting a Gaussian density into a Gaussian mixture density (GMM) is extended to allow the mixture components to be arranged along arbitrary directions. The parameters of the Gaussian mixture are chosen such that the GMM and the original Gaussian still exhibit equal central moments up to an order of four. The resulting mixtures' covariances will have eigenvalues that are smaller than those of the covariance of the original distribution, which is a desirable property in the context of non-linear state estimation, since the underlying assumptions of the extended KALMAN filter are better justified in this case. Application to speech feature enhancement in the context of noise-robust automatic speech recognition reveals the beneficial properties of the proposed approach in terms of a reduced word error rate on the Aurora 2 recognition task.

Index Terms: density splitting, moment matching

1. Introduction

The application of *model-based speech feature enhancement* to the prominent noise-robustness problem of today's automatic speech recognizer (ASR) systems has gained considerable interest in recent years. Thereby, conditional BAYESian estimation is employed to infer the uncorrupted speech feature vectors from the corrupted observations based on a priori models of the cepstral speech feature vectors $\mathbf{x}_t$ and the noise-only feature vectors $\mathbf{n}_t$ and a highly non-linear observation model relating the two to the corrupted observations $\mathbf{y}_t$ [1]:

$$\mathbf{y}_t = \mathbf{D}\log\left(e^{\mathbf{D}^+\mathbf{x}_t} + e^{\mathbf{D}^+\mathbf{n}_t} + 2\boldsymbol{\alpha}_t e^{\mathbf{D}^+\frac{(\mathbf{x}_t+\mathbf{n}_t)}{2}}\right). \quad (1)$$

Thereby $\mathbf{D}$ and $\mathbf{D}^+$ denote the discrete-cosine transform (DCT) matrix and its pseudo-inverse, respectively. Several approximations have been proposed to model the observation probability density $p(\mathbf{y}_t|\mathbf{x}_t,\mathbf{n}_t)$. The most prominent and at the same time simplest approach neglects any phase difference between the speech and the noise signal. However, it is well known and has already been verified experimentally that the more accurate model of (1), which includes the phase factor $\boldsymbol{\alpha}_t$, results in improved performance [2]. Since a numerical evaluation of the occurring integrals to obtain the observation probability $p(\mathbf{y}_t|\mathbf{x}_t,\mathbf{n}_t)$ is computationally very demanding, it usually is still approximated by a Gaussian density.

The two most prominent filters for non-linear estimation problems are the unscented [3] and the extended KALMAN filter [4]. The unscented filter determines the mean and covariance of this Gaussian by means of so-called *sigma points*, which are drawn from a joint Gaussian a priori density of $\mathbf{x}_t$ and $\mathbf{n}_t$ and later on propagated through the non-linearity. In contrast, the extended KALMAN filter utilizes lower-order terms of the Taylor-series expansion of the non-linearity to obtain estimates of the mean and the covariance and in general is very sensitive to the choice of the expansion vector. But it is only when the spectral radius of the covariance matrix of the joint Gaussian prior is sufficiently small that the underlying assumptions are approximately valid.

Driven by this idea, ALSPACH and SORENSON proposed to model the original a priori distribution by a Gaussian mixture model where the individual mixtures exhibit smaller eigenvalues than the original distribution [5]. Only recently, MERWE et al. [6] and FAUBEL et al. [7] adopted this approach. While MERWE et al. applied a weighted expectation-maximization algorithm to samples drawn from the a priori density to obtain the parameters of a GMM, FAUBEL et al. proposed to iteratively increase the number of mixture components by splitting a selected Gaussian along one of the eigenvectors of the corresponding covariance. The mixture component to split and the eigenvector the split is performed along is thereby determined by the *degree of non-linearity* [7]. Both approaches can be considered computationally expensive and thus, to the best of our knowledge, none or very few attempts have been made to apply the concept to the problem of speech feature enhancement as is done here. Further, we will extend the theory to splits along arbitrary directions, not restricted to those defined by the eigenvectors of the covariance matrix.

The paper begins with a short review of approximating a GMM by a single Gaussian by means of minimizing their KULLBACK-LEIBLER divergence. Since the same criterion cannot be applied to split a Gaussian into a GMM without any constraints on the parameter set, a splitting and weighting scheme which allows for multiple splits in arbitrary directions is introduced next, followed by a discussion on the sensitivity of the choice of the parameters. The splitting and weighting scheme is then applied to model-based speech feature enhancement and its performance is evaluated by means of recognition results on the Aurora 2 database [8].

2. Merging a GMM into a single Gaussian

Given a Gaussian mixture probability density $p(\mathbf{z}),\ \mathbf{z}\in\mathbb{R}^D$

$$p(\mathbf{z}) = \sum_{m=0}^{M-1} w_m p_m(\mathbf{z}) = \sum_{m=0}^{M-1} w_m \mathcal{N}(\mathbf{z};\ \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m), \quad (2)$$

where w_m, $\boldsymbol{\mu}_m$ and $\boldsymbol{\Sigma}_m$ denote the weight, the mean and the covariance of the m-th Gaussian mixture component $p_m(\mathbf{z})$, $m\in\{1,..,M\}$, respectively, *merging* the GMM to a single Gaussian $q(\mathbf{z})=\mathcal{N}(\mathbf{z};\ \tilde{\boldsymbol{\mu}},\tilde{\boldsymbol{\Sigma}})$ usually refers to finding its mean $\tilde{\boldsymbol{\mu}}$ and covariance $\tilde{\boldsymbol{\Sigma}}$ such that the KULLBACK-LEIBLER (KL) divergence between $p(\mathbf{z})$ and $q(\mathbf{z})$ is minimized, resulting in

$$\tilde{\boldsymbol{\mu}} = \sum_{m=0}^{M-1} w_m \boldsymbol{\mu}_m \quad (3)$$

$$\tilde{\boldsymbol{\Sigma}} = \sum_{m=0}^{M-1} w_m \boldsymbol{\Sigma}_m + \sum_{m=0}^{M-1} w_m (\boldsymbol{\mu}_m - \tilde{\boldsymbol{\mu}})(\boldsymbol{\mu}_m - \tilde{\boldsymbol{\mu}})', \quad (4)$$

where $(\cdot)'$ denotes the matrix/vector transpose operator. The covariance $\tilde{\mathbf{\Sigma}}$ can thereby be regarded as being composed of a "within class" term $\mathbf{S}_W = \sum_{m=0}^{M-1} w_m \mathbf{\Sigma}_m$ and a "between class" term $\mathbf{S}_B = \sum_{m=0}^{M-1} w_m (\boldsymbol{\mu}_m - \tilde{\boldsymbol{\mu}})(\boldsymbol{\mu}_m - \tilde{\boldsymbol{\mu}})'$. This approach is also referred to as *moment matching* for obvious reasons.

3. Splitting a Single Gaussian into a GMM

While merging a GMM into a single Gaussian by minimizing their KL-divergence is common practice, the objective here is the opposite: *splitting* a Gaussian into a GMM consisting of an arbitrary number of mixture components M. A splitting and a weighting scheme which allow for multiple splits in arbitrary directions are described next. To keep the objective of minimizing the KL-divergence, both schemes are driven by the key idea of matching the moments of the GMM to that of the Gaussian.

3.1. Splitting in an Arbitrary Direction

Splitting the Gaussian $q(\mathbf{z}) = \mathcal{N}(\mathbf{z};\ \tilde{\boldsymbol{\mu}}, \tilde{\mathbf{\Sigma}})$ into a GMM $p(\mathbf{z})$ with $M = 2K+1$ components along an arbitrary direction $\mathbf{u}_l \in \mathbb{R}^D$ can be achieved by first shifting the mean $2K$-times and second transforming the covariance of the shifted density. Arranging the means of the M mixture components symmetrically with respect to $\tilde{\boldsymbol{\mu}}$ along $\mathbf{u}_l$ with an equidistant spacing while transforming all covariances in the same manner results in a GMM that is fully specified by the parameter set

$$\begin{array}{lll} \boldsymbol{\mu}_0 = \tilde{\boldsymbol{\mu}}, & \mathbf{\Sigma}_0, & w_0, \\ \boldsymbol{\mu}_k = \tilde{\boldsymbol{\mu}} + \eta k \mathbf{u}_l, & \mathbf{\Sigma}_k = \mathbf{\Sigma}_0, & w_k, \\ \boldsymbol{\mu}_{K+k} = \tilde{\boldsymbol{\mu}} - \eta k \mathbf{u}_l, & \mathbf{\Sigma}_{K+k} = \mathbf{\Sigma}_k, & w_{K+k} = w_k, \end{array} \tag{5}$$

where $k \in \{1, .., K\}$. All weights are positive and normalized such that $w_0 + \sum_{\tilde{k}=1}^{2K} w_{\tilde{k}} = 1$. The parameter $\eta \in \mathbb{R}_{\geq 0}$ determines the distance between two adjacent means in terms of the length of the vector $\mathbf{u}_l$.

The symmetric setup ensures the first central moment of the Gaussian and the GMM to match. Applying (4) and solving for $\mathbf{\Sigma}_0$ to also match the second central moment yields

$$\mathbf{\Sigma}_0 = \tilde{\mathbf{\Sigma}} - \beta \mathbf{u}_l \mathbf{u}_l',\ \beta = 2\eta^2 \sum_{k=1}^{K} w_k k^2, \tag{6}$$

where the parameter $\beta \in \mathbb{R}_{\geq 0}$ controls the covariance reduction, which certainly has to be a function of the mean displacement η if the moments have to match. Calling for $\mathbf{\Sigma}_0$ to be symmetric and non-negative definite poses constraints on either the length of $\mathbf{u}_l$ or β. While the symmetry is obviously given by any vector $\mathbf{u}_l \in \mathbb{R}^D$, the non-negative definiteness property

$$\mathbf{z}' \left(\tilde{\mathbf{\Sigma}} - \beta \mathbf{u}_l \mathbf{u}_l' \right) \mathbf{z} \geq 0 \quad \forall \mathbf{z} \in \mathbb{R}^D \tag{7}$$

is preserved only by vectors in a certain subset of $\mathbb{R}^D$. Expressing the covariance $\tilde{\mathbf{\Sigma}}$ in terms of an orthonormal matrix composed of its eigenvectors $\mathbf{V} = [\mathbf{v}_1, ..., \mathbf{v}_D]$ and a diagonal matrix composed of its eigenvalues $\mathbf{\Lambda} = \mathrm{diag}([\lambda_1, ..., \lambda_D])$, one obtains

$$\mathbf{z}' \left(\mathbf{V} \mathbf{\Lambda} \mathbf{V}' - \beta \mathbf{u}_l \mathbf{u}_l' \right) \mathbf{z} \geq 0 \quad \forall \mathbf{z} \in \mathbb{R}^D, \tag{8}$$

which is equivalent to

$$\tilde{\mathbf{z}}' \left(\mathbf{I} - \beta \mathbf{\Lambda}^{-\frac{1}{2}} \mathbf{V}' \mathbf{u}_l \mathbf{u}_l' \mathbf{V} \mathbf{\Lambda}^{-\frac{1}{2}} \right) \tilde{\mathbf{z}} \geq 0 \quad \forall \tilde{\mathbf{z}} \in \mathbb{R}^D, \tag{9}$$

with $\tilde{\mathbf{z}} = \mathbf{\Lambda}^{\frac{1}{2}} \mathbf{V}' \mathbf{z}$, since $\mathbf{\Lambda}^{\frac{1}{2}} \mathbf{V}'$ is full-rank.

The matrix $\mathbf{I} - \beta (\mathbf{\Lambda}^{-\frac{1}{2}} \mathbf{V}' \mathbf{u}_l)(\mathbf{\Lambda}^{-\frac{1}{2}} \mathbf{V}' \mathbf{u}_l)'$ in turn is non-negative definite if the eigenvalues of $\beta (\mathbf{\Lambda}^{-\frac{1}{2}} \mathbf{V}' \mathbf{u}_l)(\mathbf{\Lambda}^{-\frac{1}{2}} \mathbf{V}' \mathbf{u}_l)'$ are smaller than or equal to one. Assuming without loss of generality that β is bound to the interval $[0, 1]$, non-negative definiteness can be ensured by normalizing the vector $\mathbf{u}_l$ by the square root of the maximum eigenvalue of $(\mathbf{\Lambda}^{-\frac{1}{2}} \mathbf{V}' \mathbf{u}_l)(\mathbf{\Lambda}^{-\frac{1}{2}} \mathbf{V}' \mathbf{u}_l)'$, which is just the length of the vector $\mathbf{\Lambda}^{-\frac{1}{2}} \mathbf{V}' \mathbf{u}_l$. The normalized vector $\breve{\mathbf{u}}_l$ is thus given by

$$\breve{\mathbf{u}}_l = \frac{\mathbf{u}_l}{\sqrt{(\mathbf{\Lambda}^{-\frac{1}{2}} \mathbf{V}' \mathbf{u}_l)'(\mathbf{\Lambda}^{-\frac{1}{2}} \mathbf{V}' \mathbf{u}_l)}} = \frac{\mathbf{u}_l}{\sqrt{\mathbf{u}_l' \tilde{\mathbf{\Sigma}}^{-1} \mathbf{u}_l}}. \tag{10}$$

Note that if $\mathbf{u}_l$ coincides with an arbitrary eigenvector $\mathbf{v}_d$ of $\tilde{\mathbf{\Sigma}}$, $\breve{\mathbf{u}}_l$ turns into $\breve{\mathbf{u}}_l = \sqrt{\lambda_d} \mathbf{v}_d / ||\mathbf{v}_d||$.

3.2. Simultaneous Splitting in Multiple Directions

Extension to a simultaneous split in L arbitrary directions $\mathbf{u}_1, ..., \mathbf{u}_L$ can be carried out in a similar fashion. The parameters of the resulting GMM consisting of $M = 2KL+1$ mixture components are given by

$$\begin{array}{lll} \boldsymbol{\mu}_0 = \tilde{\boldsymbol{\mu}}, & \mathbf{\Sigma}_0 = \tilde{\mathbf{\Sigma}} - \beta \sum_{l=1}^{L} \mathbf{u}_l \mathbf{u}_l', & w_0 \\ \boldsymbol{\mu}_k^l = \tilde{\boldsymbol{\mu}} + \eta k \mathbf{u}_l, & \mathbf{\Sigma}_k^l = \mathbf{\Sigma}_0, & w_k^l \\ \boldsymbol{\mu}_{K+k}^l = \tilde{\boldsymbol{\mu}} - \eta k \mathbf{u}_l, & \mathbf{\Sigma}_{K+k}^l = \mathbf{\Sigma}_k^l, & w_{K+k}^l = w_k^l, \end{array} \tag{11}$$

where $l \in \{1, .., L\}$ and $k \in \{1, .., K\}$. All split directions $\mathbf{u}_l$ again have to be normalized, this time by the square root of the maximum eigenvalue of $\sum_{l=1}^{L} (\mathbf{\Lambda}^{-\frac{1}{2}} \mathbf{V}' \mathbf{u}_l)(\mathbf{\Lambda}^{-\frac{1}{2}} \mathbf{V}' \mathbf{u}_l)'$ to preserve the non-negative definiteness of all covariances.

Applying (4) to match the second central moment calls for

$$\sum_{l=1}^{L} \mathbf{u}_l \mathbf{u}_l' \left[\beta - 2\eta^2 \sum_{k=1}^{K} w_k^l k^2 \right] \stackrel{!}{=} \mathbf{0}, \tag{12}$$

which can simply be met for an arbitrary number of split directions L if the weights are chosen to be independent of the index l. The parameter η determining the displacement of the mixture means and the parameter β controlling the change in the covariances are thus again (compare (6)) related by

$$\beta = 2\eta^2 \sum_{k=1}^{K} w_k^l k^2. \tag{13}$$

Since the normalization of the split directions requires $0 \leq \beta \leq 1$, the mean shift η will be bound to $\eta \in [0, \eta_{\max}]$, which in general will depend on L and K. If the weights are independent of η, a closed-form solution for the upper bound $\eta_{\max}$ can be derived from (13). However, if the weights depend on η, a closed-form solution usually does not exist, as will be outlined next.

3.3. Weighting Scheme

Equation (13) relates the covariance reduction parameter β to the mean shift η and the chosen weights w_k^l. Weighting schemes where the center weight w_0 is specified and the remaining probability mass is equally distributed among the remaining $2KL$ non-center mixtures to satisfy $w_0 + 2\sum_{l=1}^{L} \sum_{k=1}^{K} w_k^l = 1$ are quite common [3]. However, since we split a Gaussian $(2K+1)$-times along a specified direction $\mathbf{u}_l$, assigning the same weight to mixtures that are *close* to the mean $\tilde{\boldsymbol{\mu}}$ and those that are *further away* from it may not be an appropriate choice. Instead, we assign each mixture component $p_m(\mathbf{z})$ a weight that is proportional to the likelihood of its shifted mean under the initial Gaussian $q(\mathbf{z})$ as

$$w_0 = \frac{1}{C}\gamma,\ w_k^l = w_{K+k}^l = \frac{1}{C} e^{-\frac{1}{2} k^2 \eta^2 \mathbf{u}_l' \tilde{\mathbf{\Sigma}}^{-1} \mathbf{u}_l}, \tag{14}$$

with the normalization constant $C \in \mathbb{R}_{>0}$. The parameter γ determines how much weight is assigned to the centered mixture component and moreover controls the higher moments, as will be utilized later. Since the weights represent probabilities, γ thereby is constrained to be $\in \mathbb{R}_{\geq 0}$, as opposed to the unscented filter [3] where also negative center weights may be

considered. Since the weights are also required to be independent of l to meet (12) for an arbitrary L, we further require $\mathbf{u}_l'\tilde{\mathbf{\Sigma}}^{-1}\mathbf{u}_l = c,\ c \in \mathbb{R}_{>0}\ \forall l\ \in\ \{1,..,L\}$. Both non-negative definiteness and weights independent of l can be achieved by normalizing $\mathbf{u}_l$ according to

$$\breve{\mathbf{u}}_l = \frac{\mathbf{u}_l}{\sqrt{\mathbf{u}_l'\tilde{\mathbf{\Sigma}}^{-1}\mathbf{u}_l}} \tag{15}$$

first and compute the non-zero eigenvalues $\nu_1, ..., \nu_r,\ r \leq \min\{L, D\}$ of the matrix $\sum_{l=1}^{L}(\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{V}'\breve{\mathbf{u}}_l)(\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{V}'\breve{\mathbf{u}}_l)'$ afterwards. The final normalization step is then given by

$$\breve{\breve{\mathbf{u}}}_l = \sqrt{c}\,\breve{\mathbf{u}},\ c = 1/\max_r\{\nu_r\}. \tag{16}$$

If all $\mathbf{u}_l$ coincide with directions of the eigenvectors of $\tilde{\mathbf{\Sigma}}$, i.e., $\mathbf{u}_l'\mathbf{v}_i \neq 0$ and $\mathbf{u}_l'\mathbf{v}_j = 0\ \forall j \neq i, j \in \{1,..,D\}$ and a particular $i \in \{1,..,D\}$, c becomes the reciprocal of the maximum number of vectors $\mathbf{u}_l$ pointing in direction of the same eigenvector.

The parameter γ introduced in (14) can either be set heuristically or, if the splits are performed along eigenvectors of $\tilde{\mathbf{\Sigma}}$, be determined such that the fourth moment of the GMM matches the fourth moment of the Gaussian. However, with the generalized raw moment of order four of a random vector $\mathbf{z}$ given by $E[\mathbf{z}\mathbf{z}' \otimes \mathbf{z}\mathbf{z}']$ (see [9]), where $\otimes$ denotes the KRONECKER product, matching the fourth moment would require

$$E_{q(\mathbf{z})}[z_g z_h z_i z_j] = \sum_{m=0}^{M-1} w_m E_{p_m(\mathbf{z})}[z_g z_h z_i z_j] \tag{17}$$

to hold for all possible combination of $g, h, i, j \in \{1,..,D\}$ of the vector $\mathbf{z}$. Though the expectation values can be computed by applying ISSERLIS theorem [10] for any tuple (g, h, i, j), a single scalar γ cannot fulfill (17) for all tuples. Focusing on the case where $g{=}h{=}i{=}j$, Eq. (17) calls for

$$\left(\sum_{k=1}^{K} w_k^l k^2\right)^2 - \frac{1}{6}\sum_{k=1}^{K} w_k^l k^4 \stackrel{!}{=} 0. \tag{18}$$

Note that the above constraint on the center weight w_0 applies to any weighting scheme where the weights are independent of a particular split direction $\mathbf{u}_l$. Using the exponential weighting scheme (14) and solving for γ results in

$$\gamma_{\text{opt}} = 6\frac{\left(\sum_{k=1}^{K} k^2 e^{-\frac{1}{2}k^2\eta^2 c}\right)^2}{\sum_{k=1}^{K} k^4 e^{-\frac{1}{2}k^2\eta^2 c}} - 2L\sum_{k=1}^{K} e^{-\frac{1}{2}k^2\eta^2 c}, \tag{19}$$

which, dependent on L and K, may be negative and as such not a valid solution. Thus, the additional non-negativity constraint on γ may keep the fourth moments of the GMM and the Gaussian from matching. Compensation of this shortcoming by, e.g., a fixed non-negative γ and a proper choice of η may also be difficult, since η, in addition, has to be chosen such that $0{\leq}\beta{\leq}1$ to ensure non-negative definiteness of the covariances. Its bound η_{max} can, in general, not be determined analytically, however, may be determined graphically, as it is shown in Fig. 1 for $\gamma{=}1$ and $\gamma{=}\gamma_{\text{opt}}$ for $L{=}1$ and $K{\in}\{1,..,3\}$.

The choice of η may be considered to be subject to a trade-off between within and between class contribution to the merged/matched covariance $\tilde{\mathbf{\Sigma}}$. However, the major aim of the proposed splitting scheme is not to find a *shape preserving* representation of the original Gaussian distribution, but to reduce the influence of the non-linearity on the Taylor-series expansion – which may also motivate $\gamma{\neq}\gamma_{\text{opt}}$ to be chosen, regardless of a valid solution. Thus, η would have to be chosen large (β close to one) to generate mixture components that are individually as concentrated as possible.

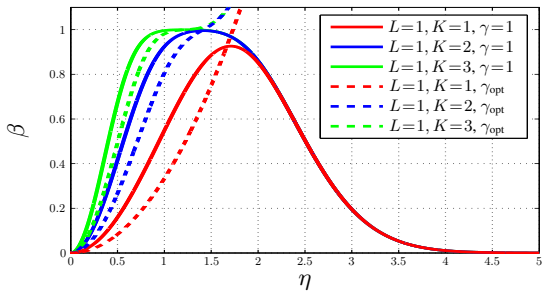

Figure 1: *The covariance scaling factor β as a function of η according to* (13) *with $\gamma{=}1$ (solid lines) and $\gamma{=}\gamma_{opt}$ (dashed lines) for $L{=}1$ and $K \in \{1,..,3\}$; η is allowed to be of the subset $\{\eta{\in}\mathbb{R}_{\geq 0}|0 \leq \beta \leq 1\}$*

4. Application to Model-based Speech Feature Enhancement

Application of the proposed splitting scheme to the model-based speech feature enhancement given in [2] calls for only slight modifications. There, a switching linear dynamic model has been used as a priori model of speech and noise. Denoting the regime variable indicating the active dynamic model at time instant t by s_t, the splitting scheme will independently be applied after the time-update step to all predictive Gaussian distributions $p(\mathbf{x}_t, \mathbf{n}_t|\mathbf{y}_{1:t-1}, s_t{=}j)$, $j \in [1, J]$. This is followed by a total of $J(2KL{+}1)$ measurement-update steps, here carried out by the iterated extended KALMAN filter as given in [2] and denoted by IEKF-α. It is only after all measurement update steps have been performed that we apply the moment matching to match the mixture of $J(2KL{+}1)$ posterior distributions to a single Gaussian, as required by the applied multi-model inference scheme. Denoting the measurement-update instance under the j-th dynamic model and its m-th split distribution by the index pair (j, m), we as such do not merge the single estimated observation densities $p(\mathbf{y}_t|\mathbf{x}_t, \mathbf{n}_t, (j, m))$, as carried out by FAUBEL et al. [7], but the resulting posterior distributions $p(\mathbf{x}_t, \mathbf{n}_t|\mathbf{y}_{1:t}, (j, m))$. This calls for the calculation of the posterior probabilities of all $J(2KL{+}1)$ measurement-update instances, as opposed to the J model posterior probabilities under the standard multi-model inference scheme.

5. Experimental Results

The experiments were conducted on the test set A of the Aurora 2 database. The IEKF-α is applied in the cepstral domain, followed by a cepstral mean subtraction carried out on the enhanced feature vectors prior to calculation of the dynamic components. Splitting is performed along the L dominant eigenvectors of the predictive Gaussian distributions $p(\mathbf{x}_t, \mathbf{n}_t|\mathbf{y}_{1:t-1}, s_t{=}j)$ for $J{=}16$ linear dynamic models.

In a first experiment we investigate the influence of the mean shift η on the recognition performance.Therefore, the parameter η is varied between 0 and its bound η_{max}. Since we expect the splitting to be most beneficial under low signal-to-noise ratio (SNR) conditions, we focus on the 0 dB case, only. Thereby Eq. (19) is used to determine the *optimal* center weight, which has been found to be non-negative for the considered setup. Note that though the optimal value for γ depends on L and K, η_{max} solely depends on K if $\gamma{=}\gamma_{\text{opt}}$ is chosen. The resulting recognition results are given in Fig. 2.

The performance of the IEKF-α without application of the splitting scheme can be read off at $\eta/\eta_{\text{max}}{=}0$. For all examined

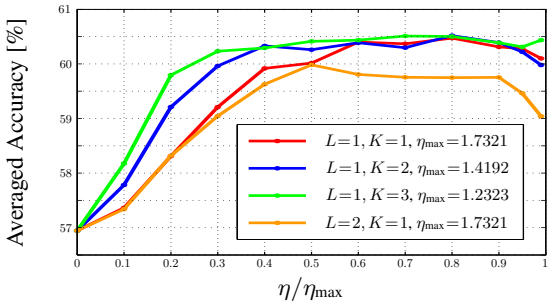

Figure 2: *Averaged recognition accuracies on test set A of the Aurora 2 database for an SNR of* 0 *dB;* $\eta \in [0, \eta_{max}]$, $\gamma = \gamma_{opt}$ *for* $L=1, K \in \{1, .., 3\}$ *and* $L=2, K=1$

setups ($L=1,\ K \in \{1, .., 3\}$ and $L=2,\ K=1$), the averaged recognition accuracies exhibit the same characteristic: the accuracy first increases with η, reaches a plateau and finally declines when approaching $\eta_{\max}$. The greatest improvement compared to the baseline ($+3.5\%$ absolute) is obtained with $L=1$ and $K=2$ at $\eta/\eta_{\max} = 0.8$. Interestingly, increasing the number of split directions from $L=1$ to $L=2$ could not improve the results.

This may be contributed to the fact that splitting (and filtering) is performed in the cepstral domain, where the energy component dominates the eigenvalues. However, the linearization of (1) with respect to the current split distribution and the estimation of the mean and covariance of the corresponding observation density requires the transformation to the logarithmic mel power spectral domain, which is done by application of the pseudo-inverse of the DCT matrix. The variance reduction in the energy component can thus be considered to *spread* along all components of the state vector in the logarithmic mel power spectral domain and thus reduce the influence of the non-linearity on the linearization for all components of the observation model in the logarithmic mel power spectral domain. Nevertheless, splitting in $L > 1$ directions may become useful for different observation models or inference problems.

With η approaching its bound $\eta_{\max}$, β approaches one and the variances along the corresponding split directions becomes zero. Though this may be reminiscent of the unscented filter, there is a subtle, but important difference to it. First, the proposed splitting scheme does not require the number of split directions to be equal to the number of (possibly augmented) state dimensions as required by the unscented filter. Thus, one can focus on those directions where the non-linearity is most severe. Second and most important, the proposed splitting scheme allows for merging the contributions of the individual splits on the posterior density level. This, however, is not possible under the unscented filter, which explicitly assumes the observation density to be Gaussian and thus renders the posterior distribution to be Gaussian, too. However, a detailed comparison to the unscented filter is not the focus of this paper and will be left for future research.

Finally, the computationally modest configuration ($L=1$, $K=1$, $\gamma=\gamma_{\text{opt}}$, $\eta/\eta_{\max}=0.8$) is used to carry out feature enhancement on the complete test set A of the Aurora 2 database. This time, we also utilize the uncertainty about the clean speech estimates provided by the filter to carry out the final recognition with *uncertainty decoding* [11]. Results for both filters, the standard IEKF-α and the filter with the Gaussian sum approximation, are given in Tab. 1. In comparison to the standard IEKF-α, the application of the splitting scheme results in an improved recognition performance which under the standard decoder already reaches the performance of the uncertainty decoder applied to the standard IEKF-α. Additional application of the uncertainty decoder finally results in an accuracy of 87.41%. Note that the major gain is almost exclusively achieved under low SNR conditions – validating the assumption our first experiment was subject to.

Table 1: *Averaged recognition accuracies on test set A of the Aurora 2 database with* standard *and* uncertainty *decoding*

SNR	**IEKF-α**		**IEKF-α + splitting**	
[dB]	standard	uncertainty	standard	uncertainty
20	98.72	98.64	98.70	98.66
15	97.18	97.31	97.21	97.30
10	93.61	94.03	93.41	93.72
5	82.82	84.12	83.16	84.66
0	56.95	58.92	60.47	62.73
∅	85.86	86.60	86.59	87.41

6. Conclusions

In this work, a splitting and weighting scheme has been described that allows for splitting a Gaussian density into a Gaussian mixture density. We were able to extend the existing theory to splits along arbitrary directions, not necessarily along the directions of the eigenvectors. The GMM and the original Gaussian exhibit equal central moments up to an order of four. However, the resulting mixtures' covariances have eigenvalues that are smaller than those of the covariance of the original distribution, thereby reducing the linearization error of a Taylor series expansion. Application to model-based speech feature enhancement on the the Aurora 2 speech recognition task confirmed this property to be beneficial in the context of non-linear state estimation. The splitting contributes most under low SNR conditions, where the non-linearity is most severe and where the covariance of the predictive distribution is *large*, such that the assumptions inherent to the extended KALMAN filter are contradicted most.

7. Acknowledgements

This work has been supported by Deutsche Forschungsgemeinschaft (DFG) under contract no. Ha3455/6-1.

8. References

[1] J. Droppo, A. Acero, and L. Deng, "A nonlinear observation model for removing noise from corrupted speech log mel-spectral energies," in *Proc. Int. Conf. Spoken Lang. Process. (ICSL)*, 2002.

[2] V. Leutnant and R. Haeb-Umbach, "An analytic derivation of a phase sensitive observation model for noise robust speech recognition," in *Proc. Annu. Conf. of ISCA (Interspeech)*, 2009.

[3] S. J. Julier and J. K. Uhlmann, "A new extension of the Kalman filter to nonlinear systems," in *Int. Symp. Aerospace/Defense Sensing, Simul. and Controls*, 1997.

[4] B. Bell and F. Cathey, "The iterated Kalman filter update as a Gauss-Newton method," *IEEE Trans. Autom. Control*, vol. 38, no. 2, 1993.

[5] D. Alspach and H. Sorenson, "Nonlinear Bayesian estimation using Gaussian sum approximations," *IEEE Trans. Autom. Control*, vol. 17, no. 4, 1972.

[6] R. van der Merwe and E. Wan, "Gaussian mixture sigma-point particle filters for sequential probabilistic inference in dynamic state-space models," in *Proc. Int. Conf. Acoust., Speech and Sig. Process. (ICASSP)*, 2003.

[7] F. Faubel and D. Klakow, "Further improvement of the adaptive level of detail transform: Splitting in direction of the nonlinearity," in *Proc. Europ. Sig. Process. Conf. (EUSIPCO)*, 2010.

[8] D. Pearce and H.-G. Hirsch, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. Int. Conf. Spoken Lang. Process. (ICSL)*, 2000.

[9] Z. Winiewski, "Vectors of kth order moments," *Bulletins of the Stanisaw Staszic Academy of Mining and Metallurgy, Geodesy b. 112*, no. 1423, 1991.

[10] L. Isserlis, "On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables," *Biometrika*, vol. 12, no. 1/2, 1918.

[11] V. Ion and R. Haeb-Umbach, "A novel uncertainty decoding rule with applications to transmission error robust speech recognition," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 16, no. 5, 2008.