On the Equivalence of Gaussian HMM and Gaussian HMM-like Hidden Conditional Random Fields

Georg Heigold, Ralf Schlüter, and Hermann Ney

RWTH Aachen University
Lehrstuhl für Informatik 6 – Computer Science Department
D-52056 Aachen, Germany
{heigold,schlueter,ney}@cs.rwth-aachen.de

Abstract

In this work we show that Gaussian HMMs (GHMMs) are *equivalent* to GHMM-like Hidden Conditional Random Fields (HCRFs). Hence, improvements of HCRFs over GHMMs found in literature are not due to a refined acoustic modeling but rather come from the more robust formulation of the underlying optimization problem or spurious local optima. Conventional GHMMs are usually estimated with a criterion on segment level whereas hybrid approaches are based on a formulation of the criterion on frame level. In contrast to CRFs, these approaches do not provide scores or do not support more than two classes in a natural way. In this work we analyze these two classes of criteria and propose a refined frame based criterion, which is shown to be an approximation of the associated criterion on segment level. Experimental results concerning these issues are reported for the German digit string recognition task Sietill and the large vocabulary English European Parliament Plenary Sessions (EPPS) task.

Index Terms: speech recognition, parameter estimation, maximum entropy methods

1. Introduction

There is growing interest in non-GHMM-like modeling techniques in Automatic Speech Recognition (ASR) like (H)CRFs, or hybrid approaches using Neural Networks (NNs) or Support Vector Machines (SVMs). All these approaches are discriminative in nature. Discriminative methods have been established in ASR and are an important technique in almost all state-of-the-art systems. The conventional approach in ASR consists of modeling the posterior of a word sequence $w_1^N = (w_1, w_2, \dots)$ given feature vectors $x_1^T = (x_1, x_2, \dots)$ by decomposing the problem into language model $p(w_1^N)$ and acoustic model $p(x_1^T|w_1^N)$

$$p(w_1^N|x_1^T) = \frac{p(w_1^N)p(x_1^T|w_1^N)}{\sum\limits_{v_1^M} p(v_1^M)p(x_1^T|v_1^M)}.$$

The acoustic model uses hidden HMM states s. The state sequences s_1^T obey the Markov property

$$p(x_1^T|w_1^N) = \sum_{s_1^T|w_1^N} \prod_{t=1}^{T} p(x_t|s_t)p(s_t|s_{t-1})$$

where $p(x|s)$ denotes the emission probability and $p(s_t|s_{t-1})$ the transition probability. Usually the emission probability is represented by Gaussian mixtures. Direct models like (H)CRFs try to model the state or word sequence posteriors without the implication of emission probabilities. They have a log-linear functional structure motivated by the Maximum Entropy (ME) principle. Hybrid approaches transform the emission probabilities into HMM state posteriors $p(s|x)$ by Bayes' rule and estimate these quantities. In recognition, the state priors $p(s)$ are required to determine the emission probabilities.

It is well-known in literature that GHMMs and other Gaussian based models can be represented as (H)CRFs [1, 2, 3]. However, it is believed that in general the opposite direction is not possible [2, 3] due to the parameter constraints of Gaussian models, e.g. the normalization of mixture weights or the positivity of variances. In this work we show that these constraints do not restrict the flexibility of log-linear models, i.e., *any* log-linear model can be transformed into an equivalent GHMM.

The parameter estimation of these models requires a training criterion. Here we focus on the Maximum Mutual Information (MMI) criterion both on segment and frame level. Examples for frame based approaches can be found in different flavors. Frame discrimination (FD) based on generative models was proposed by [4], [5] uses a criterion on state level to estimate Maximum Entropy Markov Models (MEMM) that are similar to the more general CRFs [6], and hybrid approaches [7] use NNs or SVMs to model the HMM state posteriors. These approaches assume that the *true* state sequence is known. In practice a time alignment is used. A very attractive property of CRFs (but not HCRFs) is that the associated objective function in parameter estimation has a single global maximum as long as the alignment is kept fixed. An open question is how the criteria on segment and frame level are related, i.e., is there a way to work on frame level without loosing any context information provided by other state-of-the-art criteria? In this work we show that this is possible using time *and* state dependent priors, instead of using only state dependent priors.

Recent publications imply that the robust reestimation of the parameters may be an issue in discriminative training. In particular for systems using density specific variances it has been shown that recognition performance depends on the choice

This material is partly based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001-06-C-0023, and was partly funded by the European Union under the integrated project TC-STAR (FP6-506738). 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.

of optimization technique [3, 8]. As a by-product this work provides some results on this issue.

The structure of the remaining paper is as follows. In Sec. 2 GHMMs and GHMM-like HCRFs are proven to be equivalent. Sec. 3 provides the formulation of training criteria used in this work and an analysis of these criteria, leading to frame based MMI with so-called *context priors*. Finally, Sec. 4 gives comparative experimental results to validate our findings. Our choice of feature functions allows for a direct comparison between GHMMs and GHMM-like HCRFs, cf. Sec. 2.

2. Log-Linear Models

Introductions to (H)CRFs can be found in [1, 2, 3, 5, 6]. Log-linear models always have linear decision boundaries w.r.t. the feature functions. Non-linearities in the original feature space are obtained either by introducing hidden variables [3] or by a suitable choice of the feature functions [1, 5]. Here, the latter approach is pursued and non-linearities are modeled by polynomials of degree n, i.e., the feature functions are n-th order features (e.g. full second order features) which are basically monomials of degree n for a given time and state. For zeroth and first order features, for instance, the state posteriors read

$$p_\Lambda(s|x) \quad = \quad \frac{\exp\left(\sum\limits_{d=1}^{D}\lambda_{s,d}x_d + \alpha_s\right)}{\sum_{s'}\exp\left(\sum\limits_{d=1}^{D}\lambda_{s',d}x_d + \alpha_{s'}\right)}$$

where Λ describes the log-linear parameters. Higher order features can be added similarly. The state prior $p(s)$ is incorporated into α_s for training. For recognition α_s needs to be corrected because the emission probability is the state posterior without prior. Zeroth order features correspond to the terms of the exponential function that do not depend on x, e.g. priors. The first order features correspond to the means. In case of a globally pooled covariance matrix the quadratic terms $x^T\Sigma x$ do not depend on s and thus, cancel. To represent Gaussians with density specific covariance matrices, second order features are necessary.

2.1. Equivalence of GHMMs and GHMM-like HCRFs

As already mentioned the determination of log-linear parameters from Gaussian parameters is straightforward and well-known [1, 3]. For the back transformation the equations can be solved for the Gaussian parameters, which define a proper Gaussian model as long as the constraints are satisfied. Here we show how to impose the constraints on log-linear models by means of invariance transformations.

The model parameters of the Gaussian model are uniquely defined in the Maximum Likelihood (ML) framework. This is no longer valid for log-linear models which are ambiguous in the sense that there are distinct Λ and Λ' such that the resulting discriminative models are identical, i.e., all posteriors are the same. It can be shown that two log-linear models are identical if and only if $(\lambda'_s - \lambda_s)^T x$ is independent of s for all x. In all non-degenerate cases invariance is implied for $\lambda'_s = \lambda_s + \Delta\lambda$, for any $\Delta\lambda \in \mathrm{I\!R}^{\mathrm{D}}$. Below, we will use the invariance transformations $\alpha_s \rightarrow \alpha_s + \Delta\alpha$ ($\Delta\alpha \in \mathrm{I\!R}$) to normalize probabilities like priors $p(s)$ and $\Lambda_s \rightarrow \Lambda_s + \Delta\Lambda$ ($\Delta\Lambda \in \mathrm{I\!R}^{\mathrm{D}\times\mathrm{D}}$) to impose the positivity constraint on the variances. These invariances lead to rather strange and counterintuitive behavior of Gaussian based posteriors [9]. The invariance associated with first order features, for example, implies that the means can be localized anywhere in space.

Table 1: Transformation of log-linear model parameters into (proper) GHMM parameters, $\lambda_s \in \mathrm{I\!R}^{\mathrm{D}}$, $\alpha_s \in \mathrm{I\!R}$, $\Sigma \in \mathrm{I\!R}^{\mathrm{D}\times\mathrm{D}}$

1.	Σ	=	any symmetric, positive-definite matrix
2.	μ_s	=	$\Sigma\lambda_s$
3.	$\tilde{p}(s)$	=	$\exp\left(\alpha_s + \frac{1}{2}\mu_s^T\Sigma^{-1}\mu_s + \frac{1}{2}\log\lvert 2\pi\Sigma\rvert\right)$
4.	$p(s)$	=	$\frac{\tilde{p}(s)}{\sum_{s'}\tilde{p}(s')}$

Next, these invariance transformations of the model are applied to write log-linear models in Gaussian form. For simplicity, we start with log-linear models with zeroth and first order features only (globally pooled variances) and simple priors instead of transition probabilities. First, Σ can be set to any symmetric and positive definite matrix because the second order terms in x do not have any impact on the posteriors. Setting the means does not cause any conceptual problems. Next, the pseudo-probabilities $\tilde{p}(s)$ are initialized from α_s, including corrections like the state independent normalization constant $|2\pi\Sigma|$. These (non-negative) pseudo-probabilities can be normalized because the normalization constant $\sum_{s'}\tilde{p}(s')$ does not depend on s. If the feature dimension is larger than the number of classes, all priors can be set to one, cf. the invariance evoked by first order features. The transformation rules are summarized in Tab. 1. In the case of density specific variances, the first step in Tab. 1 is more intricate because Λ_s is not guaranteed to be symmetric or negative-definite. First, Λ_s is replaced with $\frac{\Lambda_s+\Lambda_s^T}{2}$ to make it symmetric, which is always possible due to the symmetry of the second order features. Subtracting a matrix with sufficiently large eigenvalues leads to a (strictly) negative-definite matrix. By definition the resulting matrix is regular and thus, $\Sigma_s = -\frac{1}{2}\Lambda_s^{-1}$ is well defined. Remember that only invariance transformations were applied in the different steps, so we have constructed GHMMs from log-linear models *without* loosing any flexibility in the model.

In conclusion, GHMMs and GHMM-like (H)CRFs are equivalent and thus, differences in performance come from numerical instabilities or from different local optima due to different optimization schemes [1, 3]. The same strategies can be applied to other HMM-like feature functions, e.g. transition or language probabilities (conditional probabilities are derived from the respective joint probabilities using basic probabilistic rules), and on segment level as well. The result also holds for other posterior based criteria, e.g. MCE or MPE [10].

3. Parameter Estimation

According to the ME principle, the optimal parameters Λ of the log-linear model are obtained by maximizing the posteriors. Depending on the dependence assumptions, the criteria can be defined on different levels, e.g. frame or segment level. Estimation on segment level is based on the objective function

$$\mathcal{F}^{(\mathsf{MMI})}(\Lambda) = \log p_\Lambda(w_1^N|x_1^T)$$

where w_1^N stands for the spoken word sequence. In contrast, the formulation on frame level is based on the state sequence s_1^T representing the spoken word sequence

$$\mathcal{F}^{(\mathsf{frame})}(\Lambda) = \sum_{t=1}^{T}\log p_\Lambda(s_t|x_t). \qquad (1)$$

To compare these two criteria, frame based MMI requires the extension to time dependent state priors and to allow for summation over more than a single state in the calculation of the posteriors in (1). The second step can introduce local optima.

3.1. Context Priors

In this section we show the relation between MMI on sentence and frame level. The derivation does not make any assumptions on the model. θ denotes the parameters to estimate. The MMI criterion on segment level can be written in terms of $p_{\theta,t}(s, w_1^N | x_1^T \backslash x_t)$ which refers to the forward-backward (FB) probability used in discriminative training [10] including the language model score and excluding the emission probability of time frame t (cf. "$x_1^T \backslash x_t$")

$$\mathcal{F}^{(\mathrm{MMI})}(\theta) = \sum_{t=1}^{T} \log \frac{\sum_s p_{\theta,t}(s, w_1^N | x_1^T \backslash x_t) p_\theta(x_t|s)}{\sum_s p_{\theta,t}(s | x_1^T \backslash x_t) p_\theta(x_t|s)}.$$

The FB-like quantity in the denominator is obtained by marginalization of the FB-like quantity in the numerator over all competing word/state sequences. Assuming single densities and strict maximum approximation, the sum in the numerator consists of a single summand, and, thus, does not depend on θ. Next, we employ the approximation that $p_{\theta,t}(s|x_1^T \backslash x_t)$ varies only slowly in θ compared with the other terms, i.e., this quantity can be considered constant in θ. This approximation might be justified by the observation that the denominator term is a (global) average in contrast to the (local) emission probability. For this reason it is expected that this quantity does not require recalculation after each iteration. Utilizing this approximation in the above-mentioned identity, we arrive at frame based MMI with time and state dependent priors proportional to the FB-like quantities. Note that the normalization of the priors does not affect the criterion and is introduced only for aesthetic reasons. Interestingly, the priors contain the complete context information. The essential question is which assumptions on the probabilistic model are made to determine the priors. In the original frame based approach the states are assumed to be independent whereas on the segment level acoustic and lexical context is considered. In case of a single summand in the numerator, this criterion has the same structure as (1). In general the numerator consists of a weighted sum over the correct states. The number of iterations without recomputing the priors is called a *period*, i.e., MMI corresponds to frame based MMI with context priors and period 1. It can be shown that frame based MMI with context priors can be considered a weak auxiliary function of MMI at θ' (identical derivatives at θ') for which the FB probabilities are computed. So the two approaches have the same optimum if the true priors (oracle) are known.

This approximation is faster than MMI on segment level once the priors are calculated (comparable with an MMI iteration). This fact makes this approximation interesting in the context of algorithms which are hard to parallelize. Furthermore, the frame based formulation allows for frame based concepts which are difficult to define on a coarser, say segment level. Similar relations can be derived for the MWE and the MPE criterion, too.

3.2. Optimization

For GHMMs the Extended Baum Welch (EBW) algorithm is usually employed to optimize the parameters. Unfortunately, this algorithm is not suitable for log-linear models because they do not have variance-like parameters which are required to reach reasonably fast convergence [10]. Generalized Iterative Scaling (GIS) is applicable but turned out to be inefficient [1]. General gradient based procedures like RProp, QProp, or more sophisticated versions thereof have proven to be efficient [3, 8]. Our choice was a QProp-like optimization scheme. Note that these algorithms do not find the global optimum in general - at best they provide a local optimum which is an additional difficulty [3]. CRFs have a single (global) maximum and are rather simple to optimize for this reason.

According to [3, 8], the EBW reestimation of GHMMs using density specific variances might be inferior to other optimization techniques. On the one hand, the choice of globally pooled variances (as used in our system) is expected to alleviate this problem. On the other hand, EBW sets the iteration constants such that the variances remain positive [10] although the quadratic terms in x cancel in the sentence posterior probability, and thus, are arbitrary.

4. Experimental Results

Experiments were performed on the German digit string recognition task Sietill and the large vocabulary EPPS English 2006 task. The baseline systems are based on GHMMs with globally pooled variances and HMM states are modeled by single densities. Sietill uses whole-word HMMs and single densities for each HMM state. The vocabulary comprises the German digits. EPPS English has a vocabulary with about 50,000 entries and uses CART tied triphone states modeled by mixtures. All discriminative trainings were initialized with ML models. In fact the systems under consideration are not completely equivalent because the GHMM imposes the constraint that the mixture weights are normalized. Experiments enforcing this constraint, however, have shown that the improvements by this effect are marginal, if any at all. For this reason the reported results were produced without this additional constraint.

First, and diagonal/full second order features are abbreviated by '1', 'd2', and 'f2'. Different n-th order features are combined with '+'. The zeroth order feature is always included, cf. Tab. 2 and 3.

In frame based training the priors were set to the relative occurrences in the training corpus. It turned out that the proper handling of priors, in particular of silence and noise is essential. MMI on frame level tends to converge slower than the other criteria but convergence is smoother. It looks that mixtures allow for a more selective modeling than n-th order features. There is ongoing work with other non-linear feature functions to improve the current results.

So far we have not found any evidence supporting the hypothesis that GHMMs using globally pooled variances and mixtures are not reliably estimated with EBW. Frame based MMI with context priors with period ∞ seems to be a reasonable approximation in this setting, too. Details on these experiments are beyond the scope of this paper, and are considered for a later publication.

4.1. Sietill

The recognition system is based on gender-dependent whole-word HMMs. For each gender 214 distinct states plus one for silence are used. The vocabulary consists of the 11 German digits (including 'zwo'). The observation vectors consist of 12 cepstral features without any derivatives. The gender-independent Linear Discriminant Analysis (LDA) is applied to 5 consecutive frames and projects the resulting feature vector to 25 dimensions. The training corpus consists of 11.3h audio data/42,857 spoken digits with a silence proportion of 55%. The test corpus has 11.4h audio data, corresponding to 43,086 spoken digits. The ML baseline system uses single Gaussians with globally pooled variances and yields 3.8% WER. The MMI system with

Table 2: Word Error Rates (WER) in % for Sietill test corpus, 'Context-2' denotes context priors with period 2

Criterion	Model	Feat.	Param.	WER
ML	Gauss	1	2×5k	3.8
MMI	Gauss	1	2×5k	3.0
		1	2×90k	1.9
	Log-lin.	1	2×5k	2.9
		1+f2	2×75k	2.1
Context-2	Log-lin.	1	2×5k	2.9
Frame	Log-lin.	1	2×5k	3.0
		1+d2	2×11k	2.8
		1+f2	2×75k	2.3

16 densities/state has 2×90k parameters, compared with 2×75k parameters for single densities and full second order features, and yields 1.9% WER. The initial alignment was taken from the baseline. After convergence, a few realignments in turn with reestimation were performed. Realignments reduce the WER by 0.1-0.2%. The results are summarized in Tab. 2. The accumulation of the frame based approach was about 5 times faster than the segment based approach.

4.2. EPPS English

This task contains recordings from the European Parliament Plenary Sessions (EPPS). 87.5h of speech recordings/704,883 running words were manually transcribed, which are used for training of the acoustic models [11]. The non-speech proportion is roughly 30%. The acoustic front end comprises MFCC features augmented by a voicing feature. 9 consecutive frames are concatenated and the resulting vector is projected to 45 dimensions by means of LDA. The MFCC features are warped using a fast variant of the Vocal Tract Length Normalization (VTLN). The triphones are clustered using CART, resulting in 4,501 generalized triphone states. The acoustic models are trained on the complete manually transcribed data. The development and evaluation data from the evaluation campaign 2006 comprise 3.2h/27,029 running words and 3.2h/29,829 running words, respectively. For recognition the vocabulary size is 52,429 and a 4-gram language model is used. The ML baseline achieves 24.7% WER and the MMI trained GHMM system 21.9% WER on the evaluation corpus. The number of parameters of the GHMM with 32 densities per HMM state (6,621k parameters) is comparable with that of the log-linear system using full second order features with 4,866k parameters (see '1+f2' in Tab. 3). See Tab. 3 for further results. The alignment for frame based training was the same as for the estimation of the ML trained model and was not changed during training.

5. Conclusions

This work proves equivalence of GHMMs and GHMM-like HCRFs. This could be substantiated experimentally on Sietill and the EPPS English corpus. From this result we conclude that GHMM parameters can be estimated without numerical stability problems using standard optimization techniques. In addition we have shown that under certain assumptions MMI on frame level can be considered an approximation of MMI on segment level. This considerably speeds up accumulation time. Experiments show that this approximation is valid as long as the models do not change too much, say by relative 10% in WER.

Table 3: Word Error Rates (WER) in % for EPPS English

Criterion	Model	Feat.	Param.	WER	
				Dev	Eval
ML	Gauss	1	207k	28.9	24.7
		1	6,621k	18.9	16.1
MMI	Gauss	1	207k	24.7	21.9
Frame	Log-lin.	1	207k	26.1	22.0
		1+d2	410k	24.9	20.5
		1+f2	4,866k	20.8	16.8

6. References

[1] W. Macherey and H. Ney, "A comparative study on maximum entropy and discriminative training for acoustic modeling in automatic speech recognition," in *Proc. European Conference on Speech Communication and Technology*, Geneva, Switzerland, Sept. 2003, pp. 493 – 496.

[2] C. Sutton and A. McCallum, "An introduction to conditional random fields for relational learning," in *Introduction to Statistical Relational Learning*. MIT Press, 2006.

[3] A. Gunawardana, M. Mahajan, A. Acero, and J. C. Platt, "Hidden conditional random fields for phone classification," in *Proc. of the Int. Conf. on Spoken Language Processing (ICSLP)*, Lisbon, Portugal, Sept. 2005.

[4] D. Povey and P.C. Woodland, "Frame discrimination training for HMMs for large vocabulary speech recognition," in *Proc. of the Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Phoenix, AZ, Mar. 1999, pp. 333 – 336.

[5] 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, pp. 873 – 881, 2006.

[6] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: probabilistic models for segmenting and labeling sequence data," in *Proc. 18th Int. Cong. Machine Learning*, San Francisco, CA, 2001, pp. 282 – 289.

[7] D. Kershaw, T. Robinson, and M. Hochberg, "Context-dependent classes in a hybrid recurrent network-HMM speech recognition system," in *Advances Neural Information Processing Systems 8 (NIPS'8)*, 1996, pp. 750 – 756.

[8] E. McDermott, T.J. Hazen, J. Le Roux, A. Nakamura, and S. Katagiri, "Discriminative training for large vocabulary speech recognition using Minimum Classification Error," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 203 – 223, 2007.

[9] E. S. Ristad and P. N. Yianilos, "Towards EM-style algorithms for *a posteriori* optimization of normal mixtures," in *Proc. IEEE Symposium on Information Theory*, August 1998.

[10] W. Macherey, R. Schlüter, and H. Ney, "Discriminative training with tied covariance matrices," in *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, Jeju Island, Korea, Oct. 2004, pp. 681 – 684.

[11] J. Lööf, M. Bisani, Ch. Gollan, G. Heigold, B. Hoffmeister, Ch. Plahl, R. Schlüter, and H. Ney, "The 2006 RWTH parliamentary speeches transcription system," in *Proc. of the Int. Conf. on Spoken Language Processing (ICSLP)*, Pittsburgh, PA, Sept. 2006, pp. 105 – 108.