interspeech 2008 incorporating SST 08

Speech Recognition in Noisy Environments using a Switching Linear Dynamic Model for Feature Enhancement

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

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

10.21437/Interspeech.2008-492

Abstract

The performance of automatic speech recognition systems strongly decreases whenever the speech signal is disturbed by background noise. We aim to improve noise robustness focusing on all major levels of speech recognition: feature extraction, feature enhancement, and speech modeling. Different auditory modeling concepts, speech enhancement techniques, training strategies, and model architectures are implemented in an in-car digit and spelling recognition task. We prove that joint speech and noise modeling with a global Switching Linear Dynamic Model (SLDM) capturing the dynamics of speech, and a Linear Dynamic Model (LDM) for noise, prevails over state-of-the-art speech enhancement techniques. Furthermore we show that the baseline recognizer of the Interspeech Consonant Challenge 2008 can be outperformed by SLDM feature enhancement for almost all of the noisy testsets.

Index Terms: ASR, SLDM, feature enhancement, Consonant Challenge

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.

Another technique for noise robust speech recognition is introduced in [8] 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. The AR-SLDS includes an explicit noise model by modeling the dynamics of both the raw speech signal in the time domain and the noise.

This paper examines a model based preprocessing approach to enhance noisy features as it is proposed in [9]. 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.

The paper is organized as follows: Section 2 outlines the SLDM used for feature enhancement in this work, while in Section 3 the concept is evaluated in an isolated digit and spelling recognition task. Section 4 introduces a the noisy speech database of the Interspeech Consonant Challenge 2008 [10] and compares the performance of the Consonant Challenge baseline recognizer and the recognizer using SLDM feature enhancement.

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 [9] 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 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

Accepted after peer review of full paper

(LDM) obeying the following system equation:

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

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.

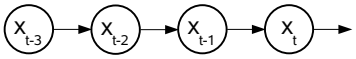

Figure 1: *Linear Dynamic Model for noise*

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 [9] and the models used in [11] 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:

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

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

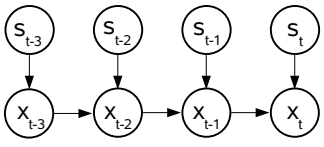

Figure 2: *Switching Linear Dynamic Model for speech*

To train the parameters $A(s)$, $b(s)$ and $C(s)$ of the SLDM conventional EM techniques are used [12]. 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 [13]. 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.

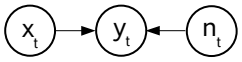

Figure 3: *Observation model for noisy speech y_t*

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 [14]. 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 [9].
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 the clean features are assumed to be the Minimum Mean Square Error (MMSE) estimate $E[x_t|y_{1:t}]$.

3. Isolated Digit and Spelling Recognition

3.1. Speech Database

The digits "zero" to "nine" as well as the letters "A" to "Z" from the TI 46 Speaker Dependent Isolated Word Corpus [15] 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 [8], 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.

3.2. Noise Database

In order to cover a wide spectrum of in-car noise conditions, speech is superposed by noise recorded in four different BMW vehicles (530i, 645Ci, M5, and Mini) at three different conditions: driving over a smooth city road at 50 km/h (CTY), driving over big cobbles at 30 km/h (COB), and driving on a highway at 120 km/h (HWY). The car noises are the same as in [16]. Table 1 shows the mean SNR levels for all four car types at each driving condition.

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

Table 1: *Mean SNR levels for noisy speech utterances*

Car Noise	SNR	Car noise	SNR
530i, CTY	-8 dB	645Ci, CTY	-3 dB
530i, HWY	-15 dB	645Ci, HWY	-13 dB
530i, COB	-23 dB	645Ci, COB	-19 dB
M5, CTY	-4 dB	Mini, CTY	-5 dB
M5, HWY	-11 dB	Mini, HWY	-15 dB
M5, COB	-21 dB	Mini, COB	-24 dB

spectrum of speech and noise.
Apart from car noises (CAR), two further noise types are used in this work: 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 (AWGN) has been used (SNR levels 20 dB, 10 dB, and 0 dB).

3.3. Experiments and Results

For every digit an HMM was trained, whereas each HMM consists of 8 states with a mixture of three Gaussians per state. 13 Mel-frequency cepstral coefficients (MFCC) as well as their first and second order derivatives were extracted. In addition the usage of Perceptual Linear Prediction (PLP) features [17] instead of MFCC was evaluated. Attempting to remove the effects of noise, various speech enhancement strategies were applied: Cepstral Mean Subtraction, Mean and Variance Normalization, Histogram Equalization, Unsupervised Spectral Subtraction, and Advanced Front-End Wiener Filtering [1]. However, as can be seen in Table 2, 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 [9] for speech and a Linear Dynamic Model for noise.

Table 2: *Mean isolated digit recognition rates for different noise types, noise compensation strategies, and features (training on clean data)*

$\textbf{Strategy}_{\textbf{feat.}}$	clean	CAR	BAB	AWGN
SLDM_{MFCC}	99.9%	99.5%	99.3%	87.8%
HEQ_{MFCC}	99.9%	98.2%	96.5%	77.5%
CMS_{PLP}	99.8%	97.7%	97.9%	72.7%
MVN_{MFCC}	99.8%	94.9%	93.3%	79.1%
CMS_{MFCC}	99.8%	97.0%	97.2%	72.2%
HEQ_{PLP}	99.9%	97.2%	95.3%	66.5%
USS_{MFCC}	99.1%	93.5%	92.3%	53.2%
AFE_{MFCC}	100.0%	87.9%	92.8%	64.1%
none_{PLP}	99.9%	81.1%	90.6%	67.7%
none_{MFCC}	99.9%	75.1%	88.4%	63.3%
$\mathrm{AR{-}SLDS}_{none}$	97.4%	47.2%	78.5%	93.3%

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 Model (AR-SLDS) introduced in [8], where the noisy speech signal is modeled in the time domain as an autoregressive process. 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 modeling.
In case an HMM recognizer without feature enhancement is applied, PLP features perform slightly better than MFCC.
Table 3 summarizes the mean recognition rates of an HMM recognizer without feature enhancement for three different training strategies: training on clean data, Mismatched Conditions Training, and Matched Conditions Training. 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 respectively). Matched Conditions Training means training and testing with exactly identical noise types and noise conditions.
The best MFCC feature enhancement methods were also applied in the spelling recognition task (see Table 4). Again, for noisy test data, SLDM perform better than conventional techniques like HEQ.

Table 3: *Mean isolated digit recognition rates of an HMM recognizer without feature enhancement for different noise types and training strategies: Matched Conditions (MC), Mismatched Conditions (MMC) and with clean data*

Training	clean	CAR	BAB	AWGN
clean data	99.9%	75.1%	88.4%	63.3%
MMC	79.4%	96.9%	98.7%	68.5%
MC	99.9%	99.7%	99.7%	99.2%

Table 4: *Mean spelling recognition rates for different noise types and noise compensation strategies (training on clean data)*

$\textbf{Strategy}_{\textbf{feat.}}$	clean	CAR	BAB	AWGN
SLDM_{MFCC}	92.7%	83.0%	81.6%	64.2%
HEQ_{MFCC}	91.8%	70.2%	69.4%	48.2%
CMS_{MFCC}	93.0%	73.8%	69.8%	47.1%
none_{MFCC}	91.0%	58.8%	66.6%	44.3%

4. Consonant Challenge

Since the SLDM speech modeling concept prevailed for the recognition tasks described in Section 3, it was also applied to the challenging task of consonant recognition which is outlined in [10] (Interspeech Consonant Challenge 2008). The database was designed to compare human and automatic speech recognition performance. Each speech utterance consists of a vowel-consonant-vowel (VCV) combination, whereas different stress conditions are used. The overall speech corpus contains 10368 tokens (24 speakers $\cdot$ 24 consonants $\cdot$ 2 stress types $\cdot$ 9 vowel contexts). The training material consists of all utterances spoken by 8 female and 8 male speakers whereas the samples of the remaining 4 female and 4 male speakers is used as testset.
In contrast to the noisy speech sequences used in Section 3 where the spectrum of speech and noise do not overlap completely, the VCV utterances are superposed by noises whose spectral characteristics are similar to the spectrum of

speech. Thereby the SNR levels of the noisy speech testsets vary between 0 dB and -6 dB.
To enhance the MFCC features of the noisy VCV utterances, a global SLDM which captures the dynamics of speech was trained using the whole clean Consonant Challenge training corpus. As for the digit and spelling recognition task, the SLDM consisted of 32 hidden states and the history parameter $r = 1$ was used. The Linear Dynamic Model for noise was derived separately for each noisy test utterance using the first and last 10 frames of the corresponding sequence.

Table 5: *Consonant recognition accuracies of the Consonant Challenge baseline recognizer with and without SLDM feature enhancement for clean (testset 1) and noisy data (testsets 2 to 6)*

testset	SNR	baseline	SLDM
1	∞	88.5%	78.9%
2	-6 dB	12.0%	10.4%
3	-2 dB	5.5%	12.2%
4	-2 dB	4.2%	11.5%
5	0 dB	4.2%	7.0%
6	-6 dB	7.6%	17.2%
7	-3 dB	7.8%	10.4%

Table 5 compares the consonant recognition accuracies of the baseline recognizer described in [10], using MFCC features and HMM with 3 states and 24 Gaussian mixtures, with a recognizer applying additional feature enhancement with a Switching Linear Dynamic Model as explained before. The HMM settings were the same as for the baseline recognizer. For the clean testset recognition accuracy decreases through the SLDM algorithm, however, when noise is added to the test utterances, SLDM feature enhancement leads to improved accuracies. The best improvement could be obtained for testset 6, where the recognition accuracy was increased by almost 10%. The only exception is testset 2 for which accuracy slightly decreased after feature enhancement. Nevertheless the absolute values of the accuracies for the baseline recognizer as well as for the recognizer with additional SLDM feature enhancement are at a low level as they are still too close to the probability of guessing (4.2%). A reason for this are the negative SNR levels and the similarity of the spectrum of speech and noise which hinders the separation of the speech signal from the noise source.

5. Conclusion

The digit and spelling recognition task in this work examines various auditory modeling, feature enhancement, speech modeling, and training strategies for a wide range of different noise types. Thereby speech enhancement with a Switching Linear Dynamic Model prevails for lowpass car noises, whereas autoregressive speech modeling using an AR-SLDS was proven to be the best technique for white noise. Mismatched Conditions Training is able to improve noisy speech recognition rates with respect to clean training. An upper border for the recognition performance is determined when using Matched Conditions Training, which assumes perfect knowledge of the noise properties.
The effect of applying a Switching Linear Dynamic Model as a technique of model based feature enhancement was also evaluated on the noisy consonant recognition task of the Interspeech Consonant Challenge 2008, where the SLDM concept was able to improve the recognition accuracies of 5 out of 6 noisy testsets with respect to the Consonant Challenge baseline recognizer.

6. Acknowledgements

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

7. 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., and Boulard, H., "Channel normalization for unsupervised spectral subtraction", Proceedings of ASRU (2005)

[3] Rahim, M. G., Juang, B. H., Chou, W., and Buhrke, E., "Signal conditioning techniques for robust speech recognition", IEEE Signal Processing Letters (1996) 107–109

[4] Lathoud, G., Magimia-Doss, M., Mesot, B., and Boulard, H., "Unsupervised spectral subtraction for noise-robust ASR", Proceedings of ASRU (2005) 189–194

[5] Viikki, O. and Laurila, K., "Cepstral domain segmental feature vector normalization for noise robust speech recognition", Speech Communication (1998) 133–147

[6] de la Torre, A., Peinado, A. M., Segura, J. C., Perez-Cordoba, J. L., Benitez, M. C., and Rubio, A. J., "Histogram equalization of speech representation for robust speech recognition", IEEE Transactions on Speech and Audio Processing (2005) 355–366

[7] Hilger, F. and Ney, H., "Quantile based histogram equalization for noise robust speech recognition", Eurospeech (2001) 1135–1138

[8] Mesot, B. and Barber, D., "Switching linear dynamical systems for noise robust speech recognition", IEEE Transactions on Audio, Speech and Language Processing (2007)

[9] Droppo, J. and Acero, A., "Noise robust speech recognition with a switching linear dynamic model", Proceedings of the International Conference on Acoustics, Speech and Signal Processing (2004)

[10] Cooke, M. and Scharenborg, O., "The Interspeech 2008 Consonant Challenge", Interspeech (2008)

[11] Deng, J., Bouchard, M., and Yeap, T. H., "Noisy speech feature estimation on the Aurora2 database using a switching linear dynamic model", Journal of Multimedia (2007) 47–52

[12] Dempster, A. P., Laird, N. M., and Rubin, D. B., "Maximum likelihood from incomplete data via the EM algorithm", Journal of the Royal Statistical Society, Series B (1977) 1–38

[13] Droppo, J., Deng, L., and Acero, A., "A comparison of three nonlinear observation models for noisy speech features", Eurospeech (2003) 681–684

[14] Bar-Shalom, Y., and Li, X. R., "Estimation and tracking: principles, techniques, and software", Artech House, Norwood, MA (1993)

[15] Doddington, G. R. and Schalk, T. B., "Speech recognition: turning theory to practice", IEEE Spectrum (1981) 26–32

[16] Grimm, M., Kroschel, K., Harris, H., Nass, C., Schuller, B., Rigoll, G., and Moosmayr, T., "On the necessity and feasibility of detecting a driver's emotional state while driving", Proceedings of ACII 2007, 2nd Int. Conf. on Affective Computing and Intelligent Interaction (2007) 126–138

[17] Hermansky, H., "Perceptual linear predictive (PLP) analysis of speech", J. Acoust. Soc. Amer. (1990) 1738–1752