MODELING THE DYNAMICS OF SPEECH AND NOISE FOR SPEECH FEATURE ENHANCEMENT IN ASR

Stefan Windmann, Reinhold Haeb-Umbach

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

ABSTRACT

In this paper a Switching Linear Dynamical Model (SLDM) approach for speech feature enhancement is improved by employing more accurate models for the dynamics of speech and noise. The model of the clean speech feature trajectory is improved by augmenting the state vector to capture information derived from the delta features. Further a hidden noise state variable is introduced to obtain a more elaborated model for the noise dynamics. Approximate Bayesian inference in the SLDM is carried out by a bank of Extended Kalman filters, whose outputs are combined according to the a posteriori probability of the individual state models. Experimental results on the AURORA2 database show improved recognition accuracy.

Index Terms— ASR, speech recognition, speech feature enhancement, inter-frame correlation, SLDM

1. INTRODUCTION

Recently, state-space estimation techniques have been proposed for speech feature enhancement for noisy speech recognition [1], [2], [3], [4], [5]. However, the use of state estimation techniques faces two major difficulties: First, the clean speech trajectory cannot be well modelled by a single linear dynamical model; and second, the observation model, which relates the observed noisy speech to the clean speech feature vector, is highly non-linear.

A promising approach to tackle the first issue is the use of switching linear dynamical models (SLDM), where the feature trajectory is described by a number of linear models, and a so-called regime variable (a "switch") determines which model is active at a given time [1]. Recently, inference in this model has been improved by feeding back information from the speech recognizer to estimate the value of the regime variable [5].

To address the second issue, a variety of observation models has been proposed in the literature [2], [6], [7], [8], [9]. In the context of SLDMs, a method based on an iterative improvement of a SNR variable and linear and statistical linearizations in the context of Extended and Unscented Kalman Filters have been proposed.

In this paper we present an improved state dynamics and observation model, which is based on exploiting the information present in first and second order dynamical feature vector components. To arrive at a tractable solution, Principal Component Analysis is applied to capture this information in a few additional components of the state vector. A novel dynamical model of the noise is also introduced.

The paper is organized as follows. In the next section the concept of SLDM-based speech feature enhancement is briefly summarized, where a VTS-based linearisation of the observation model leads to a bank of Extended Kalman Filters. In Section 3 the improved dynamical model is presented. A novel state-space model of the noise is introduced in Section 4, and Section 5 gives experimental results. The paper finishes with conclusions drawn in Section 6.

2. SPEECH FEATURE ENHANCEMENT WITH SLDMS

Let $\mathbf{x}_t$ denote the clean cepstral feature vector consisting of 13 static components. Its dynamics are modeled by a SLDM according to the state equation

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

where $\mathbf{A}(s_t)$, $\mathbf{b}(s_t)$ and $\mathbf{C}(s_t)$ are learnt with the EM algorithm from clean speech training data. Here, $s_t \in \{1, \ldots, M\}$ is a state or regime variable which can assume M different values. During inference it has to be estimated alongside $\mathbf{x}_t$.

In order to take into account the time-varying characteristics of the background noise we also adopt a state space model of the (cepstral) noise process:

$$\mathbf{n}_t = \mathbf{D}\mathbf{n}_{t-1} + \mathbf{e} + \mathbf{v}_t, \mathbf{v}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{F}) \quad (2)$$

with noise state transition matrix $\mathbf{D}$, bias $\mathbf{e}$ and system noise $\mathbf{v}_t$. We will discuss different options for this noise model in Section 4.

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

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

$$\mathbf{y}_t = \mathbf{h}(\mathbf{x}_t, \mathbf{n}_t) = \mathbf{x}_t + \mathbf{M}_{DCT} \log(1 + e^{\mathbf{M}_{DCT}^{-1}(\mathbf{n}_t - \mathbf{x}_t)}). \quad (3)$$

Here, $\mathbf{M}_{DCT}$ 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 elementwise on their arguments.

Eq. (3) can be linearized by Vector Taylor Series expansion around given vector points $\mathbf{x}_t^{(0)}$ and $\mathbf{n}_t^{(0)}$, leading to

$$\mathbf{y}_t \approx \mathbf{h}(\mathbf{x}_t^{(0)}, \mathbf{n}_t^{(0)}) + \mathbf{H}_x(\mathbf{x}_t - \mathbf{x}_t^{(0)}) + \mathbf{H}_n(\mathbf{n}_t - \mathbf{n}_t^{(0)}) + \mathbf{w}_t \quad (4)$$

with $\mathbf{w}_t \sim N(\mathbf{0}, \mathbf{V})$, $\mathbf{I}$ denoting the identity matrix, and the Jacobians

$$\mathbf{H}_x = \mathbf{M}_{DCT} \frac{e^{\mathbf{M}_{DCT}^{-1}\mathbf{x}_t^{(0)}}}{e^{\mathbf{M}_{DCT}^{-1}\mathbf{x}_t^{(0)}} + e^{\mathbf{M}_{DCT}^{-1}\mathbf{n}_t^{(0)}}} \mathbf{M}_{DCT}^{-1}, \mathbf{H}_n = \mathbf{I} - \mathbf{H}_x \quad (5)$$

Employing this linearization around the moments of the a priori distributions an Extended Kalman Filter can be applied for filter update of each of the M models. Let $\mathbf{z}_t = (\mathbf{x}_t^T, \mathbf{n}_t^T)^T$ denote the combined state vector of speech and noise. Then the equations of the Extended Kalman Filter can be compactly written as follows:

$$\begin{aligned}
\mathbf{z}_{t|t-1}(s_t) &= \mathbf{A}_z(s_t)\mathbf{z}_{t-1|t-1} + \mathbf{b}_z(s_t) \\
\mathbf{P}_{t|t-1}(s_t) &= \mathbf{A}_z(s_t)\mathbf{P}_{t-1|t-1} + \mathbf{C}_z(s_t) \\
\mathbf{K}_t(s_t) &= \mathbf{P}_{t|t-1}(s_t)\mathbf{H}_z^T(\mathbf{H}_z\mathbf{P}_{t|t-1}(s_t)\mathbf{H}_z^T + \mathbf{V})^{-1} \\
\mathbf{z}_{t|t}(s_t) &= \mathbf{z}_{t|t-1}(s_t) + \mathbf{K}_t(s_t)(\mathbf{y}_t - \mathbf{h}(\mathbf{z}_{t|t-1}(s_t))) \\
\mathbf{P}_{t|t}(s_t) &= (\mathbf{I} - \mathbf{K}_t(s_t)\mathbf{H}_z)\mathbf{P}_{t|t-1}(s_t)
\end{aligned} \quad (6)$$

with

$$\begin{aligned}
\mathbf{A}_z(s_t) &= \begin{bmatrix} \mathbf{A}(s_t) & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{bmatrix}, \quad \mathbf{b}_z(s_t) = \begin{bmatrix} \mathbf{b}(s_t) \\ \mathbf{e} \end{bmatrix}, \\
\mathbf{C}_z(s_t) &= \begin{bmatrix} \mathbf{C}(s_t) & \mathbf{0} \\ \mathbf{0} & \mathbf{E} \end{bmatrix}, \quad \mathbf{H}_z = \begin{bmatrix} \mathbf{H}_x & \mathbf{H}_n \end{bmatrix}
\end{aligned} \quad (7)$$

and the Kalman gain $\mathbf{K}_t(s_t)$. The variables $\mathbf{z}_{t|t}(s_t)$, $\mathbf{P}_{t|t}(s_t)$, $\mathbf{z}_{t|t-1}(s_t)$ and $\mathbf{P}_{t|t-1}(s_t)$ are the moments of the state-conditional probability density functions

$$\begin{aligned}
p(\mathbf{z}_t|\mathbf{y}_1^{t-1}, s_t) &= \mathcal{N}(\mathbf{z}_{t|t-1}(s_t), \mathbf{P}_{t|t-1}(s_t)) \\
p(\mathbf{z}_t|\mathbf{y}_1^{t}, s_t) &= \mathcal{N}(\mathbf{z}_{t|t}(s_t), \mathbf{P}_{t|t}(s_t)).
\end{aligned} \quad (8)$$

of the state vector $\mathbf{z}_t$ at frame t, given all past observations up to frame $t-1$ or t, respectively. The estimation of the overall posterior $p(\mathbf{z}_t|\mathbf{y}_1^t)$ is computationally intractable, however suboptimal solutions can be obtained by the Generalized Pseudo-Bayesian (GPB) or the Interacting Multiple Model (IMM) algorithm.

3. INCORPORATION OF THE DYNAMIC FEATURES

In HMM-based speech recognition dynamic features are usually incorporated in the feature vector in order to partly compensate for the modeling errors introduced by the conditional independence assumption, which states that dependence between successive feature vectors is captured solely by the HMM states. Indeed, dynamic features have been shown to significantly improve recognition accuracy. They are usually computed over several frames and can therefore capture dependencies beyond what a first-order Markov model can do. In this section an approach is introduced where the dynamic features are incorporated efficiently in the Extended Kalman Filter in order to obtain a robust model of the speech dynamics.

The state vector $(\mathbf{x}, \mathbf{n})$ of the Extended Kalman Filter is expanded to the vector $(\mathbf{x}, \delta\mathbf{x}, \delta^2\mathbf{x}, \mathbf{n}, \delta\mathbf{n}, \delta^2\mathbf{n})$ while the measurement vector $\mathbf{y}$ is replaced by the vector $(\mathbf{y}, \delta\mathbf{y}, \delta^2\mathbf{y})$. Here, δ and δ^2 shall indicate first- (velocity) and second-order (acceleration) time derivatives of the respective feature. An approximative measurement model for the dynamic features in the log-spectral domain has been derived in [10]:

$$\begin{aligned}
\delta\tilde{\mathbf{y}} &\approx \frac{e^{\tilde{\mathbf{x}}}}{e^{\tilde{\mathbf{x}}} + e^{\tilde{\mathbf{n}}}}\delta\tilde{\mathbf{x}} + \frac{e^{\tilde{\mathbf{n}}}}{e^{\tilde{\mathbf{x}}} + e^{\tilde{\mathbf{n}}}}\delta\tilde{\mathbf{n}} \\
\delta^2\tilde{\mathbf{y}} &\approx \frac{e^{\tilde{\mathbf{x}}}}{e^{\tilde{\mathbf{x}}} + e^{\tilde{\mathbf{n}}}}\delta^2\tilde{\mathbf{x}} + \frac{e^{\tilde{\mathbf{n}}}}{e^{\tilde{\mathbf{x}}} + e^{\tilde{\mathbf{n}}}}\delta^2\tilde{\mathbf{n}} \\
&\quad + \frac{e^{(\tilde{\mathbf{x}}+\tilde{\mathbf{n}})}}{(e^{\tilde{\mathbf{x}}} + e^{\tilde{\mathbf{n}}})^2}[(\delta\tilde{\mathbf{x}})^2 + (\delta\tilde{\mathbf{n}})^2 - 2\delta\tilde{\mathbf{x}}\delta\tilde{\mathbf{n}}]
\end{aligned} \quad (9)$$

Here, $\tilde{\mathbf{x}}$, $\tilde{\mathbf{n}}$ and $\tilde{\mathbf{y}}$ are the log-spectral feature vectors of speech, noise and noisy speech. Vector Taylor Series expansion in the Cepstrum leads to the approximative measurement model

$$\begin{bmatrix} \mathbf{y} \\ \delta\mathbf{y} \\ \delta^2\mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{h}(\mathbf{x}^{(0)}, \mathbf{n}^{(0)}) \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix} + \mathbf{H} \begin{bmatrix} \mathbf{x} - \mathbf{x}^{(0)} \\ \delta\mathbf{x} - \delta\mathbf{x}^{(0)} \\ \delta^2\mathbf{x} - \delta^2\mathbf{x}^{(0)} \\ \mathbf{n} - \mathbf{n}^{(0)} \\ \delta\mathbf{n} - \delta\mathbf{n}^{(0)} \\ \delta^2\mathbf{n} - \delta^2\mathbf{n}^{(0)} \end{bmatrix} + \begin{bmatrix} \mathbf{w} \\ \delta\mathbf{w} \\ \delta^2\mathbf{w} \end{bmatrix} \quad (10)$$

with the Jacobian

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_x & \mathbf{0} & \mathbf{0} & \mathbf{H}_n & \mathbf{0} & \mathbf{0} \\ \mathbf{H}_{21} & \mathbf{H}_x & \mathbf{0} & \mathbf{H}_{24} & \mathbf{H}_n & \mathbf{0} \\ \mathbf{H}_{31} & \mathbf{H}_{32} & \mathbf{H}_x & \mathbf{H}_{34} & \mathbf{H}_{35} & \mathbf{H}_n \end{bmatrix}, \quad (11)$$

The matrices $\mathbf{H}_{ij}$ denote the Jacobians from the i-th subvector of the observation to the j-th subvector of the state vector; e.g. $\mathbf{H}_{32} = \partial(\delta^2\mathbf{y})/\partial(\delta\mathbf{x})$.

The matrices $\mathbf{H}_{21}$, $\mathbf{H}_{31}$, $\mathbf{H}_{32}$, $\mathbf{H}_{24}$, $\mathbf{H}_{34}$ and $\mathbf{H}_{35}$ are approximated by zero for the following considerations. Note that the state vectors $(\mathbf{x}, \delta\mathbf{x}, \delta^2\mathbf{x})$ and $(\mathbf{n}, \delta\mathbf{n}, \delta^2\mathbf{n})$ contain 39

components each. Such a high dimension makes the calculations in the Extended Kalman filters very time consuming and bares the danger that the training of the state matrices is susceptible to irregularities in the training data. For this reason the subvector $(\delta\mathbf{x}, \delta^2\mathbf{x})$ is mapped into a lower-dimensional space by Principal Component Analysis (PCA):

$$\delta_{\mathbf{x}} = \mathbf{M}_{PCA} \begin{bmatrix} \delta\mathbf{x} \\ \delta^2\mathbf{x} \end{bmatrix}, \tag{12}$$

with dimension $\dim(\delta_{\mathbf{x}}) \leq \dim((\delta\mathbf{x}, \delta^2\mathbf{x}))$. The subvectors $(\delta\mathbf{y}, \delta^2\mathbf{y})$ and $(\delta\mathbf{n}, \delta^2\mathbf{n})$ are likewise reduced to $\delta_{\mathbf{y}}$ and $\delta_{\mathbf{n}}$. The PCA matrix M_{PCA} is determined from clean speech training data. The observation model for the modified state and observation vector can be written as

$$\begin{bmatrix} \mathbf{y} \\ \delta_{\mathbf{y}} \end{bmatrix} = \begin{bmatrix} \mathbf{h}(\mathbf{x}^{(0)}, \mathbf{n}^{(0)}) \\ \mathbf{0} \end{bmatrix} + \tilde{\mathbf{H}} \begin{bmatrix} \mathbf{x} - \mathbf{x}^{(0)} \\ \delta_{\mathbf{x}} - \delta_{\mathbf{x}}{}^{(0)} \\ \mathbf{n} - \mathbf{n}^{(0)} \\ \delta_{\mathbf{n}} - \delta_{\mathbf{n}}{}^{(0)} \end{bmatrix} + \begin{bmatrix} \mathbf{w} \\ \delta_{\mathbf{w}} \end{bmatrix}, \tag{13}$$

with

$$\tilde{\mathbf{H}} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_{PCA} \end{bmatrix} \mathbf{H} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_{PCA}^{-1} \end{bmatrix} \tag{14}$$

where M_{PCA}^{-1} denotes the right-inverse matrix of M_{PCA}. Correspondingly, the state model is extended to

$$\begin{bmatrix} \mathbf{x}_{t+1} \\ \delta_{\mathbf{x}t+1} \end{bmatrix} = \tilde{\mathbf{A}} \begin{bmatrix} \mathbf{x}_t \\ \delta_{\mathbf{x}t} \end{bmatrix} + \tilde{\mathbf{b}} + \tilde{\mathbf{u}}_t, \tilde{\mathbf{u}}_t \sim \mathcal{N}(\mathbf{0}, \tilde{\mathbf{C}}(s_t)) \tag{15}$$

where the parameters $\tilde{\mathbf{A}}(s_t)$, $\tilde{\mathbf{b}}(s_t)$ and $\tilde{\mathbf{C}}(s_t)$ are learned in the same manner as described in [1] from training data by application of an EM algorithm.

4. MODELLING THE NOISE DYNAMICS

In the framework of the SLDM described in Section 2 stationary noise conditions can be modeled by setting $\mathbf{D} = \mathbf{0}$, $\mathbf{e} = \boldsymbol{\mu_n}$, and $\mathbf{F}$ equal to the covariance matrix of the noise in eq. (2). Mean $\boldsymbol{\mu_n}$ and covariance of the cepstral noise can be estimated from the first and last frames in the sentence, which are assumed to contain noise only [1] [3]. This model can be extended to capture non-stationary noise by letting $\mathbf{D} \neq \mathbf{0}$ resulting in a state space model of noise dynamics which captures inter-frame correlations of the noise process [2] [4]. While in [4] the state model for the noise dynamics according to eq. (2) is employed, Kim et al. [2] used the simplified noise model:

$$\mathbf{n}_t = \mathbf{n}_{t-1} + \mathbf{v}_t, \mathbf{v}_t \sim N(\mathbf{0}, \boldsymbol{\epsilon}). \tag{16}$$

where $\boldsymbol{\epsilon}$ was a priori fixed so that no training data was required. In both approaches it was assumed that during absence of speech the noise process $\mathbf{n}_t$ is observable.

In practice, the observed noise process is highly fluctuating. If the model (2) or (16) were used, this fluctuation would have to be attributed completely to the non-stationarity of the noise. The fluctuation, however, is not necessarily due to non-stationarity, but, probably to a greater extent, due to the observation process (e.g. variance of the periodogram estimation). Therefore it seems to be reasonable to use a state model for a hidden noise variable corresponding to the time-varying expectation value of the nonstationary noise and to relate the remainder of the uncertainty to an observation process:

$$\mathbf{n}'_t = \mathbf{n}'_{t-1} + \mathbf{v}_t \tag{17}$$

$$\mathbf{n}_t = \mathbf{n}'_t + \mathbf{w}_t^{(n)} \tag{18}$$

with $\mathbf{v}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\epsilon}')$ and $\mathbf{w}_t^{(n)} \sim \mathcal{N}(\mathbf{0}, \mathbf{V}^{(n)})$.
The state equation (17) results from eq. (2) by setting

$$\mathbf{D} = \mathbf{I}, \quad \mathbf{e} = \mathbf{0}, \quad \mathbf{F} = \boldsymbol{\epsilon}' \tag{19}$$

while the variance $\mathbf{V}$ of the measurement noise in eq. (4) is increased by

$$\mathbf{H}_n \mathbf{V}^{(n)} \mathbf{H}_n^T. \tag{20}$$

to account for the observation noise in (18).

In order to take the unreliability of the noise measurement during active speech into account a VAD is applied to reset $p(\mathbf{n}'_t|\mathbf{y}_1^t, s_t)$ to the prior $p(\mathbf{n}'_t|\mathbf{y}_1^{t-1}, s_t)$ after the measurement update in speech phases.

The crucial part in the parameter estimation is the determination of the state model variance $\boldsymbol{\epsilon}'$ which is assumed to be a very small constant. For this purpose a Minimum Statistics approach in the Cepstrum is applied. A window of length $N^{(win)}$ with the center position position $P_t^{(win)}$ is slided over the data. $\mathbf{n}'_t$ is estimated to be the minimum of the cepstral values in the window around $P_t^{(win)}$, leading to the estimate

$$\boldsymbol{\epsilon}' \approx E[(\mathbf{n}'_t - \mathbf{n}'_{t-1})(\mathbf{n}'_t - \mathbf{n}'_{t-1})^T]. \tag{21}$$

In order to determine $\mathbf{V}^{(n)}$ it is assumed that the noise is approximately stationary in the first and last noise-only frames of a sentence, such that the variance can be estimated to be

$$\mathbf{V}^{(n)} = \frac{N_{start}}{N_{start} + N_{end}} \boldsymbol{\Sigma}_{\mathbf{n}}{}^{(start)} + \frac{N_{end}}{N_{start} + N_{end}} \boldsymbol{\Sigma}_{\mathbf{n}}{}^{(end)}, \tag{22}$$

where the noise covariances $\boldsymbol{\Sigma}_{\mathbf{n}}{}^{(start)}$ and $\boldsymbol{\Sigma}_{\mathbf{n}}{}^{(end)}$ are obtained from the first N_{start} and the last N_{end} noise-only frames in the sentence respectively.

5. EXPERIMENTAL RESULTS

The experiments were performed on the AURORA2 database with clean speech training data. We modified the ETSI standard front-end extraction in the same manner as in [1] 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 overall recognition accuracy was averaged over all noise conditions at

SNR levels between 0dB and 20dB. The speech recognition with the described standard frontend (SFE) yielded an overall recognition accuracy of 60.37% on test set A and 56.37% on test set B (table 1).

Set A	Subw.	Bab.	Car	Exh.	Avg.
SFE	68.06	45.87	58.34	64.76	60.37
SLDM	80.19	72.56	84.28	82.43	79.87
SLDM2	80.16	70.57	86.38	82.33	79.86
SLDM2-d	82.31	74.49	87.24	82.81	81.71
SLDM2-d2	81.51	73.32	86.82	82.17	80.95
SLDM2-dn	82.07	75.90	86.90	83.12	82.00

Set B	Rest.	Street	Airp.	Train	Avg.
SFE	52.07	65.50	52.72	55.19	56.37
SLDM	74.39	79.29	79.05	83.91	79.16
SLDM2	76.57	75.92	80.82	81.37	78.67
SLDM2-d	75.74	80.73	80.88	83.90	80.31
SLDM2-dn	76.27	79.57	82.04	83.57	80.36

Table 1. Word accuracy on the AURORA 2 database at different noise conditions

The SLDM proposed in [1] led to an accuracy of 79.87% and 79.16% respectively. With the SLDM described in Section 2 (SLDM2) approximately the same overall accuracy was achieved, while the computational costs were slightly reduced, as there was no iterative estimation of an SNR variable involved. By augmenting the state vector of clean speech by a single additional component, which captures information from the delta features according to the method described in Section 3, the recognition accuracy was improved to 81.71% and 80.31% respectively (SLDM-d). Using one more feature, i.e. $\dim(\delta_{\mathbf{x}}) = 2$ (SLDM2-d2), part of the improvement was lost again on set A so that no further experiments on set B were conducted. In control experiments (not reported in the table) we observed the same gain of about 2% on test set A when using the iterative Extended Kalman filter proposed in [7] or when carrying out smoothing, i.e. estimating the posterior $p(\mathbf{z}_t|\mathbf{y}_1^T)$, which is conditioned on all frames of an utterance, rather than $p(\mathbf{z}_t|\mathbf{y}_1^t)$, by employing the Rauch-Tung-Striebel algorithm. However, these two methods are computationally more expensive than the one proposed here. Unfortunately the gains of the individual methods did not add up when used in combination.

With the dynamic noise model (SLDM-dn) the recognition accuracy is improved for instationary noise conditions (e.g. Babble, Exhibition) while it is decreased at more stationary noise conditions (e.g. Subway, Car). In all experiments the parameters of the different noise models were estimated from the first and last 10 frames of each sentence.

6. CONCLUSIONS AND OUTLOOK

In this paper SLDM-based speech feature enhancement with a bank of Extended Kalman Filters was investigated. The recognition results could be significantly improved by incorporating a single dynamic feature in the state model of clean speech. Further improvements for instationary noise conditions were achieved by employing a state model for the noise dynamics which incorporates a hidden, unobservable state variable. Currently we are working on improving the state model of noise by estimating the model parameters with the EM algorithm.

7. REFERENCES

[1] J. Droppo and A. Acero, "Noise robust speech recognition with a switching linear dynamic model," in *ICASSP*, 2004.

[2] N.-S. Kim, W. Lim, and R.-M. Stern, "Feature compensation based on switching linear dynamic model," in *IEEE Signal Processing Letters*, 2005.

[3] V. Stouten, H. Van hamme, and P. Wambacq, "Kalman and unscented kalman filter feature enhancement for noise robust asr," in *Interspeech*, 2005.

[4] J. Deng, M. Bouchard, and T.H. Yeap, "Speech feature estimation under the presence of noise with a switching linear dynamic model," in *ICASSP*, 2006.

[5] S. Windmann and R. Haeb-Umbach, "An iterative approach to speech feature enhancement and recognition," in *Interspeech*, 2007.

[6] P.J. Moreno, B. Raj, and R.M. Stern, "A vector taylor series approach for environment-independent speech recognition," in *ICASSP*, 1996.

[7] B.-J. Frey, L. Deng, A. Acero, and T. Kristjansson, "Algonquin: Iterating laplace's method to remove multiple types of acoustic distortion for robust speech recognition," in *EUROSPEECH*, 2001.

[8] J. Droppo, L. Deng, and A. Acero, "A comparison of three non-linear observation models for noisy speech features," in *EUROSPEECH*, 2003.

[9] M. Afify, "Accurate compensation in the log-spectral domain for noisy speech recognition," in *ICASSP*, 2005.

[10] M.J.F. Gales, "Model-based techniques for noise robust speech recognition," in *Ph.D. Thesis, University of Cambridge*, 1995.