A novel approach to noise estimation in model-based speech feature enhancement

Stefan Windmann, Reinhold Haeb-Umbach

Dept. of Communications Engineering, University of Paderborn, 33098 Paderborn, Germany
E-Mail: {windmann,haeb}@nt.uni-paderborn.de
Web: www.nt.uni-paderborn.de

Abstract

In this paper, we consider the noise estimation for model-based speech feature enhancement in the context of automatic speech recognition (ASR). The (cepstral) noise process is described with a linear state space model. Novel EM algorithms are derived for the estimation of the noise model parameters: First, a blockwise EM algorithm operating on noise-only input data is derived. It is supposed to be used during the offline training mode of the recognizer. Second, a sequential online EM algorithm is introduced which is able to work on input data consisting of noisy speech and which can be employed in recognition mode. Experiments on the AURORA4 database lead to improved recognition results with the new state model compared to the assumption of stationary noise.

1 Introduction

Automatic speech recognition (ASR) requires in many applications a denoising of the speech. For this purpose, a model-based approach can be applied where prior knowledge about the distributions of speech and noise is exploited for speech enhancement. The paper at hand considers the estimation of the noise prior. The application of a voice activity detection (VAD) for noise estimation is unreliable at low signal-to-noise ratios (SNRs). On the other hand there exist approved methods in the spectral domain like the Minimum statistics [10] and the (Iterated) Minima-Controlled-Recursive averaging ((I)MCRA) technique [2] which yield robust point estimates of the noise. For model based speech feature enhancement, however, an estimate of the noise probability density function (pdf) is required. The parameters of the noise pdf can be estimated with block Expectation-Maximization (EM) algorithms (e.g. [9, 11]) and sequential EM algorithms [3, 8] in the feature domain. It is further possible to optimize the posterior of the noise distribution independently for each speech frame within the model-based approach [5]. While in the works referred to so far the noise process has been assumed to be stationary, some researches have considered a state space model of the noise dynamics to account for changing noise conditions. Kim [7] employs a random-walk model where the state model variance is set to a fixed value while Singh [12] uses a more general state model where the parameters are estimated from noise-only training data. We have proposed a noise model [14] where the systematic part of the noise is related to a hidden state variable while the remaining uncertainty of the noise estimate is described with an observation model. This model allows to distinguish between systematic variations and variations due to the measurement process, which is particularly important for nonstationary noise. The statistical models for speech and noise are overviewed in Section 2. A block EM algorithm for parameter estimation from training data is derived in Section 3. In Section 4, a sequential EM algorithm for online adaptation is derived. Finally, we present experimental results in Section 5 and finish with some conclusions in Section 6.

2 State space models for the noise

Let $\tilde{\mathbf{n}}_t$ denote the cepstral noise vector consisting of $N_c = 13$ static components. Its dynamics are modeled with the linear state model

$$\mathbf{n}_t = \mathbf{n}_{t-1} + \mathbf{v}_t; \quad \mathbf{v}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{V}) \tag{1}$$

$$\tilde{\mathbf{n}}_t = \mathbf{n}_t + \mathbf{w}_t^{(n)}; \quad \mathbf{w}_t^{(n)} \sim \mathcal{N}(\mathbf{0}, \mathbf{W}). \tag{2}$$

Here, $\mathbf{n}_t$ is a hidden state variable. Note, that the measurement equation is for the case of noise-only data $\tilde{\mathbf{n}}_t$. In eq. (1) and eq. (2), $\mathbf{V}$ denotes the state covariance matrix and $\mathbf{W}$ the measurement covariance matrix, respectively.

For the case of noisy speech, further the dynamics of the cepstral feature vector $\mathbf{x}_t$ must be taken into account. Its dynamics are modeled by a switching linear dynamic model (SLDM) according to the state equation

$$\mathbf{x}_t = \mathbf{A}(s_t)\mathbf{x}_{t-1} + \mathbf{b}(s_t) + \mathbf{u}_t, \quad \mathbf{u}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{C}(s_t)) \tag{3}$$

where $\mathbf{A}(s_t)$, $\mathbf{b}(s_t)$ and $\mathbf{C}(s_t)$ are learnt with an EM algorithm from clean speech training data for M different state models $s_t \in \{1, \ldots, M\}$ [4]. Modelling both speech and noise leads to the augmented state vector $\mathbf{z}_t = (\mathbf{x}_t, \mathbf{n}_t)$.

The observation model, which relates the clean speech $\mathbf{x}_t$ and the noise $\mathbf{n}_t$ to the noisy speech cepstral feature vector $\mathbf{y}_t$ is non-linear:

$$\mathbf{y}_t = \mathbf{h}(\mathbf{x}_t, \mathbf{n}_t) = \mathbf{x}_t + \mathbf{M}_{DCT} \log(1 + e^{\mathbf{M}_{DCT}^{-1}(\mathbf{n}_t - \mathbf{x}_t)}). \tag{4}$$

$\mathbf{M}_{DCT}$ and $\mathbf{M}_{DCT}^{-1}$ denote the matrix of the discrete cosine transfom and its (pseudo-) inverse, respectively. The functions log and e have to be understood to operate element-wise on their arguments. Eq. (4) can be linearized around given vector points $\mathbf{x}_t^{(0)}$ and $\mathbf{n}_t^{(0)}$, leading to [14]

$$\mathbf{y}_t \approx \mathbf{h}(\mathbf{x}_t^{(0)}, \mathbf{n}_t^{(0)}) + \mathbf{H}_x(\mathbf{x}_t - \mathbf{x}_t^{(0)}) + \mathbf{H}_n(\mathbf{n}_t - \mathbf{n}_t^{(0)}) + \mathbf{w}_t \tag{5}$$

$$\text{with} \quad \mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{W}_y), \quad \mathbf{W}_y = \mathbf{H}_n \mathbf{W} \mathbf{H}_n' + \mathbf{W}_c \tag{6}$$

and $\mathbf{W}_c$ denoting a small error covariance matrix due to the linearization. The Jacobians are given by

$$\mathbf{H}_x = \mathbf{M}_{DCT} \frac{e^{\mathbf{M}_{DCT}^{-1}\mathbf{x}_t^{(0)}}}{e^{\mathbf{M}_{DCT}^{-1}\mathbf{x}_t^{(0)}} + e^{\mathbf{M}_{DCT}^{-1}\mathbf{n}_t^{(0)}}} \mathbf{M}_{DCT}^{-1}, \quad \mathbf{H}_n = \mathbf{I} - \mathbf{H}_x, \tag{7}$$

where $\mathbf{I}$ denotes the identity matrix. Employing this linearization around the moments of the prior distributions an Extended Kalman Filter (EKF) can be applied for filter update of each of the M models [14]. In the following

the mean vector and the covariance matrix of the posterior of the state vector $\mathbf{z}_t(s_t)$ of the EKF assigned to the state model s_t given the measurements $y_1^t = \mathbf{y}_1 \dots \mathbf{y}_t$ are denoted as $\mathbf{z}_{t|1:t}(s_t)$ and $\mathbf{P}_{t|1:t}(s_t)$ respectively.

3 Block EM algorithm with noise-only training data

In the subsequent section, the parameter estimation with a blockwise EM algorithm on noise-only training data is considered. This allows to perform a smoothing on the complete utterance. Further the parameter estimation is not influenced by the speech signal. We solve the optimization problem

$$\hat{\theta} = \text{argmax}_\theta \{\log p(\tilde{\mathbf{n}}_1^T; \theta)\}, \tag{8}$$

in order to obtain the maximum likelihood estimate $\hat{\theta}$ of the parameter vector $\theta = (\mathbf{V}, \mathbf{W})$. The vector

$$(\mathbf{n}_0^T, \tilde{\mathbf{n}}_1^T) = (\mathbf{n}_0 \dots \mathbf{n}_T, \tilde{\mathbf{n}}_1 \dots \tilde{\mathbf{n}}_T) \tag{9}$$

is complete data for the optimization problem (8). Thus we obtain the EM algorithm

$$\begin{aligned} Q(\theta, \hat{\theta}^{(l)}) &= E_{\hat{\theta}}[\log p(\mathbf{n}_0^T, \tilde{\mathbf{n}}_1^T; \theta) | \hat{\theta}^{(l)}, \tilde{\mathbf{n}}_1^T] \quad (E-step) \\ \hat{\theta}^{(l+1)} &= \text{argmax}_\theta \{Q(\theta, \hat{\theta}^{(l)})\} \qquad (M-step) \end{aligned} \tag{10}$$

The E-step in iteration $l+1$ with the parameters from iteration l is carried out with a Rauch-Tung-Striebel (RTS) smoothing of the extended state vector $\eta_t = (\mathbf{n}_t, \mathbf{n}_{t-1})$. Eq. (1) leads to the state model

$$\eta_t = \mathbf{A}_\eta \eta_{t-1} + \mathbf{v}_t^{(\eta)}, \quad \mathbf{v}_t^{(\eta)} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_\eta) \tag{11}$$

$$\text{with } \mathbf{A}_\eta = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \quad \text{and} \quad \mathbf{C}_\eta = \begin{bmatrix} \mathbf{V} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}. \tag{12}$$

The observation model

$$\tilde{\mathbf{n}}_t = \mathbf{H}_\eta \eta_t + \mathbf{w}_t^{(n)}, \quad \mathbf{w}_t^{(n)} \sim \mathcal{N}(\mathbf{0}, \mathbf{W}) \tag{13}$$

with $\mathbf{H}_\eta = [\mathbf{I} \quad \mathbf{0}]$ is obtained from (2). The probability density $p(\eta_t | \tilde{\mathbf{n}}_1^T)$ is estimated by RTS smoothing [1]. First a Kalman filtering in forward direction is carried out for $t = 1 \dots T$:

$$\begin{aligned} \eta_{t|1:t-1} &= \mathbf{A}_\eta \eta_{t-1|1:t-1} \\ \mathbf{P}_{t|1:t-1}^{(\eta)} &= \mathbf{A}_\eta \mathbf{P}_{t-1|1:t-1}^{(\eta)} \mathbf{A}_\eta' + \mathbf{C}_\eta \\ \mathbf{K}_t^{(\eta)} &= \mathbf{P}_{t|1:t-1}^{(\eta)} \mathbf{H}_\eta' (\mathbf{H}_\eta \mathbf{P}_{t|1:t-1}^{(\eta)} \mathbf{H}_\eta' + \mathbf{W})^{-1} \\ \eta_{t|1:t} &= \eta_{t|1:t-1} + \mathbf{K}_t^{(\eta)} (\tilde{\mathbf{n}}_t - \mathbf{H}_\eta \eta_{t|1:t-1}) \\ \mathbf{P}_{t|1:t}^{(\eta)} &= (\mathbf{I} - \mathbf{K}_t^{(\eta)} \mathbf{H}_\eta) \mathbf{P}_{t|1:t-1}^{(\eta)}, \end{aligned} \tag{14}$$

where $\eta_{t|1:\tau}$ and $\mathbf{P}_{t|1:\tau}^{(\eta)}$ denote the mean and covariance matrix of η_t given the measurements $\tilde{\mathbf{n}}_1^\tau$. Subsequently the smoothing in backward direction is accomplished for $t = T \dots 1$:

$$\begin{aligned} \eta_{t-1|1:T} &= \eta_{t-1|1:t-1} + \mathbf{S}_{t-1}^{(\eta)} (\eta_{t|1:T} - \eta_{t|1:t-1}) \\ \mathbf{P}_{t-1|1:T}^{(\eta)} &= \mathbf{P}_{t-1|1:t-1}^{(\eta)} + \mathbf{S}_{t-1}^{(\eta)} (\mathbf{P}_{t|1:T}^{(\eta)} - \mathbf{P}_{t|1:t-1}^{(\eta)}) \mathbf{S}_{t-1}^{(\eta)\prime} \end{aligned} \tag{15}$$

with the utility variable $\mathbf{S}_{t-1}^{(\eta)} = \mathbf{P}_{t-1|1:t-1}^{(\eta)} \mathbf{A}_\eta' (\mathbf{P}_{t|1:t-1}^{(\eta)})^{-1}$. In order to conduct the M-step, the expectation value

$$\begin{aligned} E[\log p(\mathbf{n}_0^T, \tilde{\mathbf{n}}_1^T; \theta) | \hat{\theta}^{(l)}, \tilde{\mathbf{n}}_1^T] &= E[\log p(\tilde{\mathbf{n}}_1^T | \mathbf{n}_0^T; \theta) | \hat{\theta}^{(l)}, \tilde{\mathbf{n}}_1^T] \\ &\quad + E[\log p(\mathbf{n}_0^T; \theta) | \hat{\theta}^{(l)}, \tilde{\mathbf{n}}_1^T] \end{aligned} \tag{16}$$

is estimated. $E[\log p(\tilde{\mathbf{n}}_1^T | \mathbf{n}_0^T; \theta) | \hat{\theta}^{(l)}, \tilde{\mathbf{n}}_1^T]$ is independent of $\mathbf{V}$ so that $\mathbf{V}$ is obtained by maximization of

$$\begin{aligned} E[\log p(\mathbf{n}_0^T; \theta) | \hat{\theta}^{(l)}, \tilde{\mathbf{n}}_1^T] &= -\frac{N_c T}{2} \log(2\pi) - \frac{T}{2} \log |\mathbf{V}| \\ &\quad - \frac{1}{2} \sum_{t=1}^{T} E[(\mathbf{n}_t - \mathbf{n}_{t-1}) \mathbf{V}^{-1} (\mathbf{n}_t - \mathbf{n}_{t-1})' | \hat{\theta}^{(l)}, \tilde{\mathbf{n}}_1^T]. \end{aligned} \tag{17}$$

with the feature vector dimension N_c. It can be shown that the maximization of (17) leads to the ML estimates [13]

$$\hat{\mathbf{V}} = \frac{1}{T} \sum_{t=1}^{T} E[(\mathbf{n}_t - \mathbf{n}_{t-1})(\mathbf{n}_t - \mathbf{n}_{t-1})' | \hat{\theta}^{(l)}, \tilde{\mathbf{n}}_1^T]. \tag{18}$$

The correlations in eq. (18) can be obtained from the moments

$$\eta_{t|1:T} = \begin{bmatrix} \mathbf{n}_{t|1:T} \\ \mathbf{n}_{t-1|1:T} \end{bmatrix}, \quad \mathbf{P}_{t|1:T}^{(\eta)} = \begin{bmatrix} \mathbf{P}_{t,t|1:T}^{(n)} & \mathbf{P}_{t,t-1|1:T}^{(n)} \\ \mathbf{P}_{t-1,t|1:T}^{(n)} & \mathbf{P}_{t-1,t-1|1:T}^{(n)} \end{bmatrix} \tag{19}$$

of the posterior $p(\eta_t | \tilde{\mathbf{n}}_1^T)$, which is calculated in the E-step:

$$E[\mathbf{n}_{t-i} \mathbf{n}_{t-j}' | \hat{\theta}^{(l)}, \tilde{\mathbf{n}}_1^T] = \mathbf{n}_{t-i|1:T} \mathbf{n}_{t-j|1:T}' + \mathbf{P}_{t-i,t-j|1:T}^{(n)}, \tag{20}$$

with $i, j \in \{0, 1\}$.
The maximization of $E[\log p(\tilde{\mathbf{n}}_1^T | \mathbf{n}_0^T; \theta) | \tilde{\mathbf{n}}_1^T]$ leads in analogy to the estimate of the measurement covariance matrix

$$\hat{\mathbf{W}} = \frac{1}{T} \sum_{t=1}^{T} E[(\tilde{\mathbf{n}}_t - \mathbf{n}_t)(\tilde{\mathbf{n}}_t - \mathbf{n}_t)' | \tilde{\mathbf{n}}_1^T] \tag{21}$$

$$\text{with} \quad E[\mathbf{n}_t | \tilde{\mathbf{n}}_1^T] = \mathbf{n}_{t|1:T} \tag{22}$$

$$E[\mathbf{n}_t \mathbf{n}_t' | \tilde{\mathbf{n}}_1^T] = \mathbf{n}_{t|1:T} \mathbf{n}_{t|1:T}' + \mathbf{P}_{t,t|1:T}^{(n)}. \tag{23}$$

To summarize, the EM algorithm consists in the iteration of the parameter estimation according to (18) and (21) and the RTS smoothing in (14) and (15).

4 Sequential EM algorithm for online adaptation

In practice, changing noise conditions in the environment require an online adaptation of the noise model parameters. The application of an online block EM algorithm would be very time consuming due to the iterated filtering with SLDMs which was not required in Section 3 due to the assumption of noise-only data. For this reason, in the following a sequential EM algorithm is derived which allows further a causal processing and the adaptation of the noise model parameters to changing noise conditions

within a single utterance. In informal experiments an online estimate of the state covariance matrix $\mathbf{V}$ turned out to be very unreliable. Therefore the sequential EM algorithm is in the following applied for online estimation of the measurement noise covariance only while the state noise covariance is assumed to be known, i.e. its value is determined from training data as described in Section 3. The cost function for noisy input data is

$$\begin{aligned} Q(\theta,\hat{\theta}^{(l)}) = E[\log p(\mathbf{y}_1^T|\mathbf{z}_0^T;\theta)|\hat{\theta}^{(l)},\mathbf{y}_1^T] \\ + E[\log p(\mathbf{z}_0^T;\theta)|\hat{\theta}^{(l)},\mathbf{y}_1^T]. \end{aligned} \tag{24}$$

In eq. (24) only $E[\log p(\mathbf{y}_1^T|\mathbf{z}_0^T;\theta)]$ depends on the covariance matrix $\mathbf{W}$ of the observation noise, so that the other term can be dropped. Thus, we optimize the causal cost function

$$\begin{aligned} Q_t^{(W)}(\mathbf{W},\hat{\mathbf{W}}_1^{t-1}) &= E[\log p(\mathbf{y}_1^t|\mathbf{z}_0^t;\mathbf{W})|\hat{\mathbf{W}}_1^{t-1},\mathbf{y}_1^t] \\ &\approx -\sum_{\tau=1}^{t} R_\tau(\mathbf{W},\hat{\mathbf{W}}_{\tau-1}) \end{aligned} \tag{25}$$

with

$$\begin{aligned} &R_\tau(\mathbf{W},\hat{\mathbf{W}}_{\tau-1}) = -E[\log p(\mathbf{y}_\tau|\mathbf{z}_\tau;\mathbf{W})|\hat{\mathbf{W}}_{\tau-1},\mathbf{y}_1^\tau] \\ &= -E[\log \sum_{s_\tau=1}^{M} P(s_\tau|\mathbf{y}_1^\tau)p(\mathbf{y}_\tau|\mathbf{z}_\tau,s_\tau;\mathbf{W})|\hat{\mathbf{W}}_{\tau-1},\mathbf{y}_1^\tau] \\ &\approx -E[\log p(\mathbf{y}_\tau|\mathbf{z}_\tau,\hat{s}_\tau;\mathbf{W})|\hat{\mathbf{W}}_{\tau-1},\mathbf{y}_1^\tau] \\ &= \log((2\pi)^{N_c}|\mathbf{W}_y|) + E[\Delta\mathbf{y}_\tau'\mathbf{W}_y^{-1}\Delta\mathbf{y}_\tau|\hat{\mathbf{W}}_{\tau-1},\mathbf{y}_1^\tau], \end{aligned} \tag{26}$$

with the measurement error $\Delta\mathbf{y}_\tau = \mathbf{y}_\tau - \mathbf{h}(\mathbf{z}_\tau)$. In eq. (26) only measurements up to the current time instance τ are considered in order to allow a causal processing. Further, the weighted sum over the models s_τ is replaced by the most probable model

$$\hat{s}_t = \arg\max_{s_\tau}\{P(s_\tau|\mathbf{y}_1^\tau)\}. \tag{27}$$

The expectation value in (26) can be written as

$$\begin{aligned} &E[\Delta\mathbf{y}_\tau'\mathbf{W}_y^{-1}\Delta\mathbf{y}_\tau|\hat{\mathbf{W}}_{\tau-1},\mathbf{y}_1^\tau] \\ &= \mathrm{tr}(\mathbf{W}_y^{-1}E[\Delta\mathbf{y}_\tau\Delta\mathbf{y}_\tau'|\hat{\mathbf{W}}_{\tau-1},\mathbf{y}_1^\tau]) \\ &= \mathrm{tr}(\mathbf{W}_y^{-1}(\Delta\mathbf{y}_{\tau|1:\tau}\Delta\mathbf{y}_{\tau|1:\tau}' + \mathbf{P}_{\tau|1:\tau}^{(y)})). \end{aligned} \tag{28}$$

$$\text{where} \quad \mathbf{P}_{\tau|1:\tau}^{(y)} = [\mathbf{H}_x,\mathbf{H}_n]\mathbf{P}_{\tau|1:\tau}(s_\tau)[\mathbf{H}_x,\mathbf{H}_n]' \tag{29}$$

$$\Delta\mathbf{y}_{\tau|1:\tau} = \mathbf{y}_\tau - \mathbf{h}(\mathbf{z}_{\tau|1:\tau}) \tag{30}$$

are obtained from the state estimate of the most likely EKF $\hat{s}_\tau$. $Q_t^{(W)}(\mathbf{W},\hat{\mathbf{W}}_1^{t-1})$ can be written as

$$Q_t^{(W)}(\mathbf{W},\hat{\mathbf{W}}_1^{t-1}) = Q_{t-1}^{(W)}(\mathbf{W},\hat{\mathbf{W}}_1^{t-2}) - R_t(\mathbf{W},\hat{\mathbf{W}}_{t-1}). \tag{31}$$

In order to weaken the influence of incorrect measurements $\mathbf{y}_t$ on the cost function $Q_t^{(W)}(\mathbf{W},\hat{\mathbf{W}}_1^{t-1})$, the novel contribution $R_t(\mathbf{W},\mathbf{W}_{t-1})$ to the cost function $Q_t^{(W)}(\mathbf{W},\hat{\mathbf{W}}_1^{t-1})$ is deemphasized by a factor γ, $0<\gamma<1$. In the following $\mathbf{W}$ is further modelled as a diagonal matrix $\mathrm{diag}(w^{(1)},\ldots,w^{(i)},\ldots,w^{(N_c)})$, with the components $w^{(i)}$, $i=1\ldots N_c$ on the main diagonal, which are collected in a vector $\mathbf{w}=(w^{(1)},\ldots,w^{(i)},\ldots,w^{(N_c)})'$. This leads to the modified cost function $\tilde{Q}_t^{(W)}(\mathbf{w},\mathbf{w}_1^{t-1})$:

$$\tilde{Q}_t^{(W)}(\mathbf{w},\mathbf{w}_1^{t-1}) = \tilde{Q}_{t-1}^{(W)}(\mathbf{w},\mathbf{w}_1^{t-2}) - \gamma\tilde{R}_t(\mathbf{w},\mathbf{w}_{t-1}). \tag{32}$$

In analogy to [3] the following Newton-like recursion is obtained, which can be applied for a sequential update of the noise variance:

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \gamma\mathbf{K}_t^{-1}\mathbf{s}_t \tag{33}$$

with $\mathbf{s}_t = \frac{\partial\tilde{R}_t}{\partial\mathbf{w}}\Big|_{\mathbf{w}=\mathbf{w}_{t-1}}$, $\mathbf{K}_t = -\frac{\partial^2\tilde{Q}_t^{(W)}}{\partial\mathbf{w}\partial\mathbf{w}'}\Big|_{\mathbf{w}=\mathbf{w}_{t-1}}$.
The components of $\mathbf{s}_t$ are obtained by partial derivation of $\tilde{R}_t$ with respect to $w^{(i)}$ [13]:

$$\begin{aligned} \mathbf{s}_t^{(i)} = \frac{\partial\tilde{R}_t(\mathbf{w})}{\partial w^{(i)}} = -\frac{1}{2}\sum_{k=1}^{N_c}\sum_{l=1}^{N_c}(2-\delta_{kl})\frac{\partial w_y^{(k,l)}}{\partial w^{(i)}} \\ \cdot\mathbf{c}_l'\mathbf{W}_y{}^{-1}(\Delta\mathbf{y}_{t|1:t}\Delta\mathbf{y}_{t|1:t}' + \mathbf{P}_{t|1:t}^{(y)} - \mathbf{W}_y)\mathbf{W}_y{}^{-1}\mathbf{c}_k. \end{aligned} \tag{34}$$

In eq. (34), $\mathbf{c}_l = [0\ldots0\quad 1\quad 0\ldots0]'$ denotes a column vector, where the l-th component has the value one and δ_{kl} is the Kronecker symbol. The partial derivations of the components $w_y^{(l,k)}$ of the matrix $\mathbf{W}_y(\mathbf{W})$ (see (6)) with respect to $w^{(i)}$ are

$$\frac{\partial w_y^{(l,k)}}{\partial w^{(i)}} = h_n^{(l,i)}h_n^{(k,i)}. \tag{35}$$

where $h_n^{(l,i)}$ is the value in the l-th row and i-th column of $\mathbf{H}_n$. Derivation of eq. (32) leads to the recursion

$$\mathbf{K}_t = \mathbf{K}_{t-1} - \gamma\mathbf{L}_t \tag{36}$$

for $\mathbf{K}_t$ with $\mathbf{L}_t = \frac{\partial^2 R_t}{\partial\mathbf{w}\partial\mathbf{w}'}\Big|_{\mathbf{w}=\mathbf{w}_{t-1}}$. The components of $\mathbf{L}_t$ are obtained by partial derivation of R_t with respect to $w^{(i)}$ and $w^{(j)}$ [13]:

$$\begin{aligned} L_t^{(i,j)} &= \frac{\partial^2\tilde{R}_t(\mathbf{w})}{\partial w^{(i)}\partial w^{(j)}} \\ &= \frac{1}{2}\sum_{k=1}^{N_c}\sum_{l=1}^{N_c}\sum_{m=1}^{N_c}\sum_{n=1}^{N_c}(2-\delta_{kl})(2-\delta_{mn})\frac{\partial w_y^{(k,l)}}{\partial w^{(i)}}\frac{\partial w^{(m,n)}}{\partial w^{(j)}} \\ &\quad\cdot\mathbf{c}_k'[\mathbf{W}_y{}^{-1}\mathbf{c}_m\mathbf{c}_n'\mathbf{W}_y{}^{-1}(\Delta\mathbf{y}_{t|1:t}\Delta\mathbf{y}_{t|1:t}' + \mathbf{P}_{t|1:t}^{(y)})\mathbf{W}_y^{-1} \\ &\quad + \mathbf{W}_y^{-1}(\Delta\mathbf{y}_{t|1:t}\Delta\mathbf{y}_{t|1:t}' + \mathbf{P}_{t|1:t}^{(y)})\mathbf{W}_y^{-1}\mathbf{c}_m\mathbf{c}_n'\mathbf{W}_y{}^{-1} \\ &\quad - \mathbf{W}_y^{-1}\mathbf{c}_m\mathbf{c}_n'\mathbf{W}_y{}^{-1}]\mathbf{c}_l. \end{aligned} \tag{37}$$

A simplification of eq. (34) and eq. (37) can be achieved by modelling the covariance matrix $\mathbf{P}_{t|1:t}^{(y)}$ as diagonal matrix with the elements $w_y^{(k,k)}$ and neglecting the terms $\frac{\partial w_y^{(k,k)}}{\partial w^{(i)}}$ for $i\neq k$. Thus, a step of the sequential EM algorithm at time instance t consists in the operations listed in Alg. 1.

Algorithm 1 Step of the sequential EM algorithm

1: Determine the most probable model $\hat{s}_t$ of the SLDM proposed in [14] with (27).
2: Determine $\mathbf{W}^{(y)}(\hat{s}_t)$, $\mathbf{P}^{(y)}_{t|1:t}(\hat{s}_t)$ and $\Delta\mathbf{y}_{t|1:t}(\hat{s}_t)$ for state model $\hat{s}_t$ with (6), (29) and (30)).
3: Calculate $\mathbf{s}_t$ with (34).
4: Calculate $\mathbf{L}_t$ with (37).
5: Update $\mathbf{K}_t$ with (36).
6: Update $\mathbf{W}^{(n)}_t$ with (33).

5 Experimental results

The experiments were performed on the AURORA4 database. The AURORA4 test database consists of the Wallstreet Journal Nov'92 evaluation test set to which noise at varying SNR levels and varying type has been added [6]. In our experiments, we use the official AURORA4 selection test set comprising 166 utterances which was recorded with a Sennheiser microphone. The downsampled version of the data with a sampling rate of $8kHz$ was employed. We use, beside the clean data of the AURORA4 selection test set, six further versions of the test set with artificially added noise at a randomly chosen SNR between $5dB$ and $15dB$. We present the results for each noise type and the overall average obtained by using a bigram language model for the 5000-word vocabulary. Training has been carried out on clean speech. We modified the ETSI standard front-end extraction in the same manner as in [4] by replacing the energy feature with c_0 and using the squared power spectral density rather than the spectral magnitude as the input of the Mel-frequency filter-bank. The speech feature enhancement was carried out with the SLDM introduced in Section 2 which was augmented with dynamic speech features as described in [14] (EKF-d).

The results for the noise estimation methods introduced in this paper are given in Tab. 1. The error rate for the baseline method (EKF-d) under the assumption of stationary noise and noise parameter estimation from the first and last 10 frames of each utterance amounts 37.6%. The noise estimation was at first conducted with the given simplifications of (34) and (37). For $\mathbf{V} = \mathbf{0}$ (M-0it), i.e. under the assumption of a stationary noise process the adaptation of the noise covariance matrix resulted in an improved error rate of 36.2%. The estimation of $\mathbf{V}$ with 50 iterations of the block-wise EM algorithm (M-50it) resulted in an overall error rate of 35.5%, while lesser improvements were obtained for 10 and 25 iterations (M-10it, M-25it). Further

	Tra.	Air.	Bab.	Car	Str.	Res.	Cln.	Avg.
EKF-d	41.5	50.8	44.9	18.9	41.6	52.3	12.9	37.6
M-0it	39.7	50.1	41.8	18.4	40.6	49.7	12.9	36.2
M-10it	45.0	44.8	38.3	18.7	43.9	50.1	12.9	36.1
M-25it	43.0	44.2	37.9	18.4	42.7	49.9	12.9	35.6
M-50it	42.4	44.2	37.9	18.4	42.7	49.9	12.9	35.5
M-full	42.7	43.9	38.5	17.8	43.5	49.7	12.9	35.6

Table 1: Error rates on the AURORA4 database

the results with 50 iterations of the block-wise EM algorithm (M-full) for the evaluation of the complete sums in eq. (34) and eq. (37) are shown in Tab. 1. The given approximations lead approximately to the same recognition rate on the AURORA4 database. The computational complexity was even in the case of the complete implementation of the sums in eq. (34) und eq. (37) not significantly increased by the sequential EM algorithm compared to the model-based speech feature enhancement with the baseline EKF-d.

6 Conclusions

In the paper at hand, EM algorithms for noise estimation in model-based speech feature enhancement were derived. With the sequential EM algorithm even in the case of a stationary noise model significant improvements in recognition rate were obtained. Further improvements were achieved with the assumption of a non-zero state model variance which was estimated with a block-wise EM algorithm from training data. A simplification of the sequential EM algorithm was possible by neglecting off-diagonal terms. The overall computational complexity of the model-based speech feature enhancement was not significantly improved by the application of the EM algorithms.

Acknowledgment

The research was partly supported by the DFG Research Training Group GK-693 of the Paderborn Institute for Scientific Computation (PaSCo).

Literatur

[1] Rong Li X. Kirubarajan T. Bar-Shalom, Y. *Estimation with Applications to Tracking and Navigation*. John Wiley and Sons, Inc., 2001.

[2] I. Cohen and B. Berdugo. Noise estimation by minima controlled recursive averaging for robust speech enhancement. In *IEEE Signal Processing Letters*, volume 9, pages 12–15, 2002.

[3] Droppo J. Deng, L and A. Acero. Recursive estimation of nonstationary noise using iterative stochastic approximation for robust speech recognition. *Transactions on SAP*, 11:568–580, 2003.

[4] J. Droppo and A. Acero. Noise robust speech recognition with a switching linear dynamic model. In *ICASSP*, pages 953–956, 2004.

[5] Deng L. Acero A. Frey, B. and T. Kristjansson. Algonquin: Iterating laplace's method to remove multiple types of acoustic distortion for robust speech recognition. In *Eurospeech*, 2003.

[6] G. Hirsch. Experimental framework for the performance evaluation of speech recognition front-ends on a large vocabulary task, 2002. STQ AURORA DSR WORKING GROUP.

[7] N.S. Kim. Imm-based estimation for slowly evolving environments. In *IEEE Signal Processing Letters*, volume 5, pages 146–149, 1998.

[8] N.S. Kim. Nonstationary environment compensation based on sequential estimation. In *ISPL*, volume 5, pages 57–59, 1998.

[9] N.S. Kim, D.Y. Kim, B.G. Kong, and S.R. Kim. Application of vts to environment compensation with noise statistics. In *Proc. ESCA Workshop on Robust Speech Recognition for Unknown Communication Channels*, pages 99–102.

[10] R. Martin. Noise power spectral density estimation based on optimal smoothing and minimum statistics. In *IEEE Transactions on Speech and Audio Processing*, volume 9, pages 504–512, 2001.

[11] P.J. Moreno. Speech recognition in noisy environments, 1996. Ph.D. Thesis, Carnegie Mellon University.

[12] R. Singh and B. Raj. Tracking noise via dynamical systems with a continuum of states. In *IEEE ICASSP*, 2003.

[13] S. Windmann. *Exploitation of temporal redundancies of the cepstral speech features for ASR*. PhD thesis, University of Paderborn, to appear 2008.

[14] S. Windmann and R. Haeb-Umbach. Modelling the dynamics of speech and noise. In *ICASSP*, 2008.