A GIS-LIKE TRAINING ALGORITHM FOR LOG-LINEAR MODELS WITH HIDDEN VARIABLES

Georg Heigold, Thomas Deselaers, Ralf Schlüter, and Hermann Ney

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

ABSTRACT

Conditional random fields (CRFs) are often estimated using an entropy based criterion in combination with Generalized Iterative Scaling (GIS). GIS offers, upon others, the immediate advantages that it is locally convergent, completely parameter free, and guarantees an improvement of the criterion in each step. GIS, however, is limited in two aspects. GIS cannot be applied when the model incorporates hidden variables, and it can only be applied to optimize the Maxmimum Mutual Information Criterion (MMI). Here, we extend the GIS algorithm to resolve these two limitations. The new approach allows for training log-linear models with hidden variables and optimizes discriminative training criteria different from Maximum Mutual Information (MMI), including Minimum Phone Error (MPE). The proposed GIS-like method shares the above-mentioned theoretical properties of GIS. The framework is tested for optical character recognition on the USPS task, and for speech recognition on the Sietill task for continuous digit string recognition.

Index Terms— speech recognition, parameter estimation, maximum entropy, GIS, optical character recognition

1. INTRODUCTION

Log-linear models have become an important technique in various fields of pattern recognition. They appear in different terminologies and flavors, e.g., maximum entropy (Markov) models, logistic regression, and CRFs. The only input of such models are feature functions $f_i(x, c)$, which map the observation vector x and a class $c \in \{1, \ldots, C\}$ to abstract features. Given the feature functions, the log-linear functional structure is motivated by the maximum entropy principle

$$p_\Lambda(c|x) = \frac{\exp\left(\sum_{i=1}^{D} \lambda_i f_i(x, c)\right)}{\sum_{c'} \exp\left(\sum_{i=1}^{D} \lambda_i f_i(c', x)\right)} \quad (1)$$

where D denotes the number of features functions. The parameters $\Lambda = \{\lambda_1, \ldots\}$ are typically determined by maximizing the empirical entropy on the training data $\{(x_1, c_1), \ldots, (x_N, c_N)\}$ with observation vectors x_n and class labels c_n.

$$\mathcal{F}(\Lambda) = \sum_{n=1}^{N} \log p_\Lambda(c_n|x_n). \quad (2)$$

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.

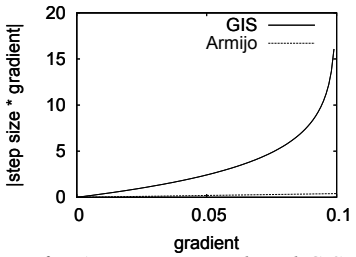

Fig. 1. *Step sizes for Armijo's approach and GIS; N=0.1, F=138, D=513 (typical values for USPS)*

The optimization is often done using GIS [1]. Many problems of practical interest like for example hidden CRFs (HCRFs) [2], however, are extensions to this purely log-linear formulation, involving hidden variables which cannot be optimized with standard GIS.

From the theoretical point of view, GIS is attractive because this algorithm does not only guarantee to converge to a critical point (the global optimum in the special case of Eq. (2)) but also guarantees to increase the objective function in each iteration. This is in contrast to general gradient based procedures for which convergence to a critical point can be proven at best, e.g. Newton method or RProp [3].

Like for example Expectation Maximization (EM) [4], GIS is based on the concept of growth transformations. Probably the most general and simplest growth transformation goes back to Armijo [5]. He showed that for objective functions with Lipschitz continuous first derivatives, global and non-vanishing step sizes exist. Examples for such functions are log-linear models or Gaussian models with floored variances. However, Armijo's step sizes turn out to be rather pessimistic compared to GIS, see Fig. 1. Other approaches decompose the problem into simpler subproblems, i.e., the overall optimization is performed by alternating the simplified problems. Typical examples for this approach are the GEM and the extension of GIS proposed in [6]. Here, we avoid such indirections and directly optimize the objective function using a single auxiliary function.

Finally, there are growth transformations to estimate generative models discriminatively, e.g., the inequality for rational functions [7, 8, 9], or the reverse Jensen inequality [10, 11]. Applying these inequalities, however, leads to purely linear growth transformations which might be problematic. The use of regularization terms, for example, avoids this problem, but then other problems occur (e.g., the reverse Jensen inequality requires non-vanishing second derivatives of the argument of the exponential). It was shown in [12] that some special cases of log-linear models can be represented by an equivalent generative model such that these growth transformations can be applied. However, because the model parameters (and thus, the iteration constants) are ambiguous [12], the optimization process heavily relies on the initial choice of the parameters. In

addition, to efficiently calculate the iteration constants for complex problems (e.g. Gaussian HMMs), typically several approximations are made [11].

This paper is organized as follows. Sec. 2 generalizes the training criterion in Eq. (2) to hidden variables, e.g. HCRFs or MPE. This generalized objective function can be optimized with an extended GIS-like algorithm as detailed in Sec. 3. Various applications of this criterion are discussed in Sec. 4. Finally, Sec. 5 provides experimental results and the paper is concluded in Sec. 6.

2. GENERALIZED OBJECTIVE FUNCTION

Many important problems (e.g. HCRFs) do not match the simple objective function in Eq. (2). They often involve hidden variables in some sense, requiring a more general formulation of the training criterion. Using prior-like and sample dependent weights $q_n(c), p_n(c) \geq 0$ (not necessarily normalized), the extension

$$\mathcal{F}^{(hidden)}(\Lambda) \quad = \quad \sum_n \log\left(\frac{\sum_c q_n(c)\exp\left(\sum_i \lambda_i f_i(x_n,c)\right)}{\sum_c p_n(c)\exp\left(\sum_i \lambda_i f_i(x_n,c)\right)}\right) \qquad (3)$$

shall be considered.In the same way as Eq. (2), this objective function is maximized. The major difference between Eq. (2) and Eq. (3) is the (weighted) sum over the classes in the numerator. Eq. (3) reduces to Eq. (2) for $q_n(c) = \delta_{cc_n}$ and $p_n(c) = 1$, in which case the sum in the numerator consists of a single summand and GIS can be applied. For more complex examples beyond this simple special case, the reader is referred to Sec. 4.

In the next section we propose an auxiliary function for this generalized criterion. For this purpose, it is convenient to rewrite the criterion as the sum of two objective functions $\mathcal{F}^{(hidden)}(\Lambda) = \mathcal{F}^{(num)}(\Lambda) - \mathcal{F}^{(den)}(\Lambda)$ with

$$\mathcal{F}^{(num)}(\Lambda) \quad = \quad \sum_n \log\left(\sum_c q_n(c)\exp\left(\sum_i \lambda_i f_i(x_n,c)\right)\right). \qquad (4)$$

The objective function $\mathcal{F}^{(den)}(\Lambda)$ is obtained from Eq. (4) by replacing $q_n(c)$ with $p_n(c)$.

3. GIS-LIKE OPTIMIZATION

In this section, we derive an auxiliary function for the generalized objective function introduced in the last section. The idea consists of building an auxiliary function by decomposing the problem into well-known subproblems and then combining these partial auxiliary functions to a complete auxiliary function of $\mathcal{F}^{(hidden)}$. First, we give a summary on the notion of auxiliary functions. This concept is then used to construct an auxiliary function for the desired objective function.

3.1. Basic Definitions and Concept

The following lemmata are based on the concept of auxiliary functions. According to [13], they are defined based on the old (given) and new (to estimate) parameters Λ' and Λ, respectively.

Definition. *Assume an objective function $\mathcal{F}(\Lambda)$ to be maximized. An* auxiliary function $A(\Lambda|\Lambda')$ *(in the strong sense) of $\mathcal{F}(\Lambda)$ at Λ' satisfies the inequality $\mathcal{F}(\Lambda) - \mathcal{F}(\Lambda') \geq \mathcal{A}(\Lambda|\Lambda')$ with equality for $\Lambda = \Lambda'$.*

The aim of an auxiliary function is to replace the original optimization problem by a related optimization problem which can be tackled more easily. In the ideal case an analytical solution exists, e.g. GIS. Typically, auxiliary functions decouple the parameters such that the optimization for the parameters are independent. Under rather mild assumptions on the objective and auxiliary functions (e.g. sufficiently smooth and bounded above), it can be shown that the sequence over the iteration index k $\{\Lambda_k|\Lambda_{k+1} := \operatorname{argmax}_\Lambda \mathcal{A}(\Lambda|\Lambda_k)\}$ converges to a critical point of the associated objective function, i.e., the gradient of the objective function vanishes at Λ_∞. The proof consists of two steps: the existence of a stationary point and the determination of possible stationary points. The existence follows from the monotonicity and boundedness of the sequence. The stationary points are critical points of the objective function because it can be shown that $\nabla\mathcal{F}(\Lambda') = \nabla\mathcal{A}(\Lambda'|\Lambda')$. The next lemma is rather simple but useful because it allows us to build auxiliary functions by combining partial auxiliary functions.

Lemma (Concatenation). *Let $\mathcal{B}$ be an auxiliary function of $\mathcal{F}$ at Λ' and let $\mathcal{A}$ be an auxiliary function of $\mathcal{B}$ at Λ'. Then $\mathcal{A}$ is also an auxiliary function of $\mathcal{F}$ at Λ'.*

All bounds in the next subsection are based on the well-known Jensen's inequality. The results are stated in terms of generalized numerator posteriors

$$q_\Lambda(c|x_n) \quad = \quad \frac{q_n(c)\exp\left(\sum_i \lambda_i f_i(x_n,c)\right)}{\sum_{c'} q_n(c')\exp\left(\sum_i \lambda_i f_i(x_n,c')\right)} \qquad (5)$$

and generalized denominator posteriors $p_\Lambda(c|x_n)$, which are defined analogously. In addition, we shall use the shortcut $\Delta\lambda_i$ to denote the difference of old and new parameters, $\lambda_i - \lambda_i'$.

3.2. Auxiliary Function

Assuming that the objective function in Eq. (3) is bounded above, it is sufficient to find a suitable auxiliary function as defined above. Then, the following lemma concerning $\mathcal{F}^{(num)}$ in Eq. (4) is valid:

Lemma (EM).

$$\mathcal{A}^{(EM)}(\Lambda|\Lambda') \quad = \quad \sum_n \sum_c q_{\Lambda'}(c|x_n) \sum_i \Delta\lambda_i f_i(x_n,c)$$

is an auxiliary function of $\mathcal{F}^{(num)}$ at Λ'.

Proof. Basically, the same inequality as for EM [4] is used:

$$\begin{aligned}
&\mathcal{F}^{(num)}(\Lambda) - \mathcal{F}^{(num)}(\Lambda') \\
&\overset{(4),(5)}{=} \quad \sum_n \log\left(\sum_c q_{\Lambda'}(c|x_n)\exp\left(\sum_i \Delta\lambda_i f_i(x_n,c)\right)\right) \\
&\overset{\text{Jensen}}{\geq} \quad \sum_n \sum_c q_{\Lambda'}(c|x_n) \sum_i \Delta\lambda_i f_i(x_n,c) \\
&=: \quad \mathcal{A}^{(EM)}(\Lambda|\Lambda').
\end{aligned}$$

Equality holds for $\Lambda = \Lambda'$. □

The next lemma requires non-negative feature functions. This, however, is not a restriction because negative feature functions can be transformed by a suitable componentwise affine transformation to satisfy these constraints without changing the posteriors.

Lemma (GIS). *Suppose that $f_i(x_n,c) \geq 0$ $(\forall i,n,c)$, and that $\sum_i f_i(x_n,c) \equiv F$ $(\forall n,c)$. Then*

$$\mathcal{A}^{(GIS)}(\Lambda|\Lambda') \quad = \quad N - \sum_n \sum_c p_{\Lambda'}(c|x_n) \sum_i \frac{f_i(x_n,c)}{F}\exp(F\Delta\lambda_i)$$

is an auxiliary function of $-\mathcal{F}^{(den)}$ at Λ'.

Proof. Basically, the same inequalities as for GIS [1] are used:

$$
\begin{aligned}
&-\left(\mathcal{F}^{(den)}(\Lambda)-\mathcal{F}^{(den)}(\Lambda')\right)\\
&\overset{(4),(5)\text{ for }p_n(c)}{=} -\sum_n \log\left(\sum_c p_{\Lambda'}(c|x_n)\exp\left(\sum_i \Delta\lambda_i f_i(x_n,c)\right)\right)\\
&\overset{\log x\le x-1}{\ge} N-\sum_n\sum_c p_{\Lambda'}(c|x_n)\exp\left(\sum_i F\Delta\lambda_i\frac{f_i(x_n,c)}{F}\right)\\
&\overset{\text{Jensen}}{\ge} N-\sum_n\sum_c p_{\Lambda'}(c|x_n)\sum_i\frac{f_i(x_n,c)}{F}\exp\left(F\Delta\lambda_i\right)\\
&=: \mathcal{A}^{(GIS)}(\Lambda|\Lambda').
\end{aligned}
$$

Equality holds for $\Lambda=\Lambda'$. □

So far, we have built two separate auxiliary functions for the numerator and denominator. Applying Lemma (Concatenation), we obtain the complete auxiliary function as desired.

Corollary. *Let $f_i(x,c)$ be feature functions subject to the constraints in Lemma (GIS). Then $\mathcal{A}^{(hidden)}=\mathcal{A}^{(EM)}+\mathcal{A}^{(GIS)}$ is an auxiliary function of $\mathcal{F}^{(hidden)}$ at Λ'.*

Proof. Set $\mathcal{F}=\mathcal{F}^{(num)}+\mathcal{F}^{(den)}$, $\mathcal{B}=\mathcal{A}^{(EM)}+\mathcal{F}^{(den)}$, and $\mathcal{A}=\mathcal{A}^{(EM)}+\mathcal{A}^{(GIS)}$ in Lemma (Concatenation). □

Notice that extensions like Improved Iterative Scaling (IIS) are compatible with this extension. It is also possible to incorporate a regularization term based on the p-norm into these formulae.

Setting the first derivatives of $\mathcal{A}^{(hidden)}(\Lambda|\Lambda')=0$ and solving these equations for λ_i provides the (unique) solution $\Delta\lambda_i=\frac{1}{F}\log\left(\frac{N_i(\Lambda')}{Q_i(\Lambda')}\right)$ with slightly modified numerator statistics

$$N_i(\Lambda') \quad = \quad \sum_n\sum_c q_{\Lambda'}(c|x_n)f_i(x_n,c).$$

The denominator statistics $Q_i(\Lambda')$ are defined in the same fashion. The equations have the same structure as for standard GIS except that now, N_i can depend on Λ' and in general, $q_n(c)$ and $p_n(c)$ are not true posteriors. Using n-th order features the accumulation statistics simplify greatly (similarly for $Q_c(\Lambda')$)

$$N_c(\Lambda') \quad = \quad \sum_n q_{\Lambda'}(c|x_n)x_n.$$

In ASR, these quantities can be calculated efficiently by generalized forward/backward (FB) probabilities, using a suitable semiring depending on how the weights $q_n(c)$ and $p_n(c)$ are defined. As an example, the common FB probabilities [13] for word sequence v_1^M and HMM state s at frame t, given the feature vectors x_1^T

$$p_t(s,v_1^M|x_1^T)=p(v_1^M)\sum_{s_1^T:s_t=s}\prod_{\tau=1}^{T}p(s_\tau|s_{\tau-1},v_1^M)p(x_\tau|s_\tau,v_1^M) \qquad (6)$$

are associated with the probability semiring. Note that in this paper these are used to calculate the context priors, see next section.

4. APPLICATIONS

There are several examples of practical interest which can be reduced to the generalized objective function, cf. Eq. (3).

Mixtures [2]: Given a mixture s with densities l, the class c stands for the index pair (s,l) and n corresponds to the observation number. By means of the numerator weights $q_n(s,l)=\delta_{l\in s}$, the required densities for a specific mixture are filtered out. The denominator weight $p_n(s,l)$ is set to one for all densities. The mixture weights are represented by a feature function, resulting in a unified treatment of the parameters. This avoids the indirection proposed in [6].

HMMs [2]: HMMs are an extension to the simple mixture models. In this case, the features are defined on segment rather than on frame level. The algorithm copes with additional scaling factors (e.g. language model scale) which can be absorbed by the log-linear parameters and thus, do not change the log-linear model structure.

Context Priors [12]: In hybrid approaches, the state posteriors are estimated with a suitable static classifier, e.g. SVMs or NNs. Here, we employ a log-linear model to represent the state posteriors and estimate the parameters by maximizing the entropy on frame level. This approach has the disadvantage that it relies on a single state sequence to represent the correct word sequence. The frame based MMI criterion using context priors [12] offers a principled way to smooth over competing states. The frame dependent context priors $p_t(s,v_1^M)$ are defined to be the FB probabilities from Eq. (6) without the emission score $p(x_t|s_t,v_1^M)$ of the frame under consideration

$$\mathcal{F}^{(\text{frame})}(\Lambda)=\sum_{t=1}^{T}\log\left(\frac{\sum_s p_t(s,w_1^N)p_\Lambda(x_t|s,w_1^N)}{\sum_{v_1^M}\sum_s p_t(s,v_1^M)p_\Lambda(x_t|s,v_1^M)}\right).$$

The sum in the numerator is basically over all possible states given the word sequence w_1^N and thus, we have a non-trivial sum in the numerator even in the case of single densities. Setting the denominator weights to the context priors $p_t(s,v_1^M)$ and the numerator weights to the context priors for the correct word sequence and to zero otherwise, leads to the generalized criterion in Eq. (3).

Risk Based Criteria [13]: MPE and similar risk based criteria maximize the expectation of a predefined accuracy function, e.g. the phone accuracy $A[v_1^M|w_1^N]\ge 0$ of word sequence v_1^M given word sequence w_1^N [13]. In this case n is obsolete because the criterion is defined on word sequences over the complete corpus and not only over single segments, cf. c. Then, such criteria conform with Eq. (3) for $p_1(v_1^M)=1$ and $q_1(v_1^M)=A[v_1^M|w_1^N]$. Remember that the feature count F (see above) can be calculated on segment level because the denominator is the same as for MMI. Finally, it can be shown that 1-best MCE is an instance of Eq. (3) as well.

5. EXPERIMENTAL RESULTS

The proposed algorithm ('hiddenGIS') is applied to mixtures on the well-known United States Postal Service (USPS) database containing handwritten digits and to context priors on the German digit string recognition task Sietill. Both these applications go beyond standard GIS because of the densities (USPS) or the HMM states (Sietill), see Sec. 4 for more details.

5.1. USPS

The well-known USPS Handwritten Digit Database consists of isolated and normalized images of handwritten digits taken from US mail envelopes scaled to 16 x 16 pixels. The database contains a separate training and test set, with 7,291 and 2,007 images, respectively[1]. One disadvantage of the USPS corpus is that no development test set exists, resulting in the possible underestimation of error rates for all of the reported results. Note that this disadvantage holds for almost all data sets available for image object recognition. The US Postal Service task is still one of the most widely used reference data sets for handwritten character recognition and allows fast experiments due to its small size. The test set contains a large amount

[1] Data available from `ftp://ftp.kyb.tuebingen.mpg.de/pub/bs`

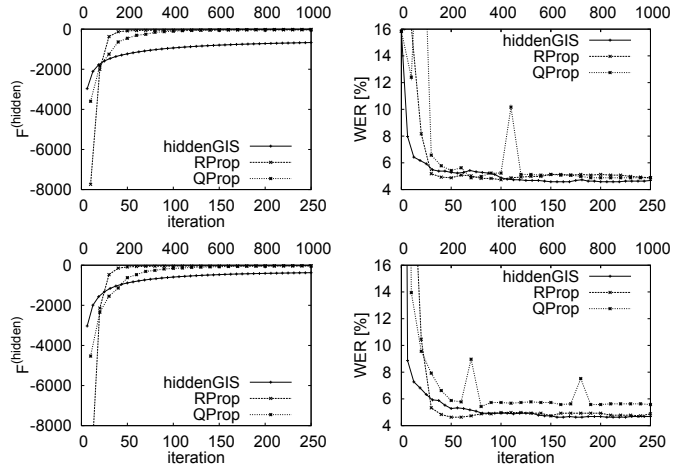

Fig. 2. *USPS, mixtures, different optimization schemes (hiddenGIS, RProp, QProp) and initializations (upper: GMM, lower: random). Left: evolution of $\mathcal{F}^{(hidden)}$ on training corpus. Right: evolution of word error rate WER [%] on test corpus. Note the different scaling of the x axis for hiddenGIS and QProp/RProp.*

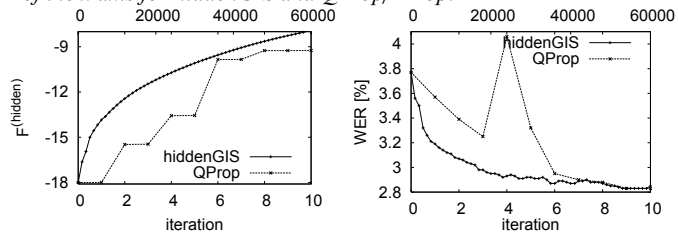

Fig. 3. *Sietill, context priors (see text for explanation), period=2 (QProp) and 250 (hiddenGIS). Left: evolution of $\mathcal{F}^{(hidden)}$ on male training corpus. Right: evolution of word error rate WER [%] on male test corpus. Note the different scaling of the x axis for hiddenGIS and QProp.*

of image variability and is considered to be a "hard" recognition task. Good error rates are in the range of 2-3% and use advanced modeling techniques, e.g. deformation models [14]. Here, we use simple Gaussian mixture models (GMMs) with 16 densities/mixture in combination with the gray-scale features augmented with Sobel based derivatives, amounting to a total of 512 features. Regularization based on a Gaussian prior was used for a smoother convergence behavior. Comparative results are shown in Fig. 2.

5.2. Sietill

The recognition system is based on whole-word HMMs with 214 distinct states plus one for silence. The vocabulary consists of the 11 German digits (including 'zwo'). The observation vectors consist of 12 cepstral features without any derivatives. The Linear Discriminant Analysis (LDA) is applied to 5 consecutive frames and projects the resulting feature vector to 25 dimensions. Both training and test corpus consist of about 5.5h audio data/21k spoken digits. The ML baseline system uses single Gaussians with globally pooled variances and is the initialization of the HCRF for further training. The HCRF is optimized with the frame based MMI criterion using context priors. We have a non-trivial sum in the numerator due to the HMM states. This effect is particularly pronounced at the word boundaries, see Sec. 4 for more details. We compare the proposed algorithm with QProp, see Fig. 3.

6. CONCLUSIONS

We proposed an extension of the well-known GIS algorithm to hidden variables. It does not only apply to the MMI estimation of HCRFs but it also includes more refined criteria, e.g., MPE in ASR. Hence, this generalized GIS can be considered the analog for log-linear discriminative models of EM used for generative models. First results suggest that the convergence is reasonably fast for bounded feature functions (USPS) whereas it is rather slow in the case of basically unbounded feature functions (e.g. MFCC for Sietill).

7. REFERENCES

[1] J. Darroch and D. Ratcliff, "Generalized iterative scaling for log-linear models," *Ann. Math. Statist.*, vol. 43, 1972.

[2] 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.

[3] A.D. Anastasiadis, G.D. Magoulas, and M.N. Vrahatis, "New globally convergent training scheme based on the resilient propagation algorithm," *Neurocomputing*, vol. 64, 2005.

[4] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. B, 1977.

[5] L. Armijo, "Minimization of functions having Lipschitz continuous first derivatives," *Pacific Journal of Mathematics*, vol. 16, no. 1-3, 1966.

[6] L. Saul and D. Lee, "Multiplicative updates for classification by mixture models," in *Advances in Neural and Information Processing Systems*, T.G. Dietterich, S. Becker, and Z. Ghahramani, Ed. MIT Press, 2002.

[7] P.S. Gopalakrishnan, D. Kanevsky, A. Nadas, and D. Nahamoo, "An inequality for rational functions with applications to some statistical estimation problems," *IEEE Trans. on Information Theory*, vol. 37, no. 1, 1991.

[8] A. Gunawardana, "Maximum mutual information estimation of acoustic HMM emission densities," CLSP Research Note No. 40, Johns Hopkins University, Baltimore, MD, 2001.

[9] D. Kanevsky, "Extended Baum Welch transformations for general functions," in *Proc. of the Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Montreal, Canada, 2004.

[10] T. Jebara, *Discriminative, generative, and imitative learning*, Ph.D. thesis, Massachusetts Institute of Technology, 2002.

[11] M. Afify, "Extended Baum-Welch reestimation of Gaussian mixture models based on reverse Jensen inequality," in *Proc. of the Int. Conf. on Spoken Language Processing (ICSLP)*, Lisbon, Portugal, 2005.

[12] G. Heigold, R. Schlüter, and H. Ney, "On the equivalence of Gaussian HMM and Gaussian HMM-like hidden conditional random fields," in *Proc. of the Int. Conf. on Spoken Language Processing (ICSLP)*, Antwerp, Belgium, Aug. 2007.

[13] D. Povey, *Discriminative Training for Large Vocabulary Speech Recognition*, Ph.D. thesis, Cambridge, England, 2004.

[14] D. Keysers, T. Deselaers, Ch. Gollan, and H. Ney, "Deformation models for image recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 8, 2007.