Investigations on Convex Optimization Using Log-Linear HMMs for Digit String Recognition

Georg Heigold, David Rybach, Ralf Schlüter, Hermann Ney

RWTH Aachen University
Chair of Computer Science 6 – Computer Science Department
D-52056 Aachen, Germany

{heigold,rybach,schlueter,ney}@cs.rwth-aachen.de

Abstract

Discriminative methods are an important technique to refine the acoustic model in speech recognition. Conventional discriminative training is initialized with some baseline model and the parameters are re-estimated in a separate step. This approach has proven to be successful, but it includes many heuristics, approximations, and parameters to be tuned. This tuning involves much engineering and makes it difficult to reproduce and compare experiments. In contrast to the conventional training, convex optimization techniques provide a sound approach to estimate all model parameters from scratch. Such a straight approach hopefully dispense with additional heuristics, e.g. scaling of posteriors. This paper addresses the question how well this concept using log-linear models carries over to practice. Experimental results are reported for a digit string recognition task, which allows for the investigation of this issue without approximations.

Index Terms: convex optimization, conditional random fields, acoustic modeling, digit string recognition

1. Introduction

Discriminative training has been successfully employed for speech recognition. A shortcoming of the conventional discriminative training criteria is that these do not guarantee to converge to the global optimum (e.g. non-convex training criteria) but can get stuck in local optima. Hence, the outcome depends on the initialization of the discriminative training and the choice of the optimization algorithm. This situation is unsatisfactory because the parameter optimization in speech recognition requires much engineering and involves many heuristics (e.g. splitting of densities) and approximations (e.g. word lattices) to make it work well in practice. This circumstance does not only make it hard to reproduce experiments but strictly speaking, also makes the fair comparison of different algorithms questionable due to spurious local optima. For these reasons, a fool-proof training criterion without any parameters to be tuned manually would be attractive in this context.

Convex optimization appears to be a promising concept to avoid the above described problems with conventional discriminative training. This concept has been used successfully in many fields of pattern recognition, e.g. Support Vector Machines (SVMs). In speech recognition, however, the construction of convex training criteria is more tricky mainly because of the hidden variables. In conventional speech recognition systems, hidden variables occur on two different levels: the HMM state sequences accounting for time distortions and the densities in the Gaussian mixtures. In this work, Conditional Random Fields (CRFs) [1] with HMM structure [2, 3] and log-linear parameterization are used. In this approach, the density indices are hidden, and are eliminated by using single densities in combination with an increased number of features. The motivation for this design decision is the fact that it is hard to initialize the density indices without a reasonable generative model. Note that this is an important difference to the work in [4, 3, 5]. The initialization of some log-linear models by generative models is possible due to the equivalence relation of Gaussian and log-linear models [6]. For instance, such a log-linear model using first and second order features (i.e., features of the type x_d and $x_d x_{d'}$) is equivalent to a Gaussian HMM (GHMM) with density-specific full covariance matrices. Due to the parameter constraints of the GHMM, the optimization of the GHMM is much more complex [5] than for the corresponding log-linear model. Assuming that the HMM state sequence is known and kept fixed during training in addition to using single densities, a convex training criterion can then be derived for HMMs, cf. [2, 3, 5].

This leads us to the main question of this paper. Using a convex and parameter-free training criterion, can the log-linear model parameters be reliably (i.e., competitive error rates) estimated from scratch? Similar work can be found in [2, 3, 5] but with slightly different focus.

As suggested above, the definition of a convex training criterion is not unique. The standard entropy-based (MMI) training criterion is employed in [2] and [3]. The first work is based on Maximum Entropy Markov Models (MEMMs) while the latter uses CRFs. In contrast, the GHMM parameters in [5] are optimized using a margin-based training criterion. This approach is motivated by the SVM for HMMs proposed in [7] for non-speech applications. Here, the training criterion is based on modified MMI [8]. This training criterion incorporates a margin term into the existing MMI criterion, approximating multiclass SVMs [7]. This allows us to re-use our transducer-based discriminative framework [8]. In addition, we did not use the approach from [5] because for separable data for instance, the margin can be made arbitrarily large by suitably scaling the model parameters. Modified MMI uses regularization to avoid this effect and is probably safer for the estimation from scratch.

The remainder of this paper is organized as follows. Section 2 discusses several practical issues with convex optimization in speech recognition and defines the objectives of this work. Section 3 introduces the log-linear model under consideration and discusses different MMI-based training criteria. Experimental results are presented in Section 4. The paper concludes with the summary of the experimental findings.

2. Practical Issues & Objectives

The definition of a convex training criterion for speech recognition leaves open several issues in practice. First, the key assumption is that the initial alignment is known and kept fixed during training. So, how sensitive is the performance on this initial alignment? In general, the model used to generate the initial alignment is not related with the log-linear model to be estimated. This is the case if the log-linear model cannot be initialized with a corresponding GHMM as e.g. in [3, 5]. Second, it is often observed in conventional lattice-based training that the error rate goes up again if training is not stopped early enough. This can indicate that the optimum of the training criterion is not sufficiently correlated with the error rate, or that the optimization is not numerically stable. Does the convex training criterion define a reasonable optimum? This issue appears interesting particularly if model parameters of different type (e.g. emission and transition parameters) are optimized in the unified framework of log-linear models. We also require that except for the regularization constant, the optimization problem does not involve any additional parameters to be tuned manually. The scaling factor used for lattice-based conventional discriminative training [4] is a typical example of such a parameter. Third, conventional discriminative training is initialized with a reasonable acoustic model, e.g. a GHMM baseline model. Does the convex training criterion estimate the parameters reliably and robustly, independently of the initialization as expected from theory? Finally, the convex training criterion under consideration does not discriminate different word sequences but rather different HMM state sequences, which can represent the same word sequence. Does this mismatch in training and recognition reduce the performance of the convex training criterion? The goal of this work is not so much to find a log-linear model that outperforms conventional GHMMs as e.g. in [3, 2]. We rather focus on the investigation of the utility and feasibility of convex training criteria using log-linear models in speech recognition. The experiments are performed on a simple, yet competitive log-linear model, which allows for a thorough experimental investigation of the above issues.

3. Log-Linear Models & Training Criteria

Here, a simple log-linear CRF is considered. The model includes emission features and transition features, leading to a CRF with HMM structure [4, 3]. Features to represent the language model are not used because the focus is on the recognition of digit strings. For this reason and assuming that the HMM state sequences s_1^T uniquely define the word sequences, the explicit dependence on the word sequence can be dropped. This considerably simplifies the notation. In addition, the model parameters associated with the emission features depend on the gender g. The transition features are shared by the two gender-dependent models. For first order features x_1^T, the posterior probability then reads

$$p_\Lambda(s_1^T, g|x_1^T) = \frac{1}{Z}\exp\left(\sum_{t=1}^{T}\alpha(s_{t-1}, s_t) + \alpha(s_t, g) + \lambda(s_t, g)^\dagger x_t\right) \quad (1)$$

where Z denotes the normalization constant over all HMM state sequences s_1^T. To avoid confusion with the number of frames T, the dagger † is used to denote the transpose of a vector. The set of log-linear parameters to be estimated is $\Lambda = \{\alpha(\sigma, s), \alpha(s, g), \lambda(s, g)\}$. The transition and emission features are associated with the parameters $\alpha(\sigma, s)$ and $\alpha(s, g), \lambda(s, g)$, respectively. This choice of features is motivated by the Gaussian models [6]. More sophisticated emission features (e.g. second order features) can be incorporated by augmenting the first order features. Keep in mind, however, that it is beyond the scope of this paper to find more refined features. The posterior probabilities as defined in Equation (1) lead to the decision rule

$$\hat{s}_1^T = \underset{s_1^T}{\operatorname{argmax}}\left\{\max_g\left\{\sum_{t=1}^{T}\alpha(s_{t-1}, s_t) + \alpha(s_t, g) + \lambda(s_t, g)^\dagger x_t\right\}\right\} \quad (2)$$

Again, the best state sequence $\hat{s}_1^T$ uniquely defines the recognized word sequence.

Before discussing different training criteria to estimate Λ, the problem of estimating gender-dependent models in the discriminative framework is addressed.

3.1. General Issues for Training

The decision rule in Equation (2) requires that the scores of either gender are comparable. For ML, this is not an issue because the optimization problem decouples into the two gender-dependent optimization problems, i.e., the gender-dependent models can be optimized separately. If the models are optimized in the discriminative framework independently, the scores are no longer guaranteed to be comparable. This is because the numerator and denominator (i.e., normalization constant Z) of the posterior in Equation (1) can be multiplied by the same constant factor without changing the posterior [6]. For this reason, the two gender-dependent models need to be jointly optimized, i.e., the normalization in Equation (1) is also over the hypotheses of the competing gender. This is in contrast to our previous work [8] where this issue is not considered critical because the discriminative training was initialized with ML optimized GHMMs. The complexity of the combined training is approximately four times larger than for the isolated training. This increase in complexity arises from the increased amount of training data (factor of two) and from the augmented summation space (another factor of two).

For simplicity, the training criteria below are formulated without regularization. However, I-smoothing (GHMMs) or L2-norm regularization (CRFs) was used for all discriminative systems below.

3.2. Frame-Based MMI

A simple convex training criterion is the frame-based MMI training criterion. This criterion is inspired by the training strategy usually employed for the hybrid approach, e.g. [9]. Assuming that the state alignment is known and kept fixed during training, the resulting training criterion for log-linear models is convex

$$\mathcal{F}(\Lambda) = \sum_g\sum_t \log\left(\frac{\exp(\alpha(s_t, g) + \lambda(s_t, g)^\dagger x_t)}{\sum_{g,s}\exp(\alpha(s, g) + \lambda(s, g)^\dagger x_t)}\right). \quad (3)$$

This criterion discards all context information, i.e., any state sequence is allowed and the transition features do not enter the criterion. Hence, it is not possible to estimate the transition parameters, which need to be tuned manually. The normalization is on frame level over all HMM states. For recognition, the log state prior is subtracted from the parameters $\alpha(\sigma, s, g)$ [9]. Furthermore, it is essential to accumulate the silence frames with lower weight, probably due to the high silence portion. In practice, setting the total silence weight to the average weight of all other states turned out to be a good choice. Our experience is that the scaling factors badly need to be re-tuned after training.

Next, two sentence-based MMI criteria are discussed.

3.3. Lattice-Based MMI

The conventional MMI criterion is based on the posterior in Equation (1)

$$\mathcal{F}(\Lambda) = \sum_r \log \left(\sum_{s_1^{T_r} \in \mathcal{N}_r} p_\Lambda(s_1^{T_r}, g_r | x_1^{T_r}) \right) \qquad (4)$$

where the first sum is over all training sentences, r. The word lattice approximates the set of competing sequences. The numerator lattice $\mathcal{N}$ is defined to be the subset of correct sequences from the word lattice. The posterior is normalized over all hypotheses in the denominator lattice and all g. The lattice-based approach is approximate not only because only a pruned summation space is considered but also because the maximum approximation within the word boundaries is employed. The transition parameters can be estimated in this framework. This, however, was not done in this work but the transition parameters were tuned manually.

Importantly, this lattice-based MMI criterion is not convex. In the next subsection, this training criterion is modified such as to make it convex. Similar approaches can be found in [3, 5].

3.4. MMI (Convex Formulation)

Assuming the maximum approximation and realignment, the convexity of the training criterion in Equation (4) is broken for two reasons: the sum within the logarithm in Equation (4), and the incomplete sum to approximate the normalization constant. As in [3, 5], the first issue is resolved by limiting the set of correct hypotheses to the best state sequence and keeping it fixed during training. The second problem is avoided by using the complete summation space (feasible for simple tasks).

This training criterion was implemented in our transducer-based discriminative framework. A weighted finite-state transducer represents the complete set of valid state sequences, which can be of different length. The arc weights are set to the transition scores. The emission scores are stored in another transducer, having a state for each time frame and having an arc for each state and HMM state. The denominator lattice is then obtained by composition of these two transducers. The margin transducer is treated in the same way, if necessary. The resulting transducer is similar to the network used for transducer-based search. For training, however, it must be made sure that no duplicate hypotheses are contained (log vs. tropical semiring). An essential difference to lattice-based MMI is that the convex formulation of MMI discriminates between HMM state sequences even if they represent the same word sequence.

The transducer-based framework is a convenient tool to quickly implement new algorithms. The implementations are harder to optimize than a naive implementation. The complexity, however, is the same. In our implementation, convex MMI is roughly a factor of five slower than lattice-based MMI. In addition, the experiments suggest that convex MMI is slower in convergence than lattice-based MMI.

3.5. Modified Training Criteria

A margin term can be incorporated into the training criteria above by using the margin-posterior instead of the posterior defined in Equation (1). Compared with the original posterior, the margin-posterior includes some margin term in addition. For the state-based Hamming accuracy, the term $-\rho\delta(s_t, \hat{s}_t)$ is added to the argument of the exponential function in Equation (1), see [8] for further details. Here, the scaling factor ρ is kept fixed, and only the regularization constant is tuned. This modification results in the corresponding modified training criterion: MMI vs. modified MMI etc.

Table 1: Corpus statistics for SieTill.

	SieTill			
	training		test	
	male	female	male	female
Acoustic data [h]	6.0	5.3	6.0	5.3
#running words [k]	21	20	23	20

Table 2: Different MMI-based training criteria for a simple setup (single densities, first order features, transition parameters tuned manually, initialization with corresponding ML optimized GHMM).

Model	Criterion	Convex	WER [%]
GHMM	ML	no	3.8
	Lattice-based mod. MMI	no	2.7
CRF	Frame-based MMI	yes	3.0
	Lattice-based MMI	no	2.9
	Lattice-based mod. MMI	no	2.7
	MMI	yes	3.1
	Modified MMI	**yes**	**2.5**

4. Experimental Results

The presented approach was evaluated on a digit string recognition task. This task allows for a thorough experimental evaluation due to its small size. All discriminative systems were optimized using Rprop [10]. In past experiments, Rprop proved to be a flexible optimization algorithm with good convergence behavior in practice.

4.1. Digit String Recognition Task

The German digit string recognition task SieTill is used for the experiments. The recognition system is based on gender-dependent whole-word HMMs. For each gender, 214 distinct HMM states plus one for silence are used. The vocabulary consists of the 11 German digits (including the pronunciation variant 'zwo'). The observation vectors consist of 12 cepstral features without derivatives. The gender-independent Linear Discriminant Analysis (LDA) is applied to 5 consecutive frames and projects the resulting feature vector to 25 dimensions [8]. The corpus statistics is summarized in Table 1. The silence portion on all data is rather large, ≈ 55%. The ML baseline system uses Gaussian mixtures with globally pooled variances. It serves as baseline system for comparison and initialization of the log-linear models, if not estimated from scratch.

4.2. Preliminary Studies

Preliminary studies were performed on a very simple setup to check several basic issues, e.g. the choice of the convex training criterion. We used single densities with first order features only. The transition parameters were kept fixed, unless otherwise stated. The log-linear model was initialized with the associated GHMM to speed up training. The results are summarized in Table 2. Internal tests to incorporate the margin term into the frame-based training criterion were not successful, i.e., the margin term did not help significantly. In contrast to lattice-based modified MMI, modified MMI does not require a scaling factor for the posteriors. To check the estimation of the transition features, the transition features were estimated from scratch, using the system optimized with modified MMI. The resulting error

Table 3: Full model training from scratch (single densities, first and second order features, transition features), compared with the other training criteria and GHMMs with 64 dns/mix.

Model	Criterion	Convex	WER [%]
GHMM	ML	no	1.8
	Lattice-based mod. MMI	no	1.6
CRF	Frame-based MMI	yes	2.3
	Lattice-based mod. MMI	no	1.8
	Modified MMI	**yes**	**1.8**

rate does not differ significantly from that in Table 2, i.e., the optimization works but the manually tuned values are already pretty close to the optimum.

These preliminary results suggest that convex optimization may help. It is essential to define a suitable training criterion to achieve good results. Here, the convex training criterion defined on sentence level and including a margin term performs best.

4.3. Full Model Training from Scratch

Now, we are in the position to run an experiment meeting the requirements from Section 2. At this end, consider a log-linear model which includes the first and second order features, and the (global) transition features, again in combination with single densities. The initial alignment is generated using a GHMM with a single globally pooled diagonal covariance matrix and 16 densities/mixture. This setup was then used to estimate the log-linear model from scratch, using the same settings for modified MMI as in Section 4.2. The convergence of the convex MMI training criterion is shown in Figure 1. The training criterion and the error rate are well correlated. No significant increase of the error rate is observed, if not stopped early enough. In parallel to the training shown in this figure, a realignment was done after 150 iterations on the model from iteration 150 to refine the acoustic model training. The realignment had only little effect (less than 0.1% WER reduction), i.e., the initial alignment appears to be good enough. Table 3 compares this error rate with the best GHMM (64 densities/mixture, notably having over four times more parameters) and the other training criteria. Here, the ML optimized GHMM baseline with globally pooled variances (i.e., first order features only) served as initialization for frame-based MMI, and MMI refines frame-based MMI in turn. The results in Figure 1 and in Table 3 imply that convex modified MMI defines a reasonable optimum that can be estimated effectively.

5. Conclusions

Convex optimization using log-linear HMMs was investigated for a digit string recognition task. A convex optimization problem was defined that showed good performance and stable convergence at the same time. Assuming a good initial state alignment, this convex training criterion was used to successfully estimate all model parameters from scratch. Similar to SVMs, the regularization constant is the only parameter that needs to be tuned. Our observation is that a carefully optimized but relatively simple setup can achieve good performance, comparable with conventional training criteria and state-of-the-art GHMMs. This might be a good starting point for adding more sophisticated features (e.g. higher order features, posterior features) to refine the acoustic model. Of course, this is only the first step towards convex and parameter-free optimization in speech recognition. More effort needs to be spent on the incorporation of similar ideas into large vocabulary speech recognition.

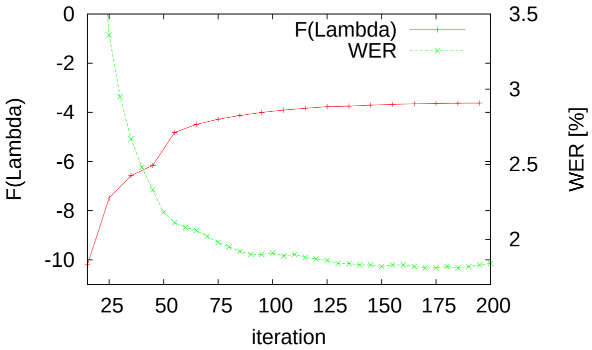

Figure 1: Progress of objective function (training) and Word Error Rate (WER, test) vs. training iteration index for SieTill.

6. Acknowledgements

This work was partly realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation, and was partly funded by the European Union under the integrated project LUNA (FP6-033549) and also partly based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the DARPA.

7. References

[1] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: probabilistic models for segmenting and labeling sequence data," in *ICML*, San Francisco, CA, 2001.

[2] H.-K. J. Kuo and Y. Gao, "Maximum entropy direct models for speech recognition," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 3, 2006.

[3] Y. Hifny Abdel-Haleem, *Conditional random fields for continuous speech recognition*, Ph.D. thesis, Faculty of Engineering, University of Sheffield, 2006.

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

[5] F. Sha and L.K. Saul, "Comparison of large margin training to other discriminative methods for phonetic recognition by hidden Markov models," in *ICASSP*, Honolulu, HI, USA, 2007.

[6] G. Heigold, P. Lehnen, R. Schlüter, and H. Ney, "On the equivalence of Gaussian and log-linear HMMs," in *Interspeech*, Brisbane, Australia, 2008.

[7] Y. Altun, I. Tsochantaridis, and T. Hofmann, "Hidden Markov support vector machines," in *ICML*, Washington, DC, USA, 2003.

[8] G. Heigold, R. Schlüter, and H. Ney, "Modified MPE/MMI in a transducer-based framework," in *ICASSP*, Taipei, Taiwan, 2009.

[9] D. Kershaw, T. Robinson, and M. Hochberg, "Context-dependent classes in a hybrid recurrent network-HMM speech recognition system," in *NIPS*, Denver, CO, USA, 1996.

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