Discriminative Adaptation for Log-linear Acoustic Models

Jonas Lööf, Ralf Schlüter, and Hermann Ney

Lehrstuhl für Informatik 6 - Computer Science Dept.
RWTH Aachen University, Aachen, Germany
{loof,schlueter,ney}@cs.rwth-aachen.de

Abstract

Log-linear models have recently been used in acoustic modeling for speech recognition systems. This has been motivated by competitive results compared to systems based on Gaussian models, and a more direct parametrisation of the posterior model. To competitively use log-linear models for speech recognition, important methods, such as speaker adaptation, have to be reformulated in a log-linear framework. In this work, an approach to log-linear affine feature transforms for speaker adaptation is described. Experiments for both supervised and unsupervised adaptation are presented, showing improvements over a maximum likelihood baseline in the form of feature space maximum likelihood linear regression for the case of supervised adaptation.

Index Terms: speech recognition, adaptation, log-linear models

1. Introduction

In recent years, discriminatively trained log-linear models have been successfully used in hidden Markov model (HMM) based acoustic models for automatic speech recognition (ASR), taking the place of Gaussian mixture models (GMMs). Log-linear models have been shown to be competitive or better than comparable discriminatively trained Gaussian models [1].

In order to successfully use log-linear modeling in a state-of-the-art speech recognition system, it is necessary to reproduce or replace all important methods used in improving such a system. Speaker adaptation is one important method to improve the performance of a speech recognition system, and especially the use of feature space maximum likelihood linear regression (fMLLR) speaker adaptive training (SAT) [2], has proved to be an important part of state-of-the-art systems [3]. Thus, it is important to develop and investigate adaptation methods for log-linear models if they are to replace Gaussian models in a state-of-the-art speech recognition system.

In this work, speaker adaptation of log-linear models, using affine feature transforms, similar to fMLLR, is introduced. This uses methods similar to the estimation of linear feature transforms for log-linear models that was introduced in [4]. Related work include discriminative adaptation of Gaussian models using the maximum mutual information criterion, as described for instance in [5].

2. Log-linear Models in ASR

The use of log-linear modeling in ASR has recently been introduced. It can be motivated as a more direct formulation of the posterior probability described by a (discriminatively trained) Gaussian model.

Assume a sequence of acoustic vectors x_1^T, and a word sequence w_1^N. From [6], the log-linear hidden Markov model (LHMM) defines a posterior probability

$$P_\Lambda(w_1^N|x_1^T) = \frac{1}{Z_\Lambda(x_1^T)} \prod_{n=1}^{N} \underbrace{\exp(\alpha''_{w_{n-1},w_n})}_{\text{language model}} \cdot \qquad (1)$$

$$\cdot \sum_{s_1^T \in w_1^N} \prod_{t=1}^{T} \underbrace{\exp(\alpha'_{s_{t-1},s_t})}_{\text{transition model}} \underbrace{\exp(\alpha_{s_t} + \lambda_{s_t}^\mathsf{T} x_t)}_{\text{emission model}}$$

for a large vocabulary ASR system, where $Z_\Lambda(x_1^T)$ denotes the normalization constant. The parameters of the model are given by the scalars $\alpha''_{v,w}$, $\alpha'_{s,s'}$, α_s, and the vectors λ_s, and the total set of all parameters is denoted by $\Lambda = \{\alpha''_{v,w}, \alpha'_{s,s'}, \alpha_s, \lambda_s\}$.

The LHMM is a linear-chain hidden conditional random field [7] with the same model structure as a Gaussian hidden Markov model, and log-linear parametrization of the submodels. This is not a log-linear model in itself, due to the sum over the state sequences in the numerator. Due to the close relationship to log-linear models it will still be included under the term *log-linear modeling* in the present work.

As described in [6], the estimation of the parameters of such a model is typically performed using the maximum mutual information (MMI) criterion, i.e by maximizing the log-posterior. Conventional lattice based MMI replaces the sum over all word sequences in the normalization constant $Z_\Lambda(x_1^T)$, by an approximation based on a word lattice. The log-posterior to be optimized over the model parameters is given by

$$Q(\Lambda) = \log P_\Lambda^{\text{lattice}}(w_1^N|x_1^T), \qquad (2)$$

Due to the sum in the numerator of the LHMM posterior this is a non-convex criterion. To achieve more robust optimization, it would be advantageous to use a convex formulation instead, since one of the advantages of LHMMs is the existence of such criteria. In [1], two convex optimization criteria are described, the first being a frame posterior based criterion, similar to the criteria used for hybrid neural network HMM models [8], the second being a convex modification of the lattice based MMI criterion. Nevertheless, experience shows that also the normal lattice based MMI criterion works well in practice, and is used for parameter estimation in the present work, since it has lower complexity than the modified convex lattice based MMI criterion.

The resulting optimization problem can be approached using any optimization method. In [6] as well as in the present work, the Rprop method [9], a gradient based method that only takes into account the sign of the gradient, is used.

3. Log-linear Model Adaptation

To devise a speaker adaptation method for LHMMs, it is instructive to look at the formulation of adaptation for the case of Gaussian models for inspiration. One important speaker adaptation method is fMLLR, that can be used both for recognition side adaptation, and speaker adaptive training. fMLLR consist of an affine transformation of the acoustic feature vector x_t, such that $x'_t = Ax_t + b$, where the matrix A and the vector b are speaker dependent parameters.

Single global feature transforms have recently been used to improve classification performance when using log-linear modeling [4]. A similar approach can be utilized for speaker adaptation.

By extending the acoustic feature vector x_t with a constant element, such that the new feature vector $\xi_t = [1 \; x^{\mathsf{T}}]^{\mathsf{T}}$, and combining the model parameters α_s and λ_s into one vector $\theta_s = [\alpha_s \; \lambda_s^{\mathsf{T}}]^{\mathsf{T}}$, the emission model part, $E(s_1^T, t)$, of Equation 1 can be rewritten such that

$$\begin{aligned} E(s_1^T, t) &= \exp\left(\alpha_{s_t} + \lambda_{s_t}^{\mathsf{T}} x_t\right) \\ &= \exp\left(\theta_{s_t}^{\mathsf{T}} \xi_t\right). \end{aligned} \qquad (3)$$

Modifying Equation 1 to include an affine transformation of the features gives

$$\begin{aligned} E(s_1^T, t) &= \exp\left(\alpha_{s_t} + \lambda_{s_t}^{\mathsf{T}} (Ax_t + b)\right) \\ &= \exp\left(\theta_{s_t}^{\mathsf{T}} W \xi_t\right) \\ &= \exp\left(\sum_{i=1}^{D}\sum_{j=1}^{D} \theta_{s_t}^{j\,\mathsf{T}} W^{ij} \xi_t^i\right). \end{aligned} \qquad (4)$$

where the extended transformation matrix W is defined as

$$W = \begin{bmatrix} 1 & 0 \\ b & A \end{bmatrix} \qquad (5)$$

to give a parametrization identical to fMLLR. It would also be possible to use an unconstrained matrix W, but this was not done in the present work. From the explicit component form of Equation 4 it can be seen that if the model parameters θ_s are kept fixed, the resulting emission model is log-linear in the matrix components W_{ij}.

As in the previous section, the estimation of the parameters of the transformation matrix can be performed using different criteria. In the present work the segment-wise lattice based MMI criterion is used. By including the adaptation matrix W, the criterion from Equation 2 changes to

$$Q(\Lambda, W) = \log P_{\Lambda,W}^{\text{lattice}}(w_1^N | x_1^T), \qquad (6)$$

where $P_{\Lambda,W}^{\text{lattice}}(w_1^N | x_1^T)$ is equivalent to Equation 1, but with the emission model exchanged with $E(s_1^T, t)$ from Equation 4, and the normalization approximated over a word lattice. As with the estimation of the regular emission model parameters, this criterion is optimized using the Rprop criterion.

From Gaussian models we know that there exists an equivalence between fMLLR and so called constrained maximum likelihood linear regression, that is a transformation of the model, where the mean and covariances are transformed using the same matrix. Using straightforward matrix algebra, a similar equivalence can be shown also for this case;

$$E(s_1^T, t) = \exp\left(\theta_{s_t}^{\mathsf{T}} W \xi_t\right) = \exp\left(\left(W^{\mathsf{T}} \theta_{s_t}\right)^{\mathsf{T}} \xi_t\right). \qquad (7)$$

It is thus clear that the transformation $\xi'_t = W\xi_t$ of the (extended) features, is equivalent to the transformation $\theta'_s = W^{\mathsf{T}}\theta_s$ of the model. Note that no offset vector is included in the model transformation formulation. This is to be expected since the output of a LHMM is invariant to a global offset to the emission parameters.

4. Experiments

Experiments were conducted based on the European Parliament plenary sessions (EPPS) corpus, from the TC-STAR project [3]. The EPPS task is a transcription task, where adaptation normally is used in an unsupervised framework. In the present work, it was decided to evaluate the use of both supervised and unsupervised adaptation. To facilitate this, the original EPPS 2006 development corpus was split into two corpora, one adaptation corpus, used for supervised adaptation, and one test corpus, on which the speech recognition results were produced. For the current experiments, the (automatic) segmentation and speaker clustering were taken as given, and for each speaker cluster, the first half of the segments were assigned to the adaptation corpus, and the rest to the test corpus. Table 1 gives the statistics of the resulting corpora.

Table 1: *Adaptation and test corpora*

	Adapt	Test
Net Duration	1.55h	1.64h
# Segments	356	370
# Speaker clust.	32	
# Running words	13704	14784
Perplexity	107.0	112.7
OOV Rate	0.63	0.78

The baseline acoustic model used for the adaptation experiments was taken from a system developed during the 2007 TC-STAR evaluation. The system used a MFCC front-end augmented with a single voicedness feature, and the acoustic models used 4500 states. Furthermore, in all experiments a one pass VTLN method, using a classifier for warping factor estimation, was used.

The goal of this work is to demonstrate the effectiveness of speaker adaptation for LHMMs. To facilitate the conversion of the acoustic model to log-linear form, it was decided to use single Gaussian models, sharing a single globally pooled covariance (as opposed to the Gaussian mixture models used in the final system in the evaluation.) It should be noted that the recognition performance of the system used in the following experiments are lower than those for a state of the art speech recognition system. Since the eventual goal of LHMM adaptation is to use it with models trained from scratch in a log-linear framework, such as those described in [10], the following experiments should be seen as a demonstration of feasibility.

In all the experiments, per-speaker fMLLR matrices were first estimated, using the original Gaussian models. These maximum likelihood estimated matrices were used to initialize the discriminative log-linear adaptation matrix estimation. The log-linear adaptation matrices were estimated based on the lattice based MMI criterion, using the Rprop optimization criterion. The resulting matrices were used as feature transforms, and the performance of the discriminative adaptation was evaluated after each iteration.

4.1. Supervised Adaptation

For the supervised experiments, the acoustic model was adapted on the adaptation corpus for each speaker cluster, using the available manual transcriptions. Table 2 shows the baseline and fMLLR results for the case of supervised adaptation. In figure 1, the word error rate (WER) is plotted as a function of the number of iterations of log-linear based MMI estimation performed, where iteration zero means the performance with the initializing fMLLR transform. Figure 2 show the development of the sum of the objective function of Equation 6 over all speaker clusters, as a function of the number of estimation iterations.

Table 2: *Baseline and supervised fMLLR results*

Iteration	WER[%]
Baseline	31.9
fMLLR Iter. 1	27.2
fMLLR Iter. 2	26.8
fMLLR Iter. 3	26.8
fMLLR Iter. 4	26.8

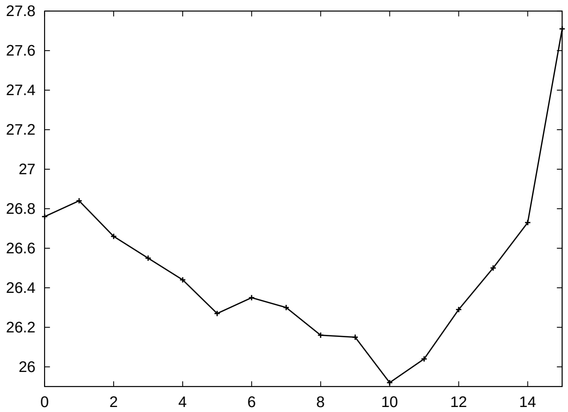

Figure 1: WER [%] over iterations, supervised adaptation

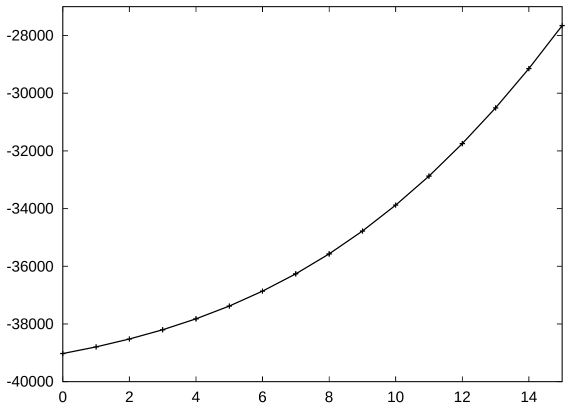

Figure 2: Objective function over iterations, supervised adaptation

4.2. Unsupervised Adaptation

Unsupervised discriminative training does not typically lead to large improvements, compared to maximum likelihood adaptation. Nevertheless, if one wants to use exclusively log-linear modelling in a typical ASR system, where unsupervised adaptation plays an important part, its use will be needed. Table 3 shows the baseline and maximum likelihood results. In figures 3 and 4, the WER and objective function for successive iterations of the MMI estimation are shown.

Table 3: *Baseline and unsupervised fMLLR results*

Iteration	WER[%]
Baseline	31.9
fMLLR Iter. 1	27.8
fMLLR Iter. 2	27.5
fMLLR Iter. 3	27.5
fMLLR Iter. 4	27.5

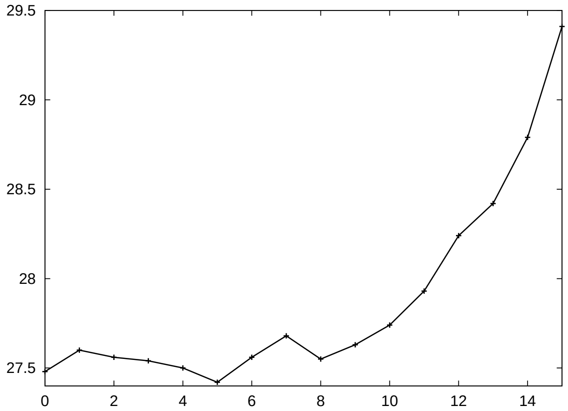

Figure 3: WER [%] over iterations, unsupervised adaptation

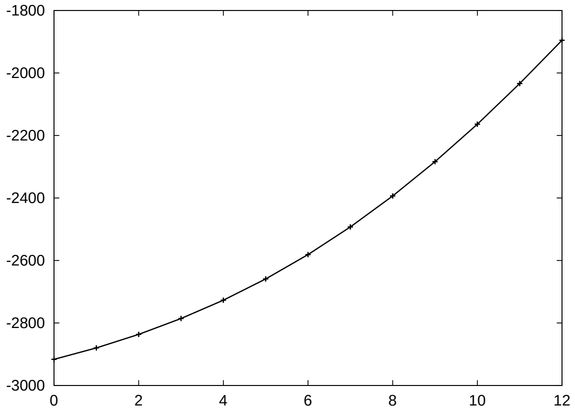

Figure 4: Objective function over iterations, unsupervised adaptation

4.3. Discussion of Results

Table 4 summarizes the important results from the experiments. It can be seen that the log-linear speaker adaptation give a sizable improvement over fMLLR, in a supervised framework. Using unsupervised adaptation though, the improvement is minimal. This agrees with previous results, where either small or no improvements have been observed from discriminative unsupervised adaptation, compared to the maximum likelihood case; see for instance [11].

Table 4: Summary of results

	Supervised	Unsupervised
Baseline	31.9	
fMLLR	26.8	27.5
LL Adapt	25.9	27.4

In both the case of supervised and unsupervised adaptation we see that the optimum in WER is reached without any convergence in the objective function being reached, indicating that severe overfitting takes place. It would be advantageous if a more direct correspondence between objective function and word error rate could be achieved, since one of the potential advantages of the log-linear formulation is the existence of convex objective functions. On the other hand, it must be remembered that no regularization was used in these experiments. In future work, the use of regularization will be investigated.

5. Conclusions

This paper presented work on discriminative speaker adaptation for log-linear acoustic models. A method for affine feature transform estimation using the maximum mutual information criterion is described, and its use for speaker adaptation is discussed.

By applying supervised log-linear discriminative features space adaptation, a substantial improvement is seen in comparison to maximum likelihood adaptation using the same parametrization. This is consistent with previous results comparing maximum likelihood and discriminative adaptation [5]. When applying log-linear adaptation in an unsupervised framework, no reliable improvements can be observed, as expected from previous results from discriminative unsupervised adaptation [11].

Future work includes investigating the different optimization criteria for LHMM adaptation, including the use of regularization. Additionally the use of different adaptation matrices for different LHMM states, as well as different parameter tyings, will be investigated.

6. Acknowledgements

This work was partly funded by the European Community's Seventh Framework Programme under the project JUMAS (FP7-214306).

7. References

[1] Heigold, G., Rybach, D., Schlüter, R., and Ney, H., "Investigations on convex optimization using log-linear HMMs for digit string recognition," in *Interspeech*, Brighton, U.K., Sep. 2009, pp. 216–219.

[2] Gales, M. J. F., "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, no. 2, pp. 75 – 98, Apr. 1998.

[3] Lööf, J., Gollan, C., Hahn, S., Heigold, G., Hoffmeister, B., Plahl, C., Rybach, D., Schlüter, R., and Ney, H., "The RWTH 2007 TC-STAR evaluation system for European English and Spanish," in *Proc. Int. Conf. on Spoken Language Processing*, Antwerp, Belgium, Aug. 2007, pp. 2145 – 2148.

[4] Tahir, M. A., Heigold, G., Plahl, C., Schlüter, R., and Ney, H., "Log-linear framework for linear feature transformations in speech recognition," in *IEEE Automatic Speech Recognition and Understanding Workshop*, Merano, Italy, Dec. 2009.

[5] Uebel, L. F. and Woodland, P. C., "Discriminative linear transforms for speaker adaptation," in *ISCA ITR-Workshop on Adaptation Methods in Speech Recognition*, Sophia Antipolis, France, Aug. 2001, pp. 61 – 64.

[6] Heigold, G., Wiesler, S., Nussbaum, M., Lehnen, P., Schlüter, R., and Ney, H., "Discriminative HMMs, log-linear models, and CRFs: What is the difference?" in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Dallas, Texas, USA, Mar. 2010, pp. 5546–5549.

[7] Gunawardana, A., Mahajan, M., Acero, A., and Platt, J., "Hidden conditional random fields for phone classification," in *Interspeech*, Lisbon, Portugal, Sep. 2005.

[8] Robinson, T., Hochberg, M., and Renals, S., "The use of recurrent networks in continuous speech recognition," in *Automatic Speech and Speaker Recognition*, K. K. Paliwal C.-H. Lee, F. K. S., Ed., Kluwer Academic Publishers, Norwell, MA, USA, 1996.

[9] Riedmiller, M. and Braun, H., "A direct adaptive method for faster backpropagation learning: The Rprop algorithm," in *ICNN*, San Francisco, CA, USA, 1993.

[10] Wiesler, S., Nußbaum, M., Heigold, G., Schlüter, R., and Ney, H., "Investigations on features for log-linear acoustic models in continuous speech recognition," in *IEEE Automatic Speech Recognition and Understanding Workshop*, Merano, Italy, Dec. 2009.

[11] Lööf, J., Schlüter, R., and Ney, H., "Efficient estimation of speaker-specific projecting feature transforms," in *Proc. Int. Conf. on Spoken Language Processing*, Antwerp, Belgium, Aug. 2007, pp. 1557 – 1560.