Switching Linear Dynamic Models for Noise Robust In-Car Speech Recognition

Björn Schuller[1], Martin Wöllmer[1], Tobias Moosmayr[2], Günther Ruske[1], and Gerhard Rigoll[1]

[1] Technische Universität München, Institute for Human-Machine Communication, 80290 München, Germany
schuller@tum.de
[2] BMW Group, Forschungs- und Innovationszentrum, Akustik, Komfort und Werterhaltung, 80788 München, Germany

1 Introduction

Aiming to counter the performance degradation of speech recognition systems in noisy surroundings, as for example the interior of a car, a variety of different concepts have been developed in recent years. The common goal of all noise compensation strategies is to minimize the mismatch between training and recognition conditions, which occurs whenever the speech signal is distorted by noise. Consequently two main methods can be distinguished: one is to reduce the mismatch by focusing on adapting the acoustic models to noisy conditions. This can be achieved by either using noisy training data or by joint speech and noise modeling. The other method is trying to determine the clean features from the noisy speech sequence while using clean training data.

Preprocessing techniques for speech enhancement aim to compensate the effects of noise before the feature-based speech representation is classified by the recognizer which has been trained on clean data. The state-of-the-art speech signal preprocessing that is used as a baseline feature extraction algorithm for noisy speech recognition problems is the Advanced Front End (AFE) two-step Wiener filtering concept introduced in [1]. As shown in [2], methods based on spectral subtraction like Cepstral Mean Subtraction (CMS) [3] or Unsupervised

Spectral Subtraction (USS) [4] reach similar performance while requiring less computational cost than Wiener filtering. Further attempts to reduce the mismatch between test and training conditions are Mean and Variance Normalization (MVN) [5] or Histogram Equalization (HEQ) [6], [7], a technique which is often used in digital image processing to improve the contrast of pictures. In speech processing HEQ is a powerful method to improve the temporal dynamics of feature vector components distorted by noise.

This paper examines a model based preprocessing approach to enhance noisy features as it is proposed in [8]. Here a Switching Linear Dynamic Model (SLDM), which can be considered as Kalman filter, is used to describe the dynamics of speech while another linear dynamic model captures the dynamics of additive noise. Both models serve to derive an observation model describing how speech and noise produce the noisy observations and to reconstruct the features of clean speech.

A second technique for noise robust speech recognition using Kalman filtering is outlined and applied in the noisy speech recognition task of this work. This method was first introduced in [9] where a Switching Autoregressive Hidden Markov Model (SAR-HMM) had been extended to an Autoregressive Switching Linear Dynamical System (AR-SLDS) for improved noise robustness. Similar to the SLDM, the AR-SLDS includes an explicit noise model by modeling the dynamics of both the raw speech signal and the noise. However, the technique does not model feature vectors like the SLDM, but the raw speech signal in the time domain.

The paper is organized as follows: Section 2 outlines the SLDM used for feature enhancement in this work, while Section 3 introduces the SAR-HMM which is embedded into an AR-SLDS in Section 4. Both Kalman filtering approaches are evaluated in a noisy isolated digit recognition task in Section 5.

2 Switching Linear Dynamic Models

Model based speech enhancement techniques are based on modeling speech and noise. Together with a model of how speech and noise produce the noisy observations, these models are used to enhance the noisy speech features. In [8] a Switching Linear Dynamic Model is used to capture the dynamics of clean speech. Similar to Hidden Markov Model (HMM) based approaches to model clean speech, the SLDM assumes that the signal passes through various states. Conditioned on the state sequence the SLDM furthermore enforces a continuous state transition in the feature space.

2.1 Modeling of Noise

Unlike speech, which is modeled applying an SLDM, the modeling of noise is done by using a simple Linear Dynamic Model (LDM) obeying the following system equation:

$$x_t = Ax_{t-1} + b + v_t \tag{1}$$

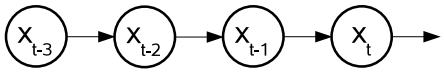

Fig. 1. Linear Dynamic Model for noise

Thereby the matrix A and the vector b simulate how the noise process evolves over time and v_t represents a Gaussian noise source driving the system. A graphical representation of this LDM can be seen in Figure 1. As LDM are time-invariant, they are suited to model signals like colored stationary Gaussian noise. Alternatively to the graphical model in Figure 1 the equations

$$p(x_t|x_{t-1}) = \mathcal{N}(x_t; Ax_{t-1} + b, C) \tag{2}$$

$$p(x_{1:T}) = p(x_1) \prod_{t=2}^{T} p(x_t|x_{t-1}) \tag{3}$$

can be used to express the LDM. Here, $\mathcal{N}(x_t; Ax_{t-1} + b, C)$ is a multivariate Gaussian with mean vector $Ax_{t-1} + b$ and covariance matrix C, whereas T denotes the length of the input sequence.

2.2 Modeling of Speech

The modeling of speech is realized by a more complex dynamic model which also includes a hidden state variable s_t at each time t. Now A and b depend on the state variable s_t:

$$x_t = A(s_t)x_{t-1} + b(s_t) + v_t \tag{4}$$

Consequently every possible state sequence $s_{1:T}$ describes an LDM which is non-stationary due to A and b changing over time. Time-varying systems like the evolution of speech features over time can be described adequately by such models. As can be seen in Figure 2, it is assumed that there are time dependencies among the continuous variables x_t, but not among the discrete state variables s_t. This is the major difference between the SLDM used in [8] and the models used in [10] where time dependencies among the hidden state variables are included. A modification like this can be seen as analogous to extending a Gaussian Mixture Model (GMM) to an HMM. The SLDM corresponding to Figure 2 can be described as follows:

$$p(x_t, s_t|x_{t-1}) = \mathcal{N}(x_t; A(s_t)x_{t-1} + b(s_t), C(s_t)) \cdot p(s_t) \tag{5}$$

$$p(x_{1:T}, s_{1:T}) = p(x_1, s_1) \prod_{t=2}^{T} p(x_t, s_t|x_{t-1}) \tag{6}$$

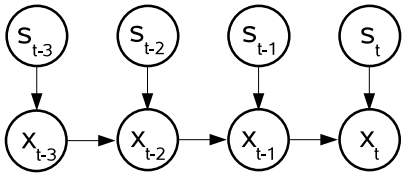

Fig. 2. Switching Linear Dynamic Model for speech

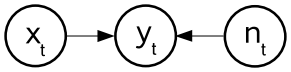

Fig. 3. Observation model for noisy speech y_t

To train the parameters $A(s)$, $b(s)$ and $C(s)$ of the SLDM, conventional EM techniques are used [11]. Setting the number of states to one corresponds to training a Linear Dynamic Model instead of an SLDM to obtain the parameters A, b and C needed for the LDM which is used to model noise.

2.3 Observation Model

In order to obtain a relationship between the noisy observation and the hidden speech and noise features, an observation model has to be defined. Figure 3 illustrates the graphical representation of the zero variance observation model with SNR inference introduced in [12]. Thereby it is assumed that speech x_t and noise n_t mix linearly in the time domain corresponding to a non-linear mixing in the cepstral domain.

2.4 Posterior Estimation and Enhancement

A possible approximation to reduce the computational complexity of posterior estimation is to restrict the size of the search space applying the generalized pseudo-Bayesian (GPB) algorithm [13]. The GPB algorithm is based on the assumption that the distinct state histories whose differences occur more than r frames in the past can be neglected. Consequently, if T denotes the length of the sequence, the inference complexity is reduced from S^T to S^r whereas $r \ll T$. Using the GPB algorithm, the three steps *collapse*, *predict* and *observe* are conducted for each speech frame [8].

The Gaussian posterior obtained in the observation step of the GPB algorithm is used to obtain estimates of the moments of x_t. Those estimates represent the de-noised speech features and can be used for speech recognition in noisy environments. Thereby the clean features are assumed to be the Minimum Mean Square Error (MMSE) estimate $E[x_t|y_{1:t}]$.

3 Switching Autoregressive Hidden Markov Models

An alternative to conventional HMM modeling of speech is the modeling of the raw signal directly in the time domain. As proven in [14] and [15], modeling the raw signal can be a reasonable alternative to feature-based approaches. In [9] a Switching Autoregressive HMM is applied for isolated digit recognition. The SAR-HMM is based on modeling the speech signal as an autoregressive (AR) process whereas the non-stationarity of human speech is captured by the switching between a number of different AR parameter sets. This is done by a discrete switch variable s_t that can be seen as analogon to the HMM states.

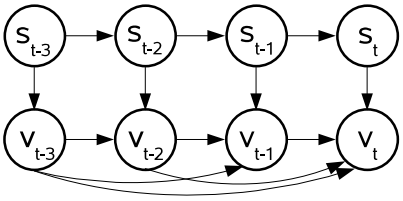

Fig. 4. Dynamic Bayesian Network structure of the SAR-HMM

One of S different states can be occupied at each time step t. Thereby the state variable indicates which AR parameter set to use at the given time instant t. Here, the time index t denotes the samples in the time domain and not the feature vectors as in Section 2. The current state only depends on the preceding state with transition probability $p(s_t|s_{t-1})$. Furthermore it is assumed that the current sample v_t is a linear combination of the R preceding samples superposed by a Gaussian distributed innovation $\eta(s_t)$. Both $\eta(s_t)$ and the AR weights $c_r(s_t)$ depend on the current state s_t:

$$v_t = -\sum_{r=1}^{R} c_r(s_t) v_{t-r} + \eta(s_t) \tag{7}$$

with

$$\eta \sim \mathcal{N}(\eta; 0, \sigma^2(s_t))$$

The purpose of $\eta(s_t)$ is not to model an independent additive noise process but to model variations from pure autoregression. For the SAR-HMM the joint probability of a sequence of length T is

$$p(s_{1:T}, v_{1:T}) = p(v_1|s_1)p(s_1) \prod_{t=2}^{T} p(v_t|v_{t-R:t-1}, s_t)p(s_t|s_{t-1}) \tag{8}$$

corresponding to the Dynamic Bayesian Network (DBN) structure illustrated in Figure 4.

As the number of samples in the time domain which are used as input for the SAR-HMM is usually a lot higher than the number of feature vectors observed by an HMM, it is necessary to ensure that the switching between the different AR models is not too fast. This is granted by forcing the model to stay in the same state for an integer multiple of K time steps.

The training of the AR parameters is realized applying the EM algorithm [11]. To infer the distributions $p(s_t|v_{1:T})$ a technique based on the forward-backward algorithm [16] is used. Due to the fact that an observation v_t depends on R preceding observations (see Figure 4) the backward pass is more complicated for the SAR-HMM than for a conventional HMM. To overcome this problem a *correction smoother* as derived in [17] is applied which means that the backward pass computes the posterior $p(s_t|v_{1:T})$ by *correcting* the output of the forward pass.

4 Autoregressive Switching Linear Dynamical Systems

To improve noise robustness, the SAR-HMM can be embedded into an AR-SLDS to include an explicit noise process as shown in [9]. The AR-SLDS interprets the observed speech sample v_t as a noisy version of a hidden clean sample. Thereby the clean signal can be obtained from the projection of a hidden vector h_t which has the dynamic properties of a Linear Dynamical System:

$$h_t = A(s_t)h_{t-1} + \eta_t^{\mathcal{H}} \tag{9}$$

with

$$\eta_t^{\mathcal{H}} \sim \mathcal{N}\big(\eta_t^{\mathcal{H}}; 0, \Sigma_{\mathcal{H}}(s_t)\big)$$

The dynamics of the hidden variable are defined by the transition matrix $A(s_t)$ which depends on the current state s_t. Variations from pure linear state dynamics are modeled by the Gaussian distributed hidden "innovation" variable $\eta_t^{\mathcal{H}}$. Similar to the variable η_t used in Equation 7 for the SAR-HMM, $\eta_t^{\mathcal{H}}$ does *not* model an independent additive noise source. To obtain the current observed sample, the vector h_t is projected onto a scalar v_t as follows:

$$v_t = Bh_t + \eta_t^{\mathcal{V}} \tag{10}$$

with

$$\eta_t^{\mathcal{V}} \sim \mathcal{N}(\eta_t^{\mathcal{V}}; 0, \sigma_{\mathcal{V}}^2)$$

The variable $\eta_t^{\mathcal{V}}$ thereby models independent additive white Gaussian noise which is supposed to corrupt the hidden clean sample Bh_t.

Figure 5 visualizes the structure of the SLDS modeling the dynamics of the hidden clean signal, as well as independent additive noise.

The SLDS parameters $A(s_t)$, B and $\Sigma_{\mathcal{H}}(s_t)$ can be defined in a way that the obtained SLDS mimics the SAR-HMM derived in Section 3 for the case $\sigma_{\mathcal{V}} = 0$ (see [9]). This has the advantage that in case $\sigma_{\mathcal{V}} \neq 0$ a noise model is included without having to train new models. Since inference calculation for the AR-SLDS is computationally intractable, the *Expectation Correction* algorithm developed in [18] is applied to reduce the complexity.

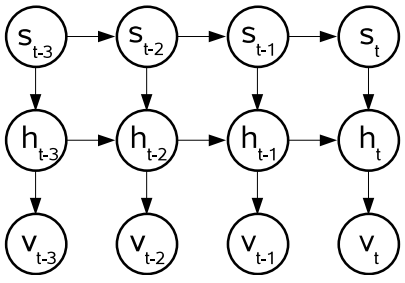

Fig. 5. Dynamic Bayesian Network structure of the AR-SLDS

5 Experiments

For noisy speech recognition experiments we use the digits "zero" to "nine" from the TI 46 Speaker Dependent Isolated Word Corpus [19]. The database contains utterances from 16 different speakers - 8 female and 8 male speakers. For the sake of better comparability with the results presented in [9], only the words which are spoken by male speakers are used. For every speaker 26 utterances were recorded per word class whereas 10 samples are used for training and 16 for testing. Consequently, the overall training corpus consists of 80 utterances per class while the test set contains 128 samples per class. As in [9], all utterances were downsampled to 8000 Hz.

The in-car noise database which was used as additive noise source in this work is the same as in [20]. The noise recordings aim to simulate a wide range of different car types and driving conditions such as driving on big cobbles ("COB") at 30 km/h, over a smooth city road surface ("CTY") at 50 km/h, and on a highway ("HWY") at 120 km/h. Thereby four different car types are considered: BMW 530i (Touring), BMW 645Ci (Convertible), BMW M5 (Sedan) and Mini Cooper S (Convertible). Even though the soft top of both convertibles was closed during recording, the worst case noise scenario is represented by the MINI convertible driving over cobbles (see Figure 6).

In spite of SNR levels below 0 dB the noisy test sequences are still well audible since the recorded noise samples are lowpass signals with most of their energy in the frequency band from 0 to 500 Hz. Consequently, there is no full overlap of the spectrum of speech and noise.

Two further noise types were used: first, a mixture of babble and street noise ("BAB") recorded in downtown Munich. This noise type is relevant for in-car speech recognition performance when driving in an urban area with open windows. The babble and street noise was superposed with the clean speech

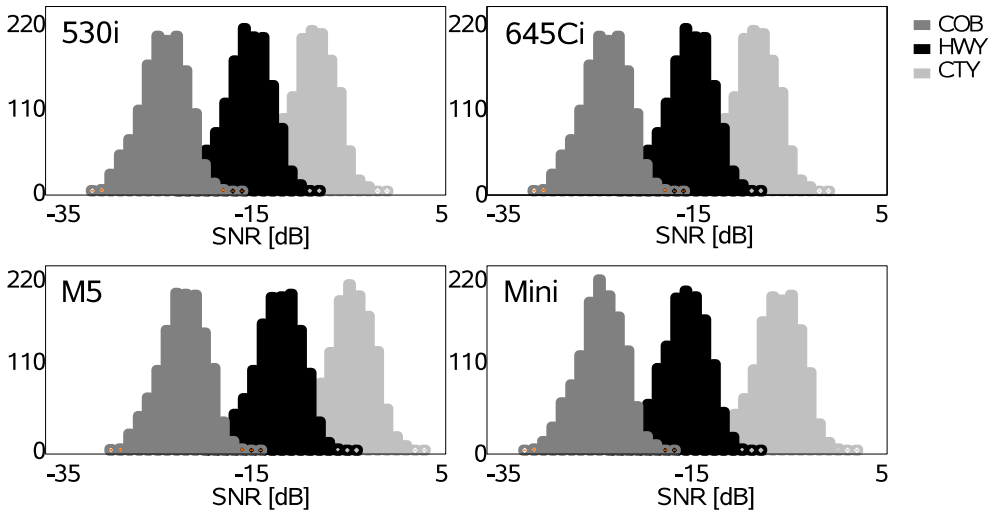

Fig. 6. SNR distribution of TI 46 (digits) utterances superposed with car noise: frequency of occurrence versus SNR level

Table 1. Mean recognition rates for different speech enhancement and modeling methods; Mean recognition rate without speech enhancement: 76.69%

Method	Recognition Rate	Method	Recognition Rate
SLDM	97.65%	USS	87.25%
HEQ	94.77%	AFE	85.53%
CMS	93.24%	AR-SLDS	63.60%
MVN	92.39%	SAR-HMM	59.18%

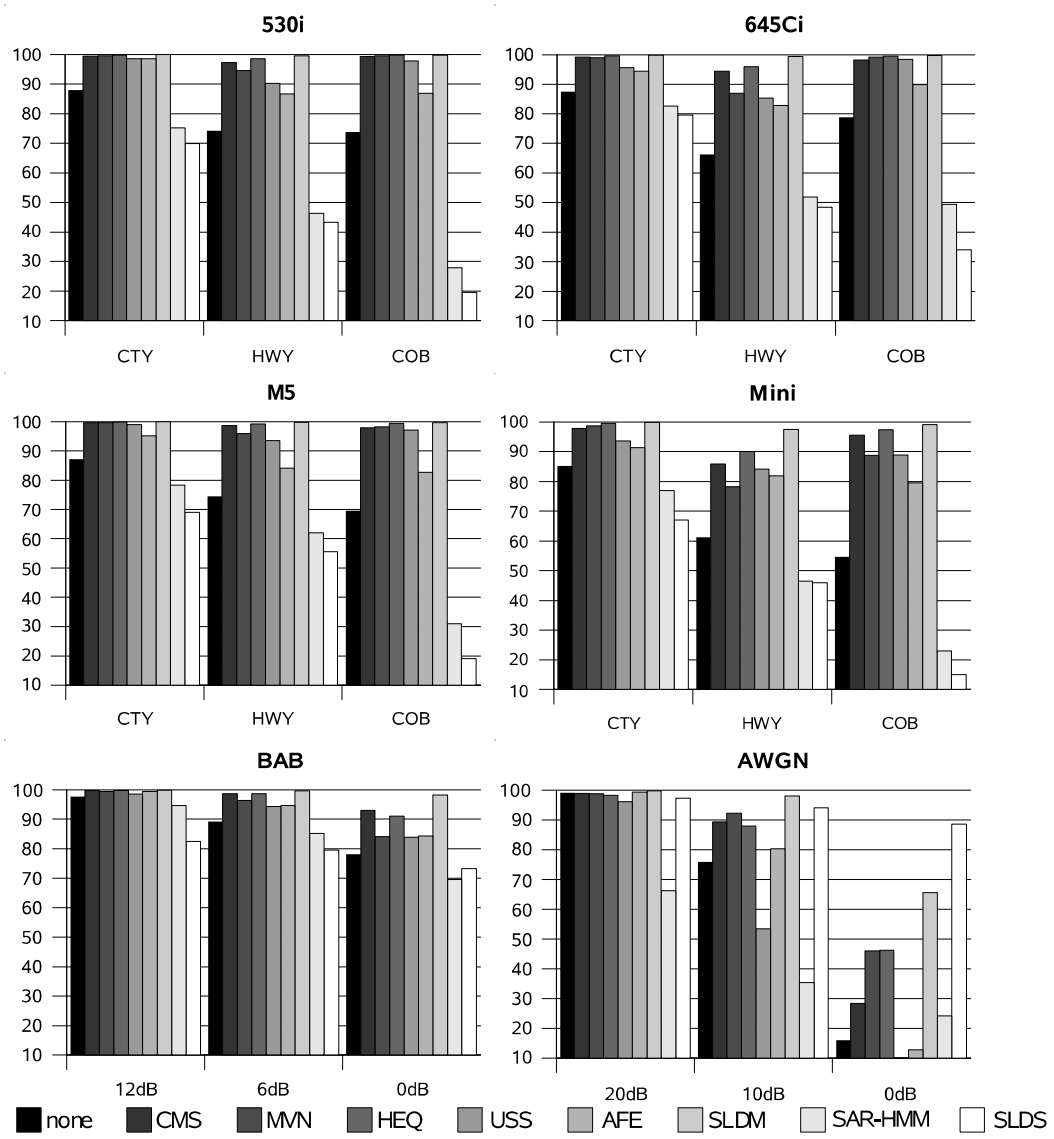

Fig. 7. Recognition rate in percent versus driving condition and SNR level respectively using different speech enhancement and modeling techniques

utterances at SNR levels 12 dB, 6 dB and 0 dB. Furthermore, additive white Gaussian noise ("AWGN") has been used in the experiments. Thereby the SNR levels 20 dB, 10 dB and 0 dB were taken into account.

For every digit from "zero" to "nine" an HMM consisting of 8 states with a mixture of 3 Gaussians per state was trained, except for the SAR-HMM and AR-SLDS experiments where speech was modeled by a 10^{th} order AR process with 10 states. Feature vectors consisted of 13 MFCC as well as the first and second order derivatives of the cepstral coefficients. A global speech SLDM of 32 hidden states was trained, whereas the enhancement algorithm was run with the history parameter $r = 1$. The LDM modeling stationary noise was trained for each noisy test sequence using the first and last 10 frames of the utterance. In addition to the Kalman filter based speech modeling and enhancement concepts explained in Section 2 and 4, a variety of different standard feature enhancement techniques as named in Section 1 were evaluated. Figure 7 shows the performance of the different speech enhancement strategies for different noise types. Thereby training was carried out using clean data. With a recognition rate of 97.65% averaged over all noise types (see Table 1), the SLDM outperformed all other feature enhancement and modeling techniques for each of the car and babble noise types, whereas for AWGN at low SNR levels the AR-SLDS performed best (recognition rate of 88.52% for AWGN at 0 dB SNR). However, the AR-SLDS was not suited to model colored noise such as noise occurring in the interior of a car.

6 Conclusion

In this paper we compared two techniques for noise robust in-car speech recognition based on Kalman filtering and joint speech and noise modeling. The strategy of describing the dynamics of speech with a Switching Linear Dynamic Model while modeling noise as linear dynamic process is able to outperform other known speech enhancement approaches like Wiener filtering or Histogram Equalization whenever speech is corrupted by colored noise produced while driving a car. Speech disturbed by white noise can best be modeled using an Autoregressive Switching Linear Dynamical System which captures the speech and noise dynamics of the raw signal in the time domain.

Acknowledgement

We would like to thank Jasha Droppo and Bertrand Mesot for providing SLDM and AR-SLDS binaries.

References

1. Speech processing, transmission and quality aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms. ETSI standard doc. ES 202 050 V1.1.5 (2007)
2. Lathoud, G., Doss, M.M., Boulard, H.: Channel normalization for unsupervised spectral subtraction. In: Proceedings of ASRU (2005)

3. Rahim, M.G., Juang, B.H., Chou, W., Buhrke, E.: Signal conditioning techniques for robust speech recognition. IEEE Signal Processing Letters, 107–109 (1996)
4. Lathoud, G., Magimia-Doss, M., Mesot, B., Boulard, H.: Unsupervised spectral subtraction for noise-robust ASR. In: Proceedings of ASRU, pp. 189–194 (2005)
5. Viikki, O., Laurila, K.: Cepstral domain segmental feature vector normalization for noise robust speech recognition. Speech Communication, 133–147 (1998)
6. de la Torre, A., Peinado, A.M., Segura, J.C., Perez-Cordoba, J.L., Benitez, M.C., Rubio, A.J.: Histogram equalization of speech representation for robust speech recognition. IEEE Transactions on Speech and Audio Processing, 355–366 (2005)
7. Hilger, F., Ney, H.: Quantile based histogram equalization for noise robust speech recognition. In: Eurospeech, pp. 1135–1138 (2001)
8. Droppo, J., Acero, A.: Noise robust speech recognition with a switching linear dynamic model. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (2004)
9. Mesot, B., Barber, D.: Switching linear dynamical systems for noise robust speech recognition. IEEE Transactions on Audio, Speech and Language Processing (2007)
10. Deng, J., Bouchard, M., Yeap, T.H.: Noisy speech feature estimation on the Aurora2 database using a switching linear dynamic model. Journal of Multimedia, 47–52 (2007)
11. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 1–38 (1977)
12. Droppo, J., Deng, L., Acero, A.: A comparison of three non-linear observation models for noisy speech features. In: Eurospeech, pp. 681–684 (2003)
13. Bar-Shalom, Y., Li, X.R.: Estimation and tracking: principles, techniques, and software. Artech House, Norwood, MA (1993)
14. Ephraim, Y., Roberts, W.J.J.: Revisiting autoregressive hidden Markov modeling of speech signals. IEEE Signal Processing Letters, 166–169 (2005)
15. Poritz, A.: Linear predictive hidden Markov models and the speech signal. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 1291–1294 (1982)
16. Baum, L.E., Petrie, T.: Statistical inference for probabilistic functions of finite state Markov chains. The Annals of Mathematical Statistics, 1554–1563 (1966)
17. Rauch, H.E., Tung, G., Striebel, C.T.: Maximum likelihood estimates of linear dynamic systems. Journal of American Instituite of Aeronautics and Astronautics, 1445–1450 (1965)
18. Barber, D.: Expectation correction for smoothed inference in switching linear dynamical systems. Journal of Machine Learning Reseach, 2515–2540 (2006)
19. Doddington, G.R., Schalk, T.B.: Speech recognition: turning theory to practice. IEEE Spectrum, 26–32 (1981)
20. Grimm, M., Kroschel, K., Harris, H., Nass, C., Schuller, B., Rigoll, G., Moosmayr, T.: On the necessity and feasibility of detecting a driver's emotional state while driving. In: Paiva, A., Prada, R., Picard, R.W. (eds.) ACII 2007. LNCS, vol. 4738, pp. 126–138. Springer, Heidelberg (2007)