Hindawi Publishing Corporation
EURASIP Journal on Audio, Speech, and Music Processing
Volume 2009, Article ID 942617, 17 pages
doi:10.1155/2009/942617

Research Article

Recognition of Noisy Speech: A Comparative Survey of Robust Model Architecture and Feature Enhancement

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

[1] *Institute for Human-Machine Communication, Technische Universität München (TUM), 80290 Munich, Germany*
[2] *BMW Group, Forschungs- und Innovationszentrum, Akustik, Komfort und Werterhaltung, 80788 München, Germany*

Correspondence should be addressed to Björn Schuller, schuller@tum.de

Received 28 October 2008; Revised 21 January 2009; Accepted 15 February 2009

Recommended by Li Deng

Performance of speech recognition systems strongly degrades in the presence of background noise, like the driving noise inside a car. In contrast to existing works, we aim to improve noise robustness focusing on all major levels of speech recognition: feature extraction, feature enhancement, speech modelling, and training. Thereby, we give an overview of promising auditory modelling concepts, speech enhancement techniques, training strategies, and model architecture, which are implemented in an in-car digit and spelling recognition task considering noises produced by various car types and driving conditions. We prove that joint speech and noise modelling with a Switching Linear Dynamic Model (SLDM) outperforms speech enhancement techniques like Histogram Equalisation (HEQ) with a mean relative error reduction of 52.7% over various noise types and levels. Embedding a Switching Linear Dynamical System (SLDS) into a Switching Autoregressive Hidden Markov Model (SAR-HMM) prevails for speech disturbed by additive white Gaussian noise.

1. Introduction

The automatic recognition of speech, enabling a natural and easy to use method of communication between human and machine, is an active area of research as it still suffers from limitations such as the restricted applicability whenever human speech is superposed with background noise [1–3]. Since the interior of a car is a popular field of application for speech recognisers, allowing hands-free operation of the centre console or text messaging, the car noises produced during driving are of great interest when designing a noise robust speech recognition system [4, 5].

To enhance recognition performance in noisy surroundings, different stages of the recognition process have to be optimised. As a first step, filtering or spectral subtraction can be applied to improve the signal before speech features are extracted. Well-known examples for such approaches are applied in the advanced front-end feature extraction (AFE) or Unsupervised Spectral Subtraction (USS). Then, suitable patterns for auditory modelling have to be extracted from the speech signal to allow a reliable distinction between the phonemes or word classes in the vocabulary of the recogniser. Apart from widely used features like Mel-frequency cepstral coefficients (MFCCs), the extraction of Perceptual Linear Prediction (PLP) coefficients is an effective method of speech representation [6].

The third stage is the enhancement of the obtained features to remove the effects of noise. Normalisation methods like Cepstral Mean Subtraction (CMS) [7], Mean and Variance Normalisation (MVN) [8], or Histogram Equalisation (HEQ) [9] are techniques to reduce distortions of the frequency domain representation of speech. Alternatively, model-based feature enhancement approaches can be applied to compensate the effects of background noise. Using a Switching Linear Dynamic Model (SLDM) to capture the dynamic behaviour of speech and another Linear Dynamic Model (LDM) to describe additive noise is the strategy of the joint speech and noise modelling concept in [10] which aims to estimate the clean speech features of the noisy signal.

The derivation of speech models can be considered as the next stage in the design of a speech recogniser. Hidden Markov Models (HMMs) [11] are commonly used for speech modelling whereas numerous alternatives, like Hidden Conditional Random Fields (HCRFs) [12], Switching Autoregressive Hidden Markov Models (SAR-HMMs)

[13], or other more general Dynamic Bayesian Network structures have been developed in recent years. Extending the SAR-HMM to an Autoregressive Switching Linear Dynamical System (AR-SLDS), as in [14], includes an explicit noise model and leads to an increased noise robustness compared to the SAR-HMM.

Speech models can be adapted to noisy conditions when the training of the recogniser is conducted using noisy training material. Since the noise conditions during the test phase of the recogniser are not known a priori, equal properties of the noises for training and testing hardly occur in reality. However, in case the recogniser is designed for a certain field of application as an in-car speech recogniser, the approximate noise conditions are known to a certain extent, for example, when using information about the current speed of the car. Therefore, the speech models can be trained using speech sequences corrupted by noise which has similar properties as the noise during testing.

In this article, the most promising approaches to increase recognition performance in noisy surroundings are implemented in an isolated digit and spelling recognition task. All denoising techniques applied in the experimental section, representing a selection of methods as simple and efficient as CMS, MVN, and HEQ but also more complex approaches like AFE, USS, and SLDM feature enhancement as well as novel noise robust model architecture such as HCRF or the AR-SLDS, are introduced in Sections 3 to 5. While it is impossible to take into account and implement all noise compensation techniques that were developed in recent years, the selection of methods in this work covers many of the different concepts that are thinkable for in-car, but also for babble and white noise scenarios with all their specific advantages and disadvantages. Since we aim to focus on in-car speech recognition, noises produced by four different cars and three different road surfaces and velocities have been recorded and superposed with the speech sequences to simulate the noise conditions during driving. However, the findings may be transferred for many similar stationary noise situations.

Section 2 briefly outlines possible approaches to enhance the noise robustness of speech recognisers. In Section 3, an explanation of the different speech signal preprocessing techniques applied in this article is given, while Section 4 focuses on the feature enhancement strategies we used. Section 5 describes the speech model architecture which are used as alternatives to Hidden Markov Models in some of the experiments of Section 6.

2. Concepts for Noise Robust Speech Recognition

Aiming to counter the performance degradation of speech recognition systems in noisy surroundings, a variety of different concepts have been developed in recent years. The common goal of all noise compensation strategies is to minimise 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 in order to enable a proper representation of speech even if the signal is corrupted by noise. This can be achieved either by using noisy training data [15] or by joint speech and noise modelling [14]. The other method is trying to determine the clean features from the noisy speech sequence while using clean training data [9, 16, 17]. For that purpose, it is necessary to extract noise robust features and to find appropriate means of signal or feature preprocessing for speech enhancement.

This section summarises selected methods for speech signal preprocessing, auditory modelling, feature enhancement, speech modelling, and model adaptation.

2.1. Speech Signal Preprocessing. Preprocessing techniques for speech enhancement aim to compensate the effects of noise before the signal or rather the feature-based speech representation is classified by the recogniser which has been trained on clean data [18–20].

A state-of-the-art speech signal preprocessing that is used as a baseline feature extraction algorithm for noisy speech recognition problems like the Aurora2 task [21] is the advanced front-end feature extraction introduced in [22]. It uses a two-step Wiener filtering technique before the features are extracted, whereas filtering is done in the time domain.

As shown in [23, 24], methods based on spectral subtraction like Unsupervised Spectral Subtraction [17] reach similar performance while requiring less computational cost than Wiener filtering. Like the two-step Wiener filtering method included in the AFE, Unsupervised Spectral Subtraction can be considered as speech signal preprocessing step; however, USS is carried out in the magnitude spectogram domain.

2.2. Auditory Modelling and Feature Extraction. The two major effects that noise has on speech representation are a distortion in the feature space and a loss of information caused by its random behaviour. This loss has to be considered as irreversible, whereas the distortion of the features can be compensated depending on the suitability of the speech representation in noisy environments [1, 4].

Widely used speech features for auditory modelling are cepstral coefficients obtained through Linear Predictive Coding (LPC). The principle is based on the assumption that the speech signal can be regarded as the output of an all-pole linear filter that simulates the human vocal tract. However, speech recognition systems which process the cepstrum calculated via LPC tend to have low performance in the presence of noise [2]. For enhanced noise robustness, the use of the Perceptual Linear Prediction analysis method is a popular approach to extract spectral patterns [6, 25]. The technique is based on a transformation of the speech spectrum to the auditory spectrum that considers multiple perceptual relationships prior to performing linear prediction analysis. Another well-known speech representation is the extraction of Mel-frequency cepstral coefficients which provide a basis for several speech signal analysis

applications [17, 26–28]. They are calculated from the logarithm of filterbank amplitudes using the Discrete Cosine Transform.

In [29], the TRAP-TANDEM features were introduced. They describe the likelihood of subword classes at a time instant by evaluating temporal trajectories of band-limited spectral densities in the vicinity of the regarded time instant. Thereby the TRAP refers to the way the linguistic information is obtained from speech, while TANDEM refers to the technique that converts the evidence of subword classes into features for HMM-based speech recognition systems. Unlike conventional feature extraction techniques, which consider time windows of about 25 milliseconds to derive spectral features, TRAP also includes relatively long time spans up to one second to extract information for the recogniser. The strategy is motivated by the finding that information about a phoneme spreads over about 300 milliseconds [30, 31]. Furthermore, this method is able to remove slow varying noise [32].

Another approach to suppress slow variations in the short-term spectrum is the RASTA-PLP concept [33, 34] that makes PLP features more robust to linear spectral distortions. The filtering of time trajectories of critical-band filter outputs enables the removal of constant spectral components caused by convolutive factors in the speech signal.

2.3. Feature Enhancement. Further attempts to reduce the mismatch between test and training conditions are Cepstral Mean Subtraction [7], Mean and Variance Normalisation [8], or the Vector Taylor Series approach [35] which is able to deal with the nonlinear effects of noise. Nonlinear distortions can also be compensated by Histogram Equalisation [9], a technique which is often used in digital image processing [36] to improve the contrast of pictures. In speech processing, HEQ is a powerful means of improving the temporal dynamics of feature vector components distorted by noise. A cepstrum-domain feature compensation algorithm aiming to decompose speech and noise had also been presented in [37].

Another preprocessing approach to enhance noisy MFCC features is proposed in [10]: here a Switching Linear Dynamic Model 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. This concept has been extended in [38] where time-dependencies among the discrete state variables of the SLDM are included. To improve the accuracy of the noise model for nonstationary noise sources, [39] employs a state model for the dynamics of noise.

An enhancement of speech features can also be attained by incremental online adaptation of the feature space as in the feature space maximum likelihood linear regression (FMLLR) approach outlined in [40]. There, an FMLLR transform is integrated into a stack decoder by collecting adaptation data during recognition in real time.

2.4. Architecture for Speech Modelling. The most popular model architecture to represent speech characteristics in automatic speech recognition is Hidden Markov Models [11]. Apart from optimising the principle of auditory modelling and the methods for speech enhancement, finding alternative model architecture that applies Dynamic Bayesian Network structures which differ from the statistic assumptions of HMM modelling is an active area of research and a promising approach to improve noise robustness [12, 14, 41].

Generative models like the Hidden Markov Model are restricted in a way that they assume that the speech feature observations are conditionally independent. This can be considered as drawback as the restriction ignores long-range dependencies between observations. On the contrary, the Conditional Random Fields (CRFs) introduced in [42] use an exponential distribution to model a sequence, given the observation sequence. In order to estimate the conditional probability of a class for an entire sequence, the Hidden Conditional Random Field [12] incorporates hidden state sequences.

Other model architecture like Long Short-Term Memory Recurrent Neural Networks [43] which, in contrast to conventional Recurrent Neural Networks, consider long-range dependencies between the observations was recently proven to be well suited for speech recognition [44]. Even static classifiers like Support Vector Machines have been successfully applied in isolated word recognition tasks [45], where a warping of the observation sequence is less essential than in continuous speech recognition.

An alternative to the feature-based HMM has been proposed in [13] where the raw speech signal is modelled in the time domain. In clean conditions, methods based on raw signal modelling like the Switching Autoregressive HMM [13] work well; however, the performance quickly degrades whenever the technique is used in noisy surroundings. To improve noise robustness, [14] extended the SAR-HMM to a Switching Linear Dynamical System (SLDS) which includes an explicit noise model by modelling the dynamics of both the raw speech signal and the noise.

2.5. Model Adaptation. Not only joint speech and noise modelling but also training with noisy data can incorporate information about potential signal distortion in the recognition process. Experiments as done in [46] prove that recognition results are highly dependent on how much the used training material reveals about the characteristics of possible background noise during a test phase. Depending on how similar the noise conditions for training and testing are, we can distinguish between low, medium, and highly matched conditions training. Multiconditions training refers to using training material with different noise types. In real world, applications matching the conditions of training and testing phase are only possible if information about the noise conditions in which the recogniser will be used is available, for example, during the design of an in-car speech recogniser as shown herein.

Apart from adapting models by using noisy training material, the research area of model adaptation also covers

widely used techniques such as maximum a posteriori (MAP) estimation [47], maximum likelihood linear regression (MLLR) [48], and minimum classification error linear regression (MCELR) [49].

3. Speech Signal Preprocessing

3.1. Advanced Front-End Feature Extraction. In the advanced front-end feature extraction (AFE) algorithm outlined in [22], noise reduction is performed before the cepstral features are calculated. The main steps of the algorithm can be seen in Figure 1. After noise reduction, the denoised waveforms are processed, and the cepstral features are calculated. Finally blind equalisation is applied to the features.

The preprocessing algorithm for noise reduction is based on a two-stage Wiener filtering concept. The denoised output signal of the first stage enters a second stage where an additional dynamic noise reduction is performed. In contrast to the first filtering stage, a gain factorisation unit is incorporated in the second stage to control the intensity of filtering dependent on the signal-to-noise ratio (SNR) of the signal. The components of the two noise reduction cycles are illustrated in Figure 2. First, the input signal is divided into frames. After estimating the linear spectrum of each frame, the power spectral density (PSD) is smoothed along the time axis in the PSD Mean block. A voice activity detector (VAD) determines whether a frame contains speech or background noise, and so both the estimated spectrum of the speech frames and the estimated noise spectrum are used to calculate the frequency domain Wiener filter coefficients. To get a Mel-warped frequency domain Wiener filter, the linear Wiener filter coefficients are smoothed along the frequency axis using a Mel-filterbank. The Mel-warped Inverse Discrete Cosine Transform (Mel IDCT) unit calculates the impulse response of the Wiener filter before the input signal is filtered and passes through a second noise reduction cycle. Finally, the constant component of the filtered signal is removed in the "OFF" block.

Focusing on the Wiener filter approach as part of the advanced front-end feature extraction algorithm, a great advantage with respect to other preprocessing techniques for enhanced noise robustness is that noise reduction is performed on a frame-by-frame basis. The Wiener filter parameters can be adapted to the current SNR which makes the approach applicable to nonstationary noise. However, a critical issue of the AFE technique is that it relies on exact voice activity detection—a precondition that can be difficult to fulfil, especially if the SNR level is negative like in our in-car speech recognition problem (cf. Section 6.). Further, compared with other noise compensation strategies, the AFE is a rather complex mechanism and sensible to errors and inaccuracies within the individual estimation and transformation steps.

3.2. Unsupervised Spectral Subtraction. Another technique of speech enhancement known as Unsupervised Spectral Subtraction had been developed in [17]. This Spectral Subtraction scheme relies on a two-mixture model approach of noisy speech and aims to distinguish speech and background noise at the magnitude spectogram level.

3.2.1. Mixture Model. To derive a probabilistic model for speech distorted by noise, a probability distribution for both speech and noise is needed. When modelling background noise on silent parts of the time-frequency plane, it is common to assume white Gaussian behaviour for real and imaginary parts [50, 51]. In the magnitude domain, this corresponds to a Rayleigh probability density function $f_N(m)$ for noise:

$$f_N(m) = \frac{m}{\sigma_N^2} e^{-m^2/2\sigma_N^2} \tag{1}$$

Apart from the Rayleigh silence model, a speech model for "activity" that models large magnitudes only has to be derived to obtain the two-mixture model. For the speech probability density function $f_S(m)$, a threshold δ_S is defined with respect to the noise distribution $f_N(m)$, so that only magnitudes $m > \delta_S$ are modelled. In [17], a threshold $\delta_S = \sigma_N$ is used, whereas σ_N is the mode of the Rayleigh PDF. Consequently, we assume that magnitudes below σ_N are background noise. Two further constraints are necessary for $f_S(m)$.

(i) The derivative $f_S'(m)$ of the "activity" PDF may not be zero when m is just above δ_S; otherwise, the threshold δ_S has no meaning since it can be set to an arbitrarily low value.

(ii) As m goes towards infinity, the decay of $f_S(m)$ should be lower than the decay of the Rayleigh PDF to ensure that $f_S(m)$ models large amplitudes.

The "shifted Erlang" PDF with $h = 2$ [52] fulfils these two criteria and, therefore, can be used to model large amplitudes which are assumed to be speech:

$$f_S(m) = 1_{m>\sigma_N} \cdot \lambda_S^2 \cdot (m - \sigma_N) \cdot e^{-\lambda_S(m-\sigma_N)} \tag{2}$$

with $1_{m>\sigma_N} = 1$ if $m > \sigma_N$ and $1_{m>\sigma_N} = 0$, otherwise.

The overall probability density function for the spectral magnitudes of the noisy speech signal is given as follows:

$$f(m) = P_N \cdot f_N(m) + P_S \cdot f_S(m). \tag{3}$$

P_N is the prior for "silence" and background noise, respectively, whereas P_S is the prior for "activity" and speech, respectively. All the parameters of the derived PDF $f(m)$ summarised in the parameter set

$$\Lambda = \{P_N, \sigma_N, P_S, \lambda_S\} \tag{4}$$

are independent of time and frequency.

3.2.2. EM Training of Mixture Parameters. The parameters Λ of the two-mixture model can be trained using an Expectation Maximisation (EM) training algorithm [53].

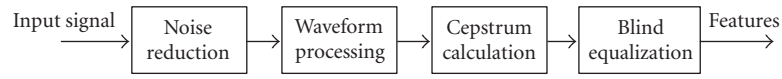

Figure 1: Feature extraction according to ETSI ES 202 050 V1.1.5.

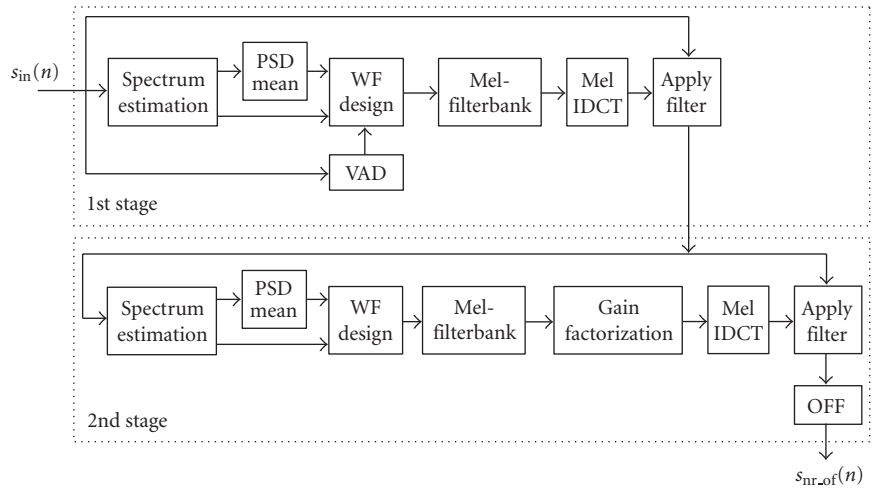

Figure 2: Two-stage Wiener filtering for noise reduction according to ETSI ES 202 050 V1.1.5.

In the "Expectation" step, the posteriors are estimated as follows:

$$p(\text{sil} \mid m_{f,t}, \Lambda) = \frac{P_N \cdot f_N(m_{f,t})}{P_N \cdot f_N(m_{f,t}) + P_S \cdot f_S(m_{f,t})}, \qquad p(\text{act} \mid m_{f,t}, \Lambda) = 1 - p(\text{sil} \mid m_{f,t}, \Lambda). \tag{5}$$

For the "Maximisation" step, the moment method is applied: all data is used to update σ_N before all data with values above the new σ_N is used to update λ_S. The method can be described by the following two update equations:

$$\hat{\sigma}_N = \frac{\left[\sum_{f,t} m_{f,t}^2 \cdot p(\text{sil} \mid m_{f,t}, \Lambda)\right]^{1/2}}{\left[2\sum_{f,t} p(\text{sil} \mid m_{f,t}, \Lambda)\right]^{1/2}}, \qquad \hat{\lambda}_S = \frac{\sum_{m_{f,t} > \hat{\sigma}_N} (m_{f,t} - \hat{\sigma}_N)^{-1} \cdot p(\text{act} \mid m_{f,t}, \Lambda)}{\sum_{m_{f,t} > \hat{\sigma}_N} p(\text{act} \mid m_{f,t}, \Lambda)}. \tag{6}$$

3.2.3. Spectral Subtraction. After the training of all mixture parameters $\Lambda = \{P_N, \sigma_N, P_S, \lambda_S\}$, Unsupervised Spectral Subtraction is applied using the parameter σ_N as floor value:

$$m_{f,t}^{\text{USS}} = \max\left(1, \frac{m_{f,t}}{\sigma_N}\right). \tag{7}$$

Flooring to a nonzero value is necessary whenever MFCC features are used, since zero magnitude values after spectral subtraction would lead to unfavourable dynamics in the cepstral coefficients.

Overall, USS is a simple and computationally efficient preprocessing strategy, allowing unsupervised EM fitting on observed data. A weakness of the approach is that it relies on appropriately estimating a speech magnitude PDF which is a difficult task. Since the PDFs do not depend on frequency and time, the applicability of USS is restricted to stationary noises. USS only models large magnitudes of speech so that low speech magnitudes cannot be distinguished from background noise.

4. Feature Enhancement

4.1. Feature Normalisation

4.1.1. Cepstral Mean Subtraction. A simple approach to remove the effects of noise and transmission channel transfer functions on the cepstral representation of speech is Cepstral Mean Subtraction [7, 54]. In many surroundings, for example, in a car where the speech signal is superposed by engine noise, the noise source can be considered as stationary, whereas the characteristics of the speech signal change relatively fast. Thus, a goal of preprocessing techniques for speech enhancement is to remove the stationary part of the input signal. As this quasi-non-varying part of the signal corresponds to a constant global shift in the cepstrum, speech can usually be enhanced by subtracting the long-term average cepstral vector

$$\mu = \frac{1}{T}\sum_{t=1}^{T} x_t \tag{8}$$

from the received distorted cepstrum vector sequence of length T:

$$X = \{x_1, x_2, \ldots, x_t, \ldots, x_T\}. \tag{9}$$

Consequently, we get a new estimate $\widetilde{x}_t$ of the signal in the cepstral domain:

$$\begin{aligned} \widetilde{x}_t &= x_t - \mu, \\ 1 &\le t \le T. \end{aligned} \tag{10}$$

This method also exploits the advantage of MFCC speech representation: if a transmission channel is inserted on the input speech, the speech spectrum is multiplied by the channel transfer function. In the logarithmic cepstral domain, this multiplication becomes an addition which can easily be removed by subtracting the cepstral mean from all input vectors. However, unlike techniques like Histogram Equalisation, CMS is not able to treat nonlinear effects of noise.

4.1.2. Mean and Variance Normalisation. Subtracting the mean of each feature vector component from the cepstral vectors (as done in CMS) corresponds to an equalisation of the first moment of the vector sequence probability distribution. In case noise also affects the variance of the speech features, a preprocessing stage for speech enhancement can profit also from normalising the variance of the vector sequence which corresponds to an equalisation of the first two moments of its probability distribution. This technique is known as Mean and Variance Normalisation and results in an estimated feature vector

$$\widetilde{x}_t = \frac{x_t - \mu}{\sigma}, \tag{11}$$

where the division by the vector σ, which contains the standard deviations of the feature vector components, is carried out elementwise. After MVN, all features have zero mean and unity variance.

4.1.3. Histogram Equalisation. Histogram Equalisation is a popular technique for digital image processing where it aims to increase the contrast of pictures. In speech processing, HEQ can be used to extend the principle of CMS and MVN to all moments of the probability distribution of the feature vector components [9, 55]. It enhances noise robustness by compensating nonlinear distortions in speech representation caused by noise and therefore reduces the mismatch between test and training data.

The main idea is to map the histogram of each component of the feature vector onto a reference histogram. The method is based on the assumption that the effect of noise can be described as a monotonic transformation of the features which can be reversed to a certain degree. As the effectiveness of HEQ is strongly dependent on the accuracy of the speech feature histograms, a sufficiently large number of speech frames have to be involved to estimate the histograms. An important difference between HEQ and other noise reduction techniques like Unsupervised Spectral Subtraction is that no analytic assumptions have to be made about the noise process. This makes HEQ effective for a wide range of different noise processes independent of how the speech signal is parameterised.

When applying HEQ, a transformation

$$\widetilde{x} = F(x) \tag{12}$$

has to be found in order to convert the probability density function $p(x)$ of a certain speech feature into a reference probability density function $\widetilde{p}(\widetilde{x}) = p_{\text{ref}}(\widetilde{x})$. If x is a unidimensional variable with probability density function $p(x)$, a transformation $\widetilde{x} = F(x)$ leads to a modification of the probability distribution, so that the new distribution of the obtained variable $\widetilde{x}$ can be expressed as

$$\widetilde{p}(\widetilde{x}) = p(G(\widetilde{x}))\frac{\partial G(\widetilde{x})}{\partial \widetilde{x}}, \tag{13}$$

with $G(\widetilde{x})$ being the inverse transformation of $F(x)$. To obtain the cumulative probabilities out of the probability density functions, we have to consider the following relationship:

$$\begin{aligned} C(x) &= \int_{-\infty}^{x} p(x')\,dx' \\ &= \int_{-\infty}^{F(x)} p(G(\widetilde{x}'))\frac{\partial G(\widetilde{x})}{\partial \widetilde{x}'}\,d\widetilde{x}' \\ &= \int_{-\infty}^{F(x)} \widetilde{p}(\widetilde{x}')\,d\widetilde{x}' \\ &= \widetilde{C}(F(x)). \end{aligned} \tag{14}$$

Consequently, the transformation converting the distribution $p(x)$ into the desired distribution $\widetilde{p}(\widetilde{x}) = p_{\text{ref}}(\widetilde{x})$ can be expressed as

$$\widetilde{x} = F(x) = \widetilde{C}^{-1}[C(x)] = C_{\text{ref}}^{-1}[C(x)], \tag{15}$$

where $C_{\text{ref}}^{-1}(\cdots)$ is the inverse cumulative probability function of the reference distribution, and $C(\cdots)$ is the cumulative probability function of the feature. To obtain the transformation for each feature vector component in our experiments, 500 uniform intervals between $\mu_i - 4\sigma_i$ and $\mu_i + 4\sigma_i$ were considered to derive the histograms, with μ_i and σ_i representing the mean and the standard deviation of the ith feature vector component. For each component, a Gaussian probability distribution with zero mean and unity variance was used as reference probability distribution.

Summing up the three feature normalisation strategies, CMS is the most simple and common technique which, however, cannot treat nonlinear effects of noise. MVN constitutes an improvement but still it only provides a linear transformation of the original variable. By contrast, HEQ compensates also nonlinear distortions. However, its effectiveness and accuracy heavily depend on the quality of the estimated feature histograms in a way that numerous speech frames are needed before HEQ can be expected to work well. Furthermore, Histogram Equalisation is intended to correct only monotonic transformations but the random behaviour of noise makes the actual transformation nonmonotonic which causes a loss of information.

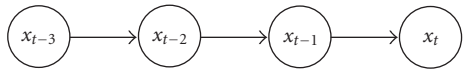

Figure 3: Linear dynamic model for noise.

4.2. Model-Based Feature Enhancement. Model-based speech enhancement techniques are based on modelling 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 [10], a Switching Linear Dynamic Model is used to capture the dynamics of clean speech. Similar to Hidden Markov Model-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.

4.2.1. Modelling of Noise. Unlike speech, which is modelled applying an SLDM, the modelling of noise is done by using a simple Linear Dynamic Model obeying the following system equation:

$$x_t = Ax_{t-1} + b + g_t. \tag{16}$$

Thereby the matrix A and the vector b simulate how the noise process evolves over time, and g_t represents a Gaussian noise source driving the system. A graphical representation of this LDM can be seen in Figure 3. As LDMs are time-invariant, they are suited to model signals like coloured stationary Gaussian noises as they occur in the interior of a car. Alternatively to the graphical model in Figure 3, the equations

$$\begin{aligned} p(x_t \mid x_{t-1}) &= \mathcal{N}(x_t; Ax_{t-1} + b, C), \\ p(x_{1:T}) &= p(x_1)\prod_{t=2}^{T} p(x_t \mid x_{t-1}) \end{aligned} \tag{17}$$

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.

4.2.2. Modelling of Speech. The modelling of speech is realised 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:

$$x_t = A(s_t)x_{t-1} + b(s_t) + g_t. \tag{18}$$

Consequently, every possible state sequence $s_{1:T}$ describes an LDM which is nonstationary 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 4, 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 [10] and the models used in [38] where time dependencies among the hidden state variables are included. A modification like this can be seen as analogous to extend a Gaussian Mixture Model (GMM) to an HMM. The SLDM corresponding to Figure 4 can be described as follows:

$$\begin{aligned} p(x_t, s_t \mid x_{t-1}) &= \mathcal{N}(x_t; A(s_t)x_{t-1} + b(s_t), C(s_t)) \cdot p(s_t), \\ p(x_{1:T}, s_{1:T}) &= p(x_1, s_1)\prod_{t=2}^{T} p(x_t, s_t \mid x_{t-1}). \end{aligned} \tag{19}$$

To train the parameters $A(s)$, $b(s)$, and $C(s)$ of the SLDM, conventional EM techniques are used. 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.

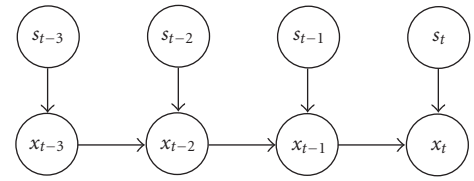

Figure 4: Switching linear dynamic model for speech.

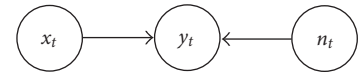

Figure 5: Observation model for noisy speech y_t.

4.2.3. 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 5 illustrates the graphical representation of the zero variance observation model with SNR inference introduced in [56]. Thereby it is assumed that speech x_t and noise n_t mix linearly in the time domain corresponding to a nonlinear mixing in the cepstral domain.

4.2.4. 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 generalised pseudo-Bayesian (GPB) algorithm [57]. 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.

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 denoised 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 $E[x_t \mid y_{1:t}]$.

Due to the noise modelling assumptions, SLDM feature enhancement has shown excellent performance also for coloured Gaussian noise even if the SNR level is negative. The linear dynamics of the speech model capture the smooth time evolution of human speech, while the switching states express the piecewise stationarity. The major limitation with respect to the noise type is that the model assumes the noise frames to be independent over time, so that only stationary noises are modelled accurately. Despite the GPB algorithm, SLDM feature enhancement is relatively time-consuming compared to simpler feature processing algorithms such as Histogram Equalisation. Another drawback is that the whole concept relies on precise voice activity detection in order to detect feature frames for the estimation of the noise LDM.

5. Model Architecture

5.1. Speech Modelling in the Feature Domain. To allow efficient speech modelling, it is common to model features extracted from the speech signal every 10 milliseconds instead of using the signal in the time domain as described in Section 5.2. As an alternative to conventional HMM modelling, the Hidden Conditional Random Field [58] will be introduced in the following and examined with respect to its noise robustness in Section 6.3.

5.1.1. Hidden Markov Models and Conditional Random Fields. Generative models like the Hidden Markov Model assume that the observations are conditionally independent, meaning that an observation is statistically independent of past observations provided that the values of the latent variables are known. Whenever there are long-range dependencies between the observations, like in human speech [30], this restriction can be too strict. Therefore, model architecture like the Conditional Random Field [42, 59, 60] makes use of an exponential distribution in order to model a sequence, given the observation sequence, and thereby drop the independence assumption between observations. Nonlocal dependencies between state and observation as well as unnormalised transition probabilities are allowed. As a Markov assumption can still be enforced, efficient inference techniques like dynamic programming can also be applied when using Conditional Random Fields. CRFs have been successfully applied in various tasks like information extraction [42] or language modelling [61].

5.1.2. Hidden Conditional Random Fields. As CRFs assign a label for each observation and each frame of a time-sequence, respectively, and, therefore, cannot directly estimate the probability of a class for an entire sequence, they need to be modified in order to be applicable for speech recognition tasks. Hence, the CRF has been extended to a Hidden Conditional Random Field which incorporates hidden state sequences [58]. The HCRF was successfully applied in various pattern recognition problems like Phone Classification [12], Gesture Recognition [62], Meeting Segmentation [63], or recognition of nonverbal vocalisations [64] where it partly outperformed HMM approaches. An advantage of HCRF is the ability to handle features that are allowed to be arbitrary functions of the observations while not requiring a more complicated training.

Similar to an HMM, the HCRF is used to model the conditional probability of a class label w representing a word, given the sequence of observations $X = x_1, x_2, \ldots, x_T$. With λ denoting the parameter vector and f being the so-called vector of sufficient statistics, the conditional probability is

$$p(w \mid X, \lambda) = \frac{1}{z(X,\lambda)} \sum_{\text{Seq} \in w} e^{\lambda \cdot f(w, \text{Seq}, X)}. \tag{20}$$

$\text{Seq} = s_1, s_2, \ldots, s_T$ represents the hidden state sequence that is run through while the conditional probability is calculated. The normalisation of the probability is realised by the function $z(X, \lambda)$ which is

$$z(X,\lambda) = \sum_{w} \sum_{\text{Seq} \in w} e^{\lambda \cdot f(w, \text{Seq}, X)}. \tag{21}$$

The vector f determines which probability to model, whereas f can be chosen in a way that the HCRF imitates a left-right HMM as shown in [12]. We restrict the HCRF to be a Markov chain; however the transition probabilities do not have to sum to one and the observations do not need to be real probability densities.

Like an HMM, an HCRF can be parameterised by transition scores a_{is} and observation scores $b_s(x_t)$:

$$\begin{aligned} a_{is} &\mathrel{\hat{=}} e^{\lambda_{is}^{(\text{Tr})}}, \\ b_s(x_t) &\mathrel{\hat{=}} e^{\lambda_s^{(\text{Occ})} + \lambda_s^{(M1)} x_t + \lambda_s^{(M2)} x_t^2}. \end{aligned} \tag{22}$$

The conditional probability can efficiently be computed when using forward and backward recursions as derived for the HMM. The forward probability is given as

$$\begin{aligned} \alpha_{s,t} &= \left[\sum_{i=1}^{S} \alpha_{i,t-1} a_{is} \right] b_s(x_t) \\ &= \left[\sum_{i=1}^{S} \alpha_{i,t-1} e^{\lambda_{is}^{(\text{Tr})}} \right] e^{\lambda_i^{(\text{Occ})} + \lambda_i^{(M1)} x_t + \lambda_i^{(M2)} x_t^2}, \end{aligned} \tag{23}$$

where S is the number of hidden states. The backward probabilities $\beta_i(t)$ can be obtained by using the recursion

$$\begin{aligned} \beta_{i,t} &= \sum_{s=1}^{S} a_{is} b_s(x_{t+1}) \beta_{s,t+1} \\ &= \sum_{s=1}^{S} e^{\lambda_{is}^{(\text{Tr})}} e^{\lambda_i^{(\text{Occ})} + \lambda_i^{(M1)} x_{t+1} + \lambda_i^{(M2)} x_{t+1}^2} \beta_{s,t+1}. \end{aligned} \tag{24}$$

Given the forward probabilities $\alpha_s(t)$, the probability $p(X \mid w, \lambda)$ that the model with parameters λ representing the word w produces observation X can be written as

$$p(X \mid w, \lambda) = \sum_{s=1}^{S} \alpha_{s,T}. \tag{25}$$

The conditional probability of a class label w given the observation X is

$$p(w \mid X, \lambda) = \frac{\sum_{s=1}^{S} \alpha_{s,T}}{\sum_{w} \sum_{s=1}^{S} \alpha_{s,T}}. \tag{26}$$

This HCRF definition makes it possible to use dynamic programming methods for decoding as with HMM. As shown in [12], a conditional probability density as for an HMM with transition probabilities a_{is}, emission means, and covariances μ_s and σ_s, respectively, can be obtained by setting the parameters λ as follows:

$$\lambda_{is}^{(\mathrm{Tr})} = \log a_{is}, \tag{27}$$

$$\lambda_{s}^{(\mathrm{Occ})} = -\frac{1}{2}\left\{ \log\left[(2\pi)^{D} \prod_{d=1}^{D} \sigma_{s,d}^{2} \right] + \sum_{d=1}^{D} \frac{\mu_{s,d}^{2}}{\sigma_{s,d}^{2}} \right\}, \tag{28}$$

$$\lambda_{s,d}^{(M1)} = \frac{\mu_{s,d}}{\sigma_{s,d}^{2}}, \tag{29}$$

$$\lambda_{s,d}^{(M2)} = -\frac{1}{2}\frac{1}{\sigma_{s,d}^{2}}. \tag{30}$$

Thereby d denotes the dimension of the D-dimensional observation, whereas i and s are states of the model. For the sake of simplicity, (27) to (30) consider only one mixture component. The extension to additional mixtures is straightforward.

5.2. Speech Modelling in the Time Domain. An alternative to conventional HMM modelling of speech is the modelling of the raw signal directly in the time domain. As proven in [13], modelling the raw signal can be a reasonable alternative to feature-based approaches. Such architecture offers the advantage that including an explicit noise model is straightforward, as can be seen in Section 5.2.2.

5.2.1. Switching Autoregressive Hidden Markov Models. In [14], a Switching Autoregressive HMM is applied for isolated digit recognition. The SAR-HMM is based on modelling the speech signal as an autoregressive (AR) process, whereas the nonstationarity 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. 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 4.2. The current state only depends on the preceding state with transition probability $p(s_t \mid s_{t-1})$. 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:

$$v_t = -\sum_{r=1}^{R} c_r(s_t) v_{t-r} + \eta(s_t) \tag{31}$$

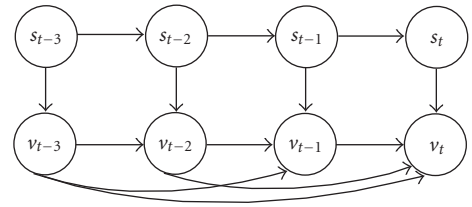

Figure 6: Dynamic bayesian network structure of the SAR-HMM.

with

$$\eta \sim \mathcal{N}(\eta; 0, \sigma^2(s_t)). \tag{32}$$

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

$$p(s_{1:T}, v_{1:T}) = p(v_1 \mid s_1)\, p(s_1) \prod_{t=2}^{T} p(v_t \mid v_{t-R:t-1}, s_t)\, p(s_t \mid s_{t-1}), \tag{33}$$

corresponding to the Dynamic Bayesian Network (DBN) structure illustrated in Figure 6.

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 realised by applying the EM algorithm. To infer the distributions $p(s_t \mid v_{1:T})$, a technique based on the forward-backward algorithm is used. Due to the fact that an observation v_t depends on R preceding observations (see Figure 6), 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 [65] is applied which means that the backward pass computes the posterior $p(s_t \mid v_{1:T})$ by "correcting" the output of the forward pass.

5.2.2. 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 [14]. 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 as follows:

$$h_t = A(s_t) h_{t-1} + \eta_t^{\mathcal{H}} \tag{34}$$

with

$$\eta_t^{\mathcal{H}} \sim \mathcal{N}(\eta_t^{\mathcal{H}}; 0, \Sigma_{\mathcal{H}}(s_t)). \tag{35}$$

The dynamics of the hidden variable are defined by the transition matrix $A(s_t)$ which depends on the current state s_t.

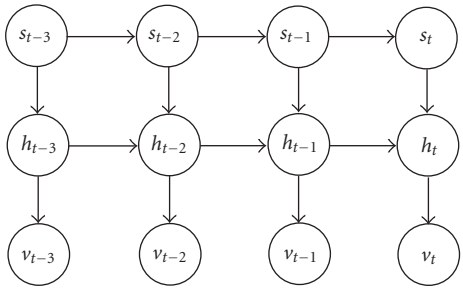

Figure 7: Dynamic bayesian network structure of the AR-SLDS.

Variations from pure linear state dynamics are modelled by the Gaussian distributed hidden "innovation" variable $\eta_t^{\mathcal{H}}$. Similar to the variable η_t used in (31) 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:

$$v_t = Bh_t + \eta_t^{\mathcal{V}} \tag{36}$$

with

$$\eta_t^{\mathcal{V}} \sim \mathcal{N}(\eta_t^{\mathcal{V}}; 0, \sigma_{\mathcal{V}}^2). \tag{37}$$

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 7 visualises the structure of the SLDS modelling 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 5.2.1 for the case $\sigma_{\mathcal{V}} = 0$ (see [14]). 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 [66] is applied to reduce the complexity. In contrast to the exact inference which requires $\mathcal{O}(S^T)$, the passes performed by the Expectation Correction algorithm are linear in T.

While the SAR-HMM has shown rather poor performance in noisy conditions, the AR-SLDS achieves excellent recognition rates for speech disturbed by white noise, as the variable $\eta_t^{\mathcal{V}}$ incorporates an additive white Gaussian noise (AWGN) model. In clean conditions, however, the performance of HMM speech modelling in the feature domain cannot be reached by the AR-SLDS, since time domain modelling is not as close to the principle of human perception as the well-established MFCC features. Also for coloured noise, the AR-SLDS cannot compete with feature domain approaches such as the SLDM. Further, computational complexity is still very high for the AR-SLDS. The Expectation Correction algorithm can reduce complexity from $\mathcal{O}(S^T)$ to $\mathcal{O}(T)$; however, for a speech utterance sampled at 16 kHz, T is 160 times higher than for a feature vector sequence extracted every 10 milliseconds.

6. Experiments

In order to compare the different speech signal preprocessing, feature enhancement, and speech modelling techniques introduced in Sections 3 to 5 with respect to their recognition performance in various noise scenarios, we implemented all of the techniques in a noisy speech recognition experiment which will be outlined in the following.

6.1. Speech Database. The digits "zero" to "nine" as well as the letters "A" to "Z" from the TI 46 Speaker Dependent Isolated Word Corpus [67] are used as speech database for the noisy digit and spelling recognition task. 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 [14], 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 digit training corpus consists of 800 utterances, while the digit test set contains 1280 samples. The same holds for the spelling database, consisting of 2080 utterances for training and 3328 for testing.

6.2. Noise Database. Even though we also considered babble and white noise scenarios, the main focus of this work lies on designing a robust speech recogniser for an in-car environment. Thus, great emphasis has been laid on simulating a wide spectrum of different noise conditions that can occur in the interior of a car. In general, interior noise can be split up into four rough groups. The first one is wind noise which is generated by air turbulence at the corners and edges of the vehicle and arises equivalently to the velocity. Another noise type is engine noise depending on load and number of revolutions. The third noise group is caused by wheels, driving, and suspension and is influenced by road surface and wheel type. Thus a rough surface causes more wheel and suspension noise than a smooth one. Finally, buzz, squeak, and rattles generated by pounding or relative movement of interior components of a vehicle have to be considered [68].

According to existing in-car speech recognition systems, the microphone would be mounted in the middle of the instrument panel. Consequently, all masking noises occurring in the interior of a car have been recorded exactly at the same point. Figure 8 illustrates the different noise sources. Note that the mouth-to-microphone transfer function had been neglected during the experiments in Section 6.3, since the masking effect of background noise was proven to be much higher than the effect of convolutional noise. In an additional experiment, the slight degradation of recognition performance in case of a convolution of the speech signal with a recorded in-car impulse response could be perfectly compensated by simple Cepstral Mean Subtraction.

As interior noise masking varies depending on vehicle class and derivates [68], speech is superposed by noise of four different vehicles as they are listed in Table 1.

Thus, a wide spectrum of car variations can be covered. Not only the vehicle type but also the road surface influences

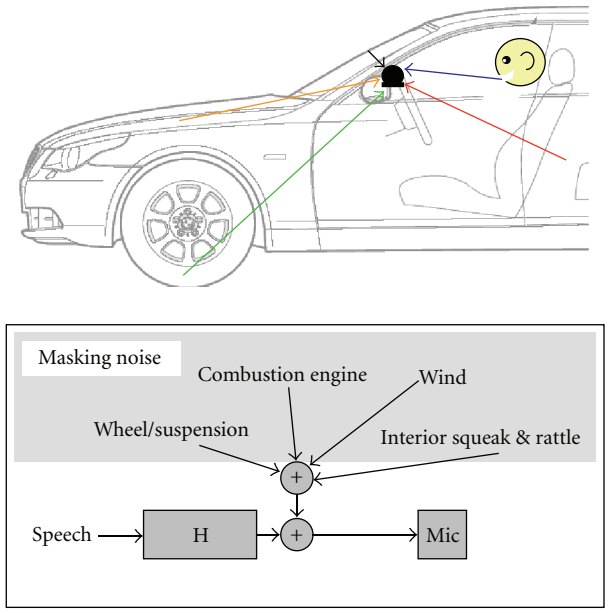

Figure 8: In-car speech and masking sound (top) and information flow (bottom).

Table 1: Considered vehicles.

Vehicle	Derivative	Class
BMW 5 series	Touring	Executive car
BMW 6 series	Convertible	Executive car
BMW M5	Sedan	Exec. sports car
MINI Cooper	Convertible	Super-mini

Table 2: Considered road surfaces and velocities.

Surface	Velocity	Abbreviation
Big cobbles	30 km/h	COB
Smooth city road	50 km/h	CTY
Highway	120 km/h	HWY

the characteristics of interior noise. Hence, three different surfaces in combination with typical velocities have been considered as shown in Table 2. The lowest excitation provides a driving over a smooth city road at 50 km/h and medium revolution (CTY). Thus, at this profile noise caused by wind, engine, wheels, and so forth has its minimum. The subsequent higher excitation is measured for a highway drive at 120 km/h (HWY). In that case, wind noise is a multiple higher than for a drive at 50 km/h. The worst and loudest sound in the interior of a car provokes a road with big cobbles (COB). At 30 km/h, wind noise can be neglected but the rough cobble surface involves dominant wheel and suspension noise. Figure 9 shows the SNR histograms of the noisy speech utterances for all four car types at each driving condition.

In spite of SNR levels below 0 dB, speech in the noisy test sequences is 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 (see Figure 10). Consequently, there is no full overlap of the spectrum of speech and noise. The extremely low SNR levels for the car noises (see Figure 9) are mainly caused by intense spectral components below the spectrum of human speech (motor drone). Filtering out those spectral components did not significantly affect recognition performance. Note that no A-weighting had been applied to estimate the SNR levels.

Apart from car noises (CAR), two further noise types are used in our experiments: first, a mixture of babble and street noise (BAB) at SNR levels 12 dB, 6 dB, and 0 dB, recorded in downtown Munich. This noise type is relevant for in-car speech recognition performance when driving with in an urban area with open windows. Furthermore, additive white Gaussian noise (WGN) has been used (SNR levels 20 dB, 10 dB, and 0 dB).

Note that heating, ventilating, and air conditioning (HVAC) noise was not examined as further potential noise source that can occur inside a car, since fan and defrost facilities were turned off during noise recording. Although it is quite evident that such additional in-car noises can further degrade speech recognition performance, we abstained from varying fan and defrost settings as those noise types can be characterised as stationary and are likely to not change the ranking of the individual noise compensation strategies but rather result in a negative "performance offset."

Contrariwise, the Lombard effect, which causes humans to speak louder when background noise is present, was also not considered since this would mostly result in a constant shift of the SNR histogram (Figure 9) towards higher SNR levels, without affecting conclusions about the effectiveness of the different denoising strategies.

6.3. Results. For every digit, a model was trained to build an isolated word recogniser. In the case of HMM and HCRF, each model consists of eight states with a mixture of three Gaussians per state. Thereby, clean utterances were used for training. 13 Mel-frequency cepstral coefficients as well as their first- and second-order derivatives were extracted. In addition, the usage of PLP features instead of MFCC was evaluated. Attempting to remove the effects of noise, various speech enhancement strategies as outlined in Section 4. were applied: Cepstral Mean Subtraction, Mean and Variance Normalisation, Histogram Equalisation, Unsupervised Spectral Subtraction, and Advanced Front-End feature extraction. In most of the experiments, the recognition rate for clean speech was around 99.9%. All parameters were tuned to achieve the best possible recognition performance.

As can be seen in Table 3, for stationary lowpass noise like the "CAR" and "BAB" noise types, the best average recognition rate can be achieved when enhancing the speech features using a global Switching Linear Dynamic Model for speech and a Linear Dynamic Model for noise (see Section 4.2). Thereby, all available clean training sequences were used to train the global SLDM which captures the dynamics of clean speech. The speech model consisted of 32 hidden states. The utterance-specific noise model consisted of a single Gaussian mixture component and was trained on the first and last 10 frames of the noisy test

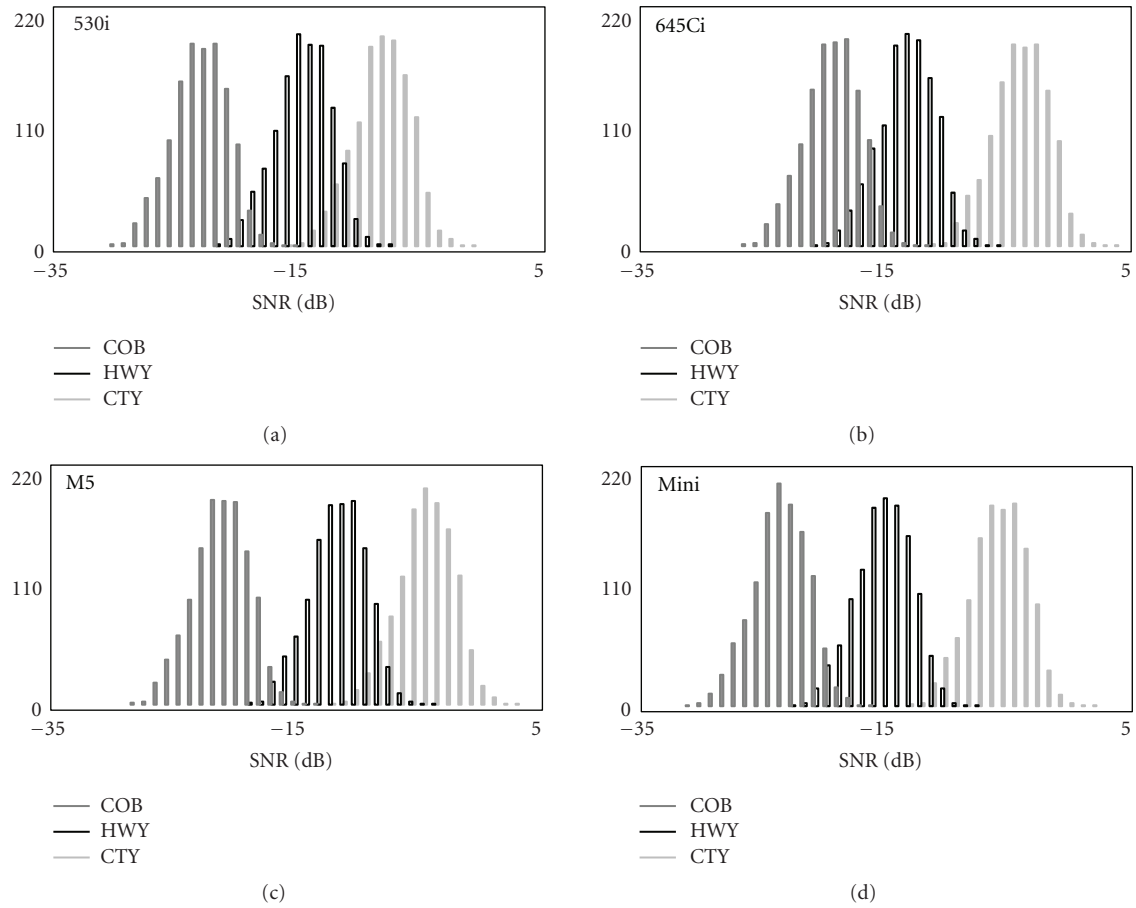

Figure 9: SNR level histograms for noisy speech utterances.

utterance. To speed up the calculation, the algorithm for speech enhancement was run with history parameter $r = 1$ (see Section 4.2.4). Also for more demanding recognition tasks like the Interspeech Consonant Challenge [69], SLDM feature enhancement was proven to increase recognition rates for noisy speech. The technique cannot compete with strategies using perfect knowledge of the local SNR of time-frequency components in the spectrogram like oracle masks [70–72]; however, compared to the Consonant Challenge HMM baseline recogniser [69], the SLDM approach can improve noisy speech recognition rates by up to 174% [73].

Applying Hidden Conditional Random Fields instead of HMM for the classification of features enhanced by CMS did not result in a better recognition rate.

For speech disturbed by white noise, the best recognition rate (93.3%, averaged over the different SNR conditions) is reached by the autoregressive Switching Linear Dynamical System explained in Section 5.2.2, where the noisy speech signal is modelled in the time domain as an autoregressive process. As explained in Section 5.2.2, the AR-SLDS constitutes the fusion of the SAR-HMM with the SLDS. The AR-SLDS used in the experiment is based on a 10th order SAR-HMM with ten states. This concept is however not suited for lowpass noise at negative SNR levels: for the "CAR" noise type a poor recognition rate of 47.2%, averaged over all car types and driving conditions, was obtained for AR-SLDS modelling. A reason for this is the assumption in (36) which expects additive noise to have a flat spectrum.

Table 3: Mean-isolated digit recognition rates in (%) for different noise types, noise compensation strategies, and features (training on clean data), sorted by mean recognition rate.

$\text{Strategy}_{\text{feat.}}$	clean	CAR	BAB	WGN
$\text{SLDM}_{\text{MFCC}}$	99.92	**99.52**	**99.29**	87.79
HEQ_{MFCC}	99.92	98.21	96.53	77.50
CMS_{PLP}	99.84	97.70	97.92	72.67
MVN_{MFCC}	99.84	94.86	93.32	79.06
CMS_{MFCC}	99.84	96.96	97.18	72.22
HEQ_{PLP}	99.92	97.20	95.27	66.51
$\text{HCRF/CMS}_{\text{MFCC}}$	99.76	95.67	94.97	70.06
USS_{MFCC}	99.05	93.52	92.27	53.19
AFE_{MFCC}	100.0	87.85	92.84	64.14
none_{PLP}	99.92	81.06	90.58	67.72
$\text{none}_{\text{MFCC}}$	99.92	75.09	88.37	63.67
$\text{AR} - \text{SLDS}_{\text{none}}$	97.37	47.24	78.51	**93.32**
$\text{SAR} - \text{HMM}_{\text{none}}$	98.10	54.26	83.16	41.91

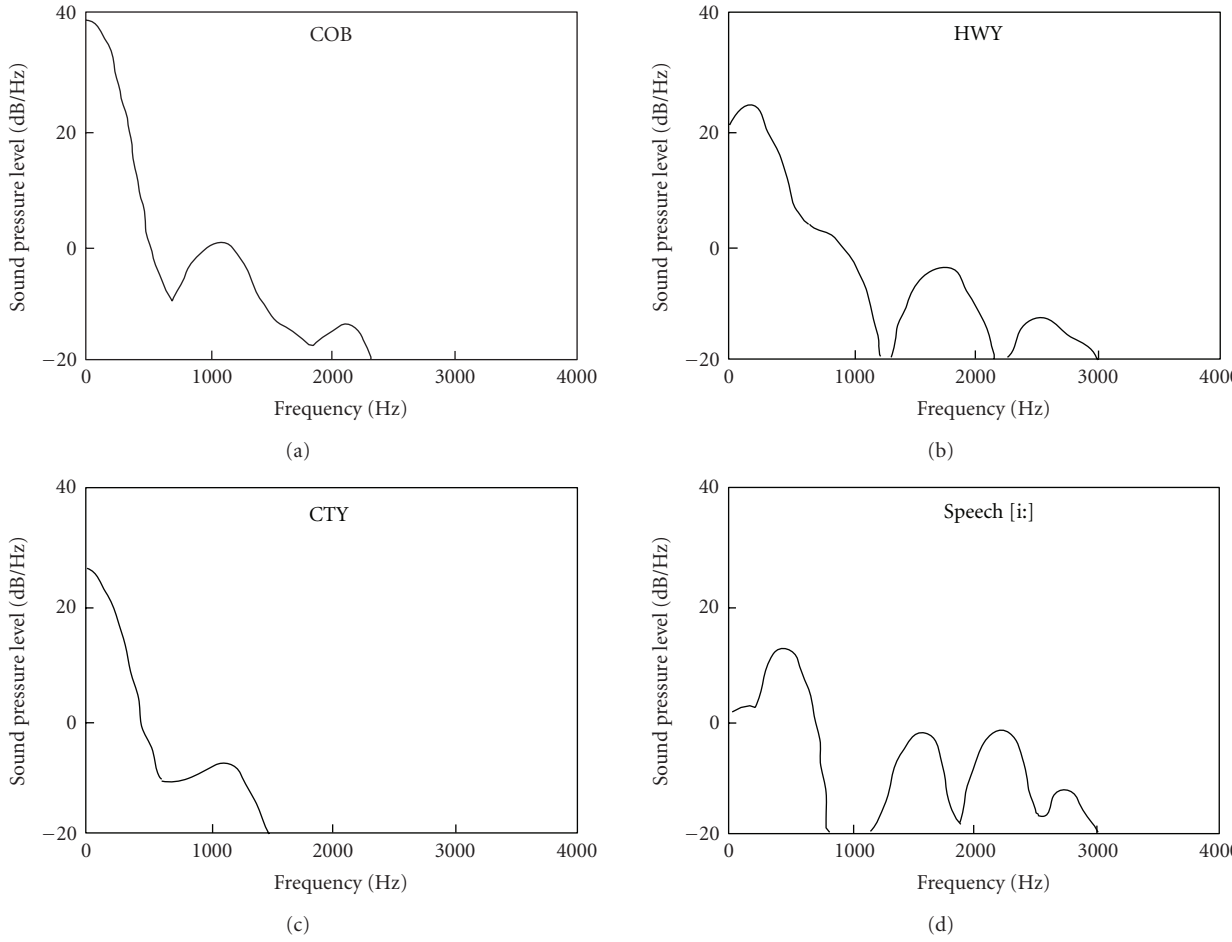

Figure 10: Long-term spectrum of the car noises COB, HWY, CTY (Mini Cooper S) and the spectral characteristics of the vowel [i:] spoken by a male speaker.

In case an HMM recogniser without feature enhancement is applied, PLP features perform slightly better than MFCC.

For white Gaussian noise, Table 4 compares the recognition rates obtained in this work with the performance reported in [14], using Unsupervised Spectral Subtraction, SAR-HMM and AR-SLDS modelling. Note that we used only 10 digits in our experiment ("zero" to "nine"), while [14] used 11 digits (including "oh"), which, together with extensive parameter tuning, should be the major reason why our SAR-HMM and AR-SLDS performance is better.

Table 5 summaries the mean recognition rates of an HMM recogniser without feature enhancement for three different training strategies: training on clean data, mismatched conditions training, and matched conditions training. Here, mismatched conditions training denotes the case when training and testing is done using speech sequences disturbed by the same noise type but at unequal noise conditions (SNR levels and driving conditions, resp.). Matched conditions training means training and testing with exactly identical noise types and noise conditions. Whenever the test sequence is disturbed by noise, mismatched conditions training outperforms a recogniser that had been trained on clean data. However, the main drawback of this approach is that for clean test sequences the mismatched conditions training strategy significantly downgrades recognition rates since in this case the noise pattern that had been learned during the training is missing when testing the recogniser. The results for matched conditions training serve as an upper benchmark for noisy speech recognition performance, as this strategy assumes perfect knowledge of the noise properties. Note that since in the matched conditions experiment one

Table 4: Isolated digit recognition rates in (%) for different SNR levels (white Gaussian noise) and noise compensation strategies (training on clean data); comparison between the results obtained in this work and the results reported in [14].

$\text{Strategy}_{\text{feat.}}$	clean	20 dB	10 dB	0 dB
USS_{MFCC}	99.1	96.1	53.5	9.9
USS_{MFCC} [14]	100.0	86.4	59.1	9.1
$\text{AR} - \text{SLDS}_{\text{none}}$	97.4	97.4	94.1	88.5
$\text{AR} - \text{SLDS}_{\text{none}}$ [14]	96.8	94.8	84.0	61.2
$\text{SAR} - \text{HMM}_{\text{none}}$	98.1	66.2	35.4	24.2
$\text{SAR} - \text{HMM}_{\text{none}}$ [14]	97.0	22.2	9.7	9.1

Table 5: Mean isolated digit recognition rates in (%) of an HMM recogniser without feature enhancement for different noise types and training strategies: matched conditions (MC) training, mismatched conditions (MMC) training, and training with clean data.

Training	clean	CAR	BAB	WGN
Clean data	99.92	75.09	88.37	63.67
MMC	79.42	96.86	98.74	68.51
MC	99.92	99.69	99.73	99.22

Table 6: Mean spelling recognition rates in (%) for different noise types and noise compensation strategies (training on clean data).

$\text{Strategy}_{\text{feat.}}$	clean	CAR	BAB	WGN
$\text{SLDM}_{\text{MFCC}}$	92.73	82.98	81.59	64.23
HEQ_{MFCC}	91.85	70.19	69.40	48.20
CMS_{MFCC}	93.09	73.79	69.78	47.06
$\text{none}_{\text{MFCC}}$	91.04	58.82	66.92	44.30

model was trained for every noise condition, this not only implies knowledge of the noise characteristics (e.g., by considering GPS or velocity information) but also higher memory requirements, as more than one model has to be stored. In the in-car scenario, this would entail one model for every driving condition, resulting in an increase of model size by factor four.

The best MFCC feature enhancement methods were also applied in the spelling recognition task (see Table 6). Again, for noisy test data, SLDM performs better than conventional techniques like HEQ.

7. Conclusion

In this article, a wide range of different techniques to improve the performance of automatic speech recognition in noisy surroundings has been implemented and evaluated in a noisy in-car isolated digit and spelling recognition task. In contrast to previous researches, diverse cars and driving conditions resulting in different spectral noise characteristics have been taken into account in order to obtain reliable conclusions about the universality of recognition performance. Thereby, four major approaches, affecting feature extraction, feature enhancement, speech decoding, and speech modelling, have been considered.

Aiming to approximate the speech recognition performance of human perception in noisy conditions, the use of PLP features as speech representation leads to a relative error reduction of 18.6% (averaged over all evaluated noise conditions) with respect to conventional MFCC. Furthermore, we proved that feature enhancement methods based on spectral subtraction and normalisation like Cepstral Mean Subtraction, Mean and Variance Normalisation, Unsupervised Spectral Subtraction, or Histogram Equalisation are able to partly remove the effects of stationary coloured noises as they occur in the interior of a car.

As a further approach to enhance speech features, a global Switching Linear Dynamic Model was used to capture the dynamics of speech enabling a model-based speech enhancement through joint speech and noise modelling. This technique prevailed for all car noise types and reached the best mean recognition rate of 96.9% for the noisy isolated digit recognition task.

The usage of Hidden Conditional Random Fields as an alternative model architecture could not outperform the conventional HMM. However, embedding a Switching Linear Dynamical System into a Switching Autoregressive HMM, and thereby modelling the raw signal in the time domain, leads to the best recognition performance for speech corrupted with additive white Gaussian noise.

Adapting the speech models by using noisy training data to build the models could also improve noise robustness. While matched conditions training is hardly possible in real life applications since the exact noise condition is not known a priori, mismatched conditions training, which uses training sequences disturbed by a noise type different from that in the test phase, outperformed training on clean data with a relative error reduction of 54.5%.

Apart from recognition performance, also computational complexity and possible fields of application have to be considered when designing a robust speech recogniser. While AFE and USS are more complex than feature normalisation techniques such as CMS or MVN, they are still suited for real-time applications. HEQ and SLDM feature enhancements achieve better recognition rates but require more computational resources. Modelling the speech signal in the time domain as done in the AR-SLDS experiment requires the most computational power and is therefore not suited for most real-life applications. For stationary noises, the SLDM is the most promising technique; however, it relies on accurate voice activity detection.

To optimise existing denoising strategies, future research effort could be spent on increasing the suitability of promising concepts like SLDM feature enhancement for the in-car speech recognition task by including discrete state transition probabilities or finding the optimum compromise between an increment of the history parameter and computational complexity. Furthermore, the AR-SLDS concept could be optimised for coloured noise to improve recognition performance when applying autoregressive speech modelling for in-car speech recognition. It might be also interesting how the implemented denoising methods perform in a continuous speech recognition task where, due to longer observation sequences, the parameters of a global SLDM as well as the cumulative histogram for the HEQ method could be estimated more precisely than in an isolated digit or spelling recognition experiment. Further improvements in noise robustness could also be achieved by combining different denoising concepts or by the application of other promising modelling concepts like Long Short-Term Memory Recurrent Neural Networks.

Speech recognition in noisy environments remains challenging; however, as shown in this article, spending effort on finding accurate techniques for auditory modelling, feature enhancement, speech modelling, and model adaption can remarkably reduce the performance gap between automatic speech recognition and human perception.

Acknowledgments

The authors would like to thank Jasha Droppo and Bertrand Mesot for providing SLDM and AR-SLDS binaries. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant agreement no. 211486 (SEMAINE).

References

[1] P. J. Moreno, *Speech recognition in noisy environments*, Ph.D. thesis, Carnegie Mellon University, Pittsburgh, Pa, USA, 1996.

[2] D.-S. Kim, S.-Y. Lee, and R. M. Kil, "Auditory processing of speech signals for robust speech recognition in real-world noisy environments," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 1, pp. 55–69, 1999.

[3] R. C. Rose, "Environmental robustness in automatic speech recognition," in *Proceedings of ISCA Workshop on Robustness in Conversational Interaction (Robust '04)*, Norwich, UK, August 2004.

[4] A. de la Torre, D. Fohr, and J. P. Haton, "Compensation of noise effects for robust speech recognition in car environments," in *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP '00)*, vol. 3, pp. 730–733, Beijing, China, October 2000.

[5] D. Langmann, A. Fischer, F. Wuppermann, R. Haeb-Umbach, and T. Eisele, "Acoustic front ends for speaker-independent digit recognition in car environments," in *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech '97)*, pp. 2571–2574, Rhodes, Greece, September 1997.

[6] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[7] M. G. Rahim, B.-H. Juang, W. Chou, and E. Buhrke, "Signal conditioning techniques for robust speech recognition," *IEEE Signal Processing Letters*, vol. 3, no. 4, pp. 107–109, 1996.

[8] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, no. 1–3, pp. 133–147, 1998.

[9] A. de La Torre, A. M. Peinado, J. C. Segura, J. L. Pérez-Córdoba, M. C. Benítez, and A. J. Rubio, "Histogram equalization of speech representation for robust speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 355–366, 2005.

[10] J. Droppo and A. Acero, "Noise robust speech recognition with a switching linear dynamic model," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '04)*, vol. 1, pp. 953–956, Montreal, Canada, May 2004.

[11] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[12] A. Gunawardana, M. Mahajan, A. Acero, and J. C. Platt, "Hidden conditional random fields for phone classification," in *Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech '05)*, pp. 1117–1120, Lisbon, Portugal, September 2005.

[13] Y. Ephraim and W. J. J. Roberts, "Revisiting autoregressive hidden Markov modeling of speech signals," *IEEE Signal Processing Letters*, vol. 12, no. 2, pp. 166–169, 2005.

[14] B. Mesot and D. Barber, "Switching linear dynamical systems for noise robust speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1850–1858, 2007.

[15] A. Sankar, A. Stolcke, T. Chung, et al., "Noise-resistant feature extraction and model training for robust speech recognition," in *Proceedings of the DARPA of CSR Workshop*, pp. 117–122, Ardenhouse, NY, USA, February 1996.

[16] N. S. Kim, "Nonstationary environment compensation based on sequential estimation," *IEEE Signal Processing Letters*, vol. 5, no. 3, pp. 57–59, 1998.

[17] G. Lathoud, M. Magimai-Doss, B. Mesot, and H. Bourlard, "Unsupervised spectral subtraction for noise-robust ASR," in *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU '05)*, pp. 343–348, San Juan, Puerto Rico, USA, November 2005.

[18] L. Szymanski and M. Bouchard, "Comb filter decomposition for robust ASR," in *Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech '05)*, pp. 2645–2648, Lisbon, Portugal, September 2005.

[19] R. Rifkin, K. Schutte, M. Saad, J. Bouvrie, and J. Glass, "Noise robust phonetic classification with linear regularized least squares and second-order features," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '07)*, vol. 4, pp. 881–884, Honolulu, Hawaii, USA, April 2007.

[20] B. Raj, L. Turicchia, B. Schmidt-Nielsen, and R. Sarpeshkar, "An FFT-based companding front end for noise-robust automatic speech recognition," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2007, Article ID 65420, 13 pages, 2007.

[21] H. G. Hirsch and D. Pierce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proceedings of the International Workshop on Automatic Speech Recognition: Challenges for the Next Millennium (ISCA ITRW ASR '00)*, Paris, France, September 2000.

[22] ETSI ES 202 050 V1.1.5, "Speech processing, transmission and quality aspects (STQ), distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," 2007.

[23] G. Lathoud, M. M. Doss, and H. Boulard, "Channel normalization for unsupervised spectral subtraction," in *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU '05)*, San Juan, Puerto Rico, USA, November 2005.

[24] S. V. Vaseghi and B. P. Milner, "Noise compensation methods for hidden Markov model speech recognition in adverse environments," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 1, pp. 11–21, 1997.

[25] J.-C. Junqua, H. Wakita, and H. Hermansky, "Evaluation and optimization of perceptually-based ASR front-end," *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 1, pp. 39–48, 1993.

[26] L. Deng, J. Droppo, and A. Acero, "Recursive estimation of nonstationary noise using iterative stochastic approximation for robust speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 568–580, 2003.

[27] Q. Zhu and A. Alwan, "On the use of variable frame rate analysis in speech recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '00)*, vol. 3, pp. 1783–1786, Istanbul, Turkey, June 2000.

[28] B. Schuller, D. Seppi, A. Batliner, A. Maier, and S. Steidl, "Towards more reality in the recognition of emotional speech," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '07)*, vol. 4, pp. 941–944, Honolulu, Hawaii, USA, April 2007.

[29] H. Hermansky, "TRAP-TANDEM: data-driven extraction of temporal features from speech," in *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU '03)*, pp. 255–260, St. Thomas, Virgin Islands, USA, November-December 2003.

[30] J. A. Bilmes, "Maximum mutual information based reduction strategies for cross-correlation based joint distributional modeling," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '98)*, vol. 1, pp. 469–472, Seattler, Wash, USA, May 1998.

[31] H. H. Yang, S. van Vuuren, S. Sharma, and H. Hermansky, "Relevance of time-frequency features for phonetic and speaker-channel classification," *Speech Communication*, vol. 31, no. 1, pp. 35–50, 2000.

[32] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel, "On the relative importance of various components of the modulation spectrum for automatic speech recognition," *Speech Communication*, vol. 28, no. 1, pp. 43–55, 1999.

[33] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "RASTA-PLP speech analysis technique," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '92)*, vol. 1, pp. 121–124, San Francisco, Calif, USA, March 1992.

[34] B. E. D. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech Communication*, vol. 25, no. 1–3, pp. 117–132, 1998.

[35] P. J. Moreno, B. Raj, and R. M. Stern, "A vector Taylor series approach for environment-independent speech recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '96)*, vol. 2, pp. 733–736, Atlanta, Ga, USA, May 1996.

[36] J.-Y. Kim, L.-S. Kim, and S.-H. Hwang, "An advanced contrast enhancement using partially overlapped sub-block histogram equalization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 4, pp. 475–484, 2001.

[37] H. K. Kim and R. C. Rose, "Cepstrum-domain acoustic feature compensation based on decomposition of speech and noise for ASR in noisy environments," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 435–446, 2003.

[38] J. Deng, M. Bouchard, and T. H. Yeap, "Noisy speech feature estimation on the Aurora2 database using a switching linear dynamic model," *Journal of Multimedia*, vol. 2, no. 2, pp. 47–52, 2007.

[39] S. Windmann and R. Haeb-Umbach, "Modeling the dynamics of speech and noise for speech feature enhancement in ASR," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '08)*, pp. 4409–4412, Las Vegas, Nev, USA, April 2008.

[40] Y. Li, H. Erdogan, Y. Gao, and E. Marcheret, "Incremental online feature space MLLR adaptation for telephony speech recognition," in *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP '02)*, pp. 1417–1420, Denver, Colo, USA, September 2002.

[41] A. Ganapathiraju, J. E. Hamaker, and J. Picone, "Applications of support vector machines to speech recognition," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2348–2355, 2004.

[42] J. D. Lafferty, A. McCallum, and F. C. N. Pereiar, "Conditional random fields: probabilistic models for segmenting and labeling sequence data," in *Proceedings of the 18th International Conference on Machine Learning (ICML '01)*, pp. 282–289, Williamstown, Mass, USA, June-July 2001.

[43] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[44] S. Fernández, A. Graves, and J. Schmidhuber, "An application of recurrent neural networks to discriminative keyword spotting," in *Proceedings of the 17th International Conference on Artificial Neural Networks (ICANN '07)*, vol. 4669 of *Lecture Notes in Computer Science*, pp. 220–229, Porto, Portugal, September 2007.

[45] A. de Andrade Bresolin, A. D. D. Neto, and P. J. Alsina, "Digit recognition using wavelet and SVM in Brazilian Portuguese," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '08)*, pp. 1545–1548, Las Vegas, Nev, USA, April 2008.

[46] D. Macho, L. Mauuray, B. Noe, et al., "Evaluation of a noise-robust DSR front-end on Aurora database," in *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP '02)*, pp. 17–20, Denver, Colo, USA, September 2002.

[47] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[48] Z. Wang, T. Schultz, and A. Waibel, "Comparison of acoustic model adaptation techniques on non-native speech," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '03)*, vol. 1, pp. 540–543, Hong Kong, April 2003.

[49] X. He and W. Chou, "Minimum classification error linear regression for acoustic model adaptation of continuous density HMMS," in *Proceedings of IEEE International Conference on Multimedia & Expo (ICME '03)*, vol. 1, pp. 397–400, Baltimore, Md, USA, July 2003.

[50] R. Martin and C. Breithaupt, "Speech enhancement in the DFT domain using Laplacian speech priors," in *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC '03)*, pp. 87–90, Kyoto, Japan, September 2003.

[51] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[52] C. M. Grinstead and J. L. Snell, *Introduction to Probability*, American Mathematical Society, Providence, RI, USA, 1997.

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

[54] C. R. Jankowski Jr., H.-D. H. Vo, and R. P. Lippmann, "Comparison of signal processing front ends for automatic word recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 286–293, 1995.

[55] F. Hilger and H. Ney, "Quantile based histogram equalization for noise robust large vocabulary speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 3, pp. 845–854, 2006.

[56] J. Droppo, L. Deng, and A. Acero, "A comparison of three non-linear observation models for noisy speech features," in *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech '03)*, vol. 2, pp. 681–684, Geneva, Switzerland, September 2003.

[57] Y. Bar-Shalom and X. R. Li, *Estimation and Tracking: Principles, Techniques, and Software*, Artech House, Norwood, Mass, USA, 1993.

[58] A. Quattoni, M. Collins, and T. Darrell, "Conditional random fields for object recognition," in *Advances in Neural Information Processing Systems 17*, pp. 1097–1104, MIT Press, Cambridge, Mass, USA, 2005.

[59] F. Sha and F. Pereira, "Shallow parsing with conditional random fields," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, vol. 1, pp. 134–141, Edmonton, Canada, May-June 2003.

[60] D. Pinto, A. McCallum, X. Wei, and W. B. Croft, "Table extraction using conditional random fields," in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '03)*, pp. 235–242, Toronto, Canada, July-August 2003.

[61] B. Roark, M. Saraclar, M. Collins, and M. Johnson, "Discriminative language modeling with conditional random fields and the perceptron algorithm," in *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL '04)*, pp. 48–55, Barcelona, Spain, July 2004.

[62] S. B. Wang, A. Quattoni, L.-P. Morency, D. Demirdjian, and T. Darrell, "Hidden conditional random fields for gesture recognition," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06*, vol. 2, pp. 1521–1527, New York, NY, USA, June 2006.

[63] S. Reiter, B. Schuller, and G. Rigoll, "Hidden conditional random fields for meeting segmentation," in *Proceedings of IEEE International Conference on Multimedia & Expo (ICME '07)*, pp. 639–642, Beijing, China, July 2007.

[64] B. Schuller, F. Eyben, and G. Rigoll, "Static and dynamic modelling for the recognition of non-verbal vocalisations in conversational speech," in *Proceedings of the 4th IEEE Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems: Perception in Multimodal Dialogue Systems (PIT '08)*, pp. 99–110, Kloster Irsee, Germany, June 2008.

[65] H. E. Rauch, G. Tung, and C. T. Striebel, "Maximum likelihood estimates of linear dynamic systems," *AIAA Journal*, vol. 3, no. 8, pp. 1445–1450, 1965.

[66] D. Barber, "Expectation correction for smoothed inference in switching linear dynamical systems," *Journal of Machine Learning Research*, vol. 7, pp. 2515–2540, 2006.

[67] G. R. Doddington and T. B. Schalk, "Speech recognition: turning theory to practice," *IEEE Spectrum*, vol. 18, no. 9, pp. 26–32, 1981.

[68] M. Grimm, K. Kroschel, H. Harris, et al., "On the necessity and feasibility of detecting a driver's emotional state while driving," in *Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interaction (ACII '07)*, pp. 126–138, Lisbon, Portugal, September 2007.

[69] M. Cooke and O. Scharenborg, "The Interspeech 2008 consonant challenge," in *Proceedings of Interspeech*, pp. 1–4, Brisbane, Australia, September 2008.

[70] B. J. Borgström and A. Alwan, "HMM-based estimation of unreliable spectral components for noise robust speech recognition," in *Proceedings of Interspeech*, pp. 1769–1772, Brisbane, Australia, September 2008.

[71] P. Jancovic and K. Münevver, "On the mask modeling and feature representation in the missing-feature ASR: evaluation on the consonant challenge," in *Proceedings of Interspeech*, pp. 1777–1780, Brisbane, Australia, September 2008.

[72] J. F. Gemmeke and B. Cranen, "Noise reduction through compressed sensing," in *Proceedings of Interspeech*, pp. 1785–1788, Brisbane, Australia, September 2008.

[73] B. Schuller, M. Wöllmer, T. Moosmayr, and G. Rigoll, "Speech recognition in noisy environments using a switching linear dynamic model for feature enhancement," in *Proceedings of Interspeech*, Brisbane, Australia, September 2008.