Switching Linear Dynamic Models for Recognition of Emotionally Colored and Noisy Speech

Martin Wöllmer, Nikolaj Klebert, Björn Schuller

Technische Universität München, Institute for Human-Machine Communication, Theresienstr. 90, 80333 München, Germany
E-Mail: {woellmer,schuller}@tum.de
Web: www.mmk.ei.tum.de

Abstract

Model-based speech feature enhancement techniques were shown to be a promising approach towards increasing the robustness of automatic speech recognition in noisy environments. Strategies that model speech with a Switching Linear Dynamic Model (SLDM) have been successfully applied to noisy speech recognition tasks, since they overcome the limitations of GMM- or HMM-based approaches. However, SLDM-based feature enhancement has so far only been investigated for the recognition of isolated words or relatively friendly scenarios such as connected digit recognition under the presence of additive noise using whole word models (e. g. the AURORA task). In order to give an impression of the effectiveness of SLDM speech modeling for more challenging ASR applications, we evaluate SLDM feature enhancement for continuous recognition of spontaneous and emotionally colored speech in the noise. As backend we use tied-state triphone models trained and evaluated on the SAL Corpus. Applying SLDM-based feature enhancement, we achieve an average relative performance gain of almost 20 % when considering diverse noise settings.

1 Introduction

Applying Automatic Speech Recognition (ASR) systems in noisy surroundings usually leads to a lower recognition accuracy when compared to ASR performance in clean conditions. In order to maintain an acceptable performance in noisy environments, many different strategies have been proposed [12, 8, 10]. Approaches to increase noise robustness can be roughly categorized in speech preprocessing, model adaptation, and feature enhancement. Speech preprocessing techniques are applied before feature extraction and comprise methods like Wiener filtering or spectral subtraction [9], whereas model adaptation approaches aim at adapting e. g. phoneme models to the noisy environment or use models trained on noisy speech. Feature enhancement operates in the feature domain, meaning that it tries to determine the clean speech features from the observed noisy features. This can be done by either using a priori knowledge about how noise affects speech features (Cepstral Mean Normalization, Histogram Equalization [2]) or by building general models for speech and noise (model-based feature enhancement). Recently, extensive evaluations of different noisy speech recognition scenarios led to the finding that modeling speech with a Switching Linear Dynamic Model (SLDM) for model-based feature enhancement as introduced in [5] leads to good results. Feature enhancement algorithms that use an SLDM for speech modeling overcome some of the drawbacks of techniques using e. g. Gaussian Mixture Models (GMM) or Hidden Markov Models (HMM), since the dynamics of the SLDM capture the smooth time evolution of speech and do not produce artifacts such as sharp single frame transitions.

So far, experiments on SLDM-based feature enhancement have focused on evaluations using a whole word model backend for the recognition of isolated or connected digits [12, 5, 13] or letters [11]. In this paper we want to give an impression of the effectiveness of SLDM feature enhancement for a more challenging ASR scenario. Therefore we evaluate the impact of noise on the recognition performance when testing a continuous speech recognition engine on spontaneous and emotionally colored speech. Using feature enhancement based on a global SLDM that is trained on clean speech and on a Linear Dynamic Model (LDM) for noise, we investigate the performance gain that can be obtained for different noise conditions. All experiments are conducted on the Sensitive Artificial Listener (SAL) corpus [4] which was recorded during natural, spontaneous, and emotional human-machine interactions.

The structure of this paper is as follows: Section 2 briefly reviews the principle of SLDM feature enhancement as it will be applied in our experiments. Section 3 describes the SAL database, as well as our experiments and results. Section 4 contains a discussion of the obtained results and concluding remarks.

2 SLDM Feature Enhancement

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. As in [5] we use a Switching Linear Dynamic Model to capture the dynamics of clean speech. Similar to Hidden Markov Model (HMM) based approaches to model clean speech, the SLDM assumes that the signal passes through various states. Conditioned on the state sequence the SLDM furthermore enforces a continuous state transition in the feature space.

2.1 Modeling of Noise

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

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

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

ISBN 978-3-8007-3300-2

like colored stationary Gaussian noise. Alternatively to the graphical model in Figure 1 the following equations can be used to express the LDM:

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

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

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.

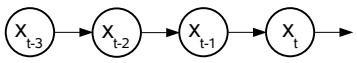

Figure 1: Linear Dynamic Model for noise

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:

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

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 shown in Figure 2 and the models used in [3] 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:

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

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

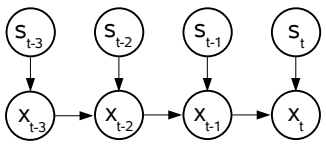

Figure 2: Switching Linear Dynamic Model for speech

To train the parameters $A(s)$, $b(s)$ and $C(s)$ of the SLDM conventional EM techniques are used. 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 [6]. 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.

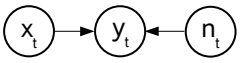

Figure 3: Observation model for noisy speech y_t

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 [1]. 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 and S represents the number of hidden states, 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 [5]. During the *collapse* step the posterior for x_{t-1} is marginalized over the states that occur r frames in the past by approximating each Gaussian mixture by a single Gaussian component via moment matching (meaning that Gaussians which share a history of $r-1$ frames are collapsed together). In the *prediction* step each of the hypothesis that remain after collapsing are branched out once for each of the S possible states s_t in order to obtain a posterior for x_t. Finally, in the *observation* step the observed noisy speech frame y_t is incorporated (see [5] for more details).

The Gaussian posterior $q(x_t, y_{1:t})$ 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 $E[x_t|y_{1:t}]$:

$$E[x_t|y_{1:t}] \cong \frac{\int x_t q(x_t, y_{1:t}) dx_t}{\int q(x_t, y_{1:t}) dx_t} \quad (7)$$

3 Experiments and Results

For all experiments we use the SAL corpus, which is a sub-set of the HUMAINE database [4]. In order to induce spontaneous, emotional speech, the four speakers (two male, two female) were interrogated separately by four virtual characters: Poppy (who is happy), Obadiah (who is gloomy), Spike (who is angry), and Prudence (who is pragmatic). The database was recorded using a Wizard-of-Oz scenario in which the four virtual characters were

ISBN 978-3-8007-3300-2

imitated by a human operator. The goal of the operator was to induce emotions in the speaker. Thereby the induced emotions should correspond to the personality of the respective character.

The corpus consists of 25 recordings with an average length of 20 minutes. The recordings were split into 1 692 speech turns with an average length of 3.5 seconds per turn. 80 % of the speech turns were randomly selected as training set, while the remaining 20 % were used for testing. All utterances of the test set were superposed with four different noises from the NOISEX database (babble, car, pink, and white noise) using different SNR levels (20, 15, 10, 5, and 0 dB).

Recognition performance with and without SLDM-based feature enhancement was evaluated for every noise condition and for the clean case. As features we used cepstral mean normalized MFCCs 0 to 12 with their first and second order delta coefficients. A global SLDM consisting of 16 states was trained on the clean training fraction of the SAL database. The utterance-specific LDM for noise was computed from the first and last ten frames of each noisy utterance. Thereby the noise model consisted of a single Gaussian mixture component. For feature enhancement we used a history parameter $r = 1$ (see Section 2.4).

As backend recognizer, we used left-to-right tied-state word internal triphone HMMs consisting of three hidden states per phoneme, whereas we used 16 Gaussian mixtures per state. All HMMs as well as a bigram language model were trained on the clean SAL training set. For non-verbal vocalizations (such as laughing, sighing, coughing, etc.) we trained monophone HMMs with 9 hidden states. The HMMs were trained and optimized using HTK [17]. Thereby the initial monophone models consisted of one Gaussian mixture per state. All initial means and variances were set to the global means and variances of all feature vector components (flat start initialization). The monophone models were then trained using four iterations of embedded Baum-Welch re-estimation. After that, the monophones were mapped to tied-state word internal triphone models with shared state transition probabilities. Two Baum-Welch iterations were performed for re-estimation of the triphone models. Finally, the number of mixture components of the triphone models was increased to 16 in four successive rounds of mixture doubling and re-estimation (four iterations in every round). In each round the newly created mixture components are copied from the existing ones, mixture weights are divided by two, and the means are shifted by plus and minus 0.2 times the standard deviation.

WA [%]	baseline	SLDM	rel. gain
clean	55.01	53.84	-2.17
car	50.38	52.74	4.48
babble	25.14	29.41	14.51
pink	14.68	22.98	36.11
white	8.43	15.60	45.93
mean	**30.73**	**34.91**	**19.77**

Table 1: Word accuracies on the SAL test set for different noise types, with and without SLDM feature enhancement (results averaged over 20 to 0 dB SNR conditions)

Table 1 shows the obtained word accuracies for the baseline HMM recognizer using standard MFCC features as well as for the HMM recognizer processing features enhanced using the SLDM approach. In clean conditions SLDM enhancement leads to a slight but hardly significant degradation of the recognition performance. In noisy conditions we can observe relative gains of between 4.48 % and 45.93 %. Note that the results for noisy test data as shown in Table 1 are averaged over 20 to 0 dB SNR conditions. When considering a potential Sensitive Artificial Listener application which presumes the recognition of emotional speech in noisy surroundings, the most interesting noise scenario is the babble noise type (e. g. when using a SAL system while other people talk in the background). In this case SLDM enhancement leads to an accuracy of between 47.98 % (20 dB SNR) and 10.70 % (0 dB SNR), corresponding to baseline MFCC results of 43.96 % and 7.71 %, respectively (see Table 2). Superposing speech with car noise leads to the lowest ASR performance degradation since this noise type is rather stationary, and due to the low-pass characteristics of car noise, there is no full spectral overlap of speech and noise. Applying SLDM enhancement, a word accuracy of over 50 % can be maintained even at an SNR level of 0 dB. White noise causes the highest performance loss when using the baseline recognition system and leads to comparably large relative gains through SLDM feature enhancement.

On average, we achieve a relative performance gain of 19.77 % which illustrates that SLDM feature enhancement is an effective approach to improve the accuracy of continuous ASR systems trained and evaluated on spontaneous and emotional speech.

WA [%]	SNR	baseline	SLDM	rel. gain
car	20 dB	54.31	53.70	-1.14
	15 dB	54.11	53.61	-0.93
	10 dB	52.52	53.17	1.22
	5 dB	48.12	52.55	8.43
	0 dB	42.82	50.65	15.46
babble	20 dB	43.96	47.98	8.38
	15 dB	36.19	40.23	10.04
	10 dB	24.40	29.35	16.87
	5 dB	13.43	18.77	28.45
	0 dB	7.71	10.70	27.94
pink	20 dB	37.65	46.30	18.68
	15 dB	22.17	33.81	34.43
	10 dB	9.47	20.67	54.18
	5 dB	3.23	9.94	67.51
	0 dB	0.88	4.16	78.85
white	20 dB	24.13	35.01	31.08
	15 dB	11.29	22.29	49.35
	10 dB	4.31	11.73	63.26
	5 dB	1.70	6.69	74.59
	0 dB	0.73	2.26	67.70

Table 2: Word accuracies on the SAL test set for different noise types and SNR levels, with and without SLDM feature enhancement

4 Discussion and Conclusion

In contrast to previous experiments on SLDM feature enhancement which focus on comparably easy ASR tasks such as recognizing digit sequences (e. g. the AURORA 2 task [7]), we investigated the effect of model-based feature enhancement on ASR performance in a challenging real-life application. Since virtual agents such as the ‘Sensitive

ISBN 978-3-8007-3300-2

Artificial Listener' are often applied in noisy surroundings, it is important to evaluate the effectiveness of feature preprocessing algorithms off-line before implementing real-time versions of the enhancement techniques. This paper shows that for the task of recognizing spontaneous emotionally colored speech, SLDM feature enhancement leads to an average relative performance gain of almost 20 %, which motivates potential real-time implementations of the SLDM approach. Of course the absolute accuracies of about 50 % as reported in this paper are far lower than typical accuracies obtained for read and well-articulated speech, however, for the SAL scenario it is sufficient to parse the recognition output for specific keywords (e. g. as in [15]) rather than to obtain the fully correct transcription. Moreover, when evaluating noise robustness, we are predominantly interested in *relative* improvements.

In future experiments we plan to compare SLDM feature enhancement to other popular techniques such as Histogram Equalization and to combine it with Long Short-Term Memory [14, 16] phoneme modeling. A promising step towards improving SLDM-based feature enhancement is a more accurate estimation of the posterior distribution by increasing the history parameter r. Furthermore, the introduction of discrete SLDM state transition probabilities might result in better feature enhancement.

Acknowledgements

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] Y. Bar-Shalom and X. R. Li. *Estimation and tracking: principles, techniques, and software*. Artech House, Norwood, MA, 1993.

[2] A. de la Torre, A. M. Peinado, J. C. Segura, J. L. Perez-Cordoba, M. C. Benitez, and A. J. Rubio. Histogram equalization of speech representation for robust speech recognition. *IEEE Transactions on Speech and Audio Processing*, 13(3):355–366, 2005.

[3] 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*, pages 47–52, 2007.

[4] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J. C. Martin, L. Devillers, S. Abrilian, A. Batliner, N. Amir, and K. Karpouzis. The HUMAINE database: addressing the collection and annotation of naturalistic and induced emotional data. In *Affective Computing and Intelligent Interaction*, volume 4738/2007. Springer, 2007.

[5] J. Droppo and A. Acero. Noise robust speech recognition with a switching linear dynamic model. In *Proc. of ICASSP*, Montreal, Canada, 2004.

[6] J. Droppo, L. Deng, and A. Acero. A comparison of three non-linear observation models for noisy speech features. In *Proc. of Eurospeech*, pages 681–684, 2003.

[7] H. G. Hirsch and D. Pearce. The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions. In *ISCA ITRW ASR2000: Automatic Speech Recognition: Challenges for the Next Millennium*, Paris, France, 2000.

[8] 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*, 7, 1999.

[9] G. Lathoud, M. Magimia-Doss, B. Mesot, and H. Boulard. Unsupervised spectral subtraction for noise-robust ASR. In *Proc. of ASRU*, San Juan, Puerto Rico, 2005.

[10] D. Li, 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*, 11:568–580, 2003.

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

[12] B. Schuller, M. Wöllmer, T. Moosmayr, and G. Rigoll. Recognition of noisy speech: A comparative survey of robust model architecture and feature enhancement. *Journal on Audio, Speech, and Music Processing*, 2009. ID 942617.

[13] S. Windmann and R. Haeb-Umbach. Modeling the dynamics of speech and noise for speech feature enhancement in ASR. In *Proc. of ICASSP*, Las Vegas, Nevada, 2008.

[14] M. Wöllmer, F. Eyben, A. Graves, B. Schuller, and G. Rigoll. Bidirectional LSTM networks for context-sensitive keyword detection in a cognitive virtual agent framework. *Cognitive Computation, Special Issue on Non-Linear and Non-Conventional Speech Processing*, 2010.

[15] M. Wöllmer, F. Eyben, J. Keshet, A. Graves, B. Schuller, and G. Rigoll. Robust discriminative keyword spotting for emotionally colored spontaneous speech using bidirectional LSTM networks. In *Proc. of ICASSP*, Taipei, Taiwan, 2009.

[16] M. Wöllmer, B. Schuller, F. Eyben, and G. Rigoll. Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening. *IEEE Journal of Selected Topics in Signal Processing, Special Issue on Speech Processing for Natural Interaction with Intelligent Environments*, 2010.

[17] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. *The HTK book (v3.4)*. Cambridge University Press, 2006.

ISBN 978-3-8007-3300-2