

A Comparative Study on Maximum Entropy and Discriminative Training for Acoustic Modeling in Automatic Speech Recognition

Wolfgang Macherey and Hermann Ney

Lehrstuhl für Informatik VI, Computer Science Department
RWTH Aachen – University of Technology, 52056 Aachen, Germany
{w.macherey, ney}@informatik.rwth-aachen.de

Abstract

While Maximum Entropy (ME) based learning procedures have been successfully applied to text based natural language processing, there are only little investigations on using ME for acoustic modeling in automatic speech recognition. In this paper we show that the well known Generalized Iterative Scaling (GIS) algorithm can be used as an alternative method to discriminatively train the parameters of a speech recognizer that is based on Gaussian densities. The approach is compared with both a conventional maximum likelihood training and a discriminative training based on the Extended Baum algorithm. Experimental results are reported on a connected digit string recognition task.

1. Introduction

Over the past years, maximum entropy (ME) based training methods have been successfully applied to the field of natural language processing, e.g. to language modeling [1], part of speech tagging, language understanding [2], and statistical machine translation [3]. Unlike natural language processing, ME based methods have hardly been investigated for automatic speech recognition. In [4], ME is employed to estimate the parameters of a direct model for a phoneme recognizer. The approach generalizes the ME Markov Models (MEMM) proposed by [5] such that sequential processes with complex contextual information can be processed.

The reason why ME is hardly ever used in acoustic modeling is that, in general the optimization problem is difficult to manage if hidden variables occur. Hidden variables may lead to non-convex objective functions and hence, the most common training algorithms, the Generalized Iterative Scaling (GIS) as well as its faster version, the Improved Iterative Scaling (IIS) cannot be applied. However, in automatic speech recognition, there are typically at least two types of events, which are not observable and thus have to be described via hidden variables: the alignment of a sequence of acoustic observations with the states of a Markov chain and the component weights of a mixture of densities. Moreover, as ME tries to optimize the class posterior probabilities, the objective function corresponds with the Maximum Mutual Information (MMI) criterion, and in ASR, there is already an effective optimization method (namely the Extended Baum (EB) algorithm) that allows for efficiently training free model parameters.

Nevertheless, as we will show in this paper, ME can still be used to discriminatively train the acoustic parameters of a speech recognizer that is based on Gaussian densities by *relaxing* some constraints of the GIS algorithm. In contrast to [4], the used features are not rank-based lists of Gaussian model

indices, but cepstral coefficients as they are provided by the signal analysis part of most common speech recognizers. Since the framework of ME takes competing classes into account, we will not only compare the achieved performance with the word error rate obtained by a maximum likelihood (ML) trained system, but also with a discriminatively trained system that is based on the MMI criterion. Experimental results are reported on a speech corpus for the recognition of telephone line recorded German connected digit strings.

The remainder of this paper is organized as follows. In Section 2 we briefly describe the basics of ME together with the GIS algorithm and show how the acoustic models can be embedded into a log-linear framework. In Section 3 we derive the update rules for the GIS algorithm and compare the resulting re-estimation formulae with those obtained from a “conventional” discriminative approach based on the EB algorithm. Experimental results and a discussion are presented in Section 4. The paper concludes with a summary and an outlook in Section 5.

2. Maximum Entropy Modeling

Given a set of training samples, the principle of ME is to choose a distribution such that it is consistent with constraints derived from the training data while making as little assumptions as possible. It can be shown that the resulting distribution is well defined and leads to a log-linear model [6]. Usually, the constraints are formulated with so-called *feature functions*. For the following presentation, it is convenient to use the following definitions of features. A feature function $f_{b,c'}(x, c)$ shall return a value $\alpha_c > 0$ iff the predicted class c' corresponds with the actual class c and the observation x satisfies the condition b :

$$f_{b,c'}(x, c) := \begin{cases} \alpha_c > 0 & \text{if } c = c' \wedge b(x) = \text{true} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

A feature is said to be *activated* if the returned value is larger than 0; otherwise it is called *inactive*. The solution of the ME approach has a log-linear or exponential form [6]:

$$p_{\Lambda}(c|x) = \frac{1}{Z(x)} \exp \left[\sum_i \lambda_i f_i(x, c) \right] \quad (2)$$

with $\Lambda = \{\lambda_i\}$ denoting the set of free parameters and $Z(x)$ as the normalization term given by

$$Z(x) = \sum_{c'} \exp \left[\sum_i \lambda_i f_i(x, c') \right] \quad (3)$$

The optimal parameter setting Λ can be estimated within a maximum likelihood framework. Given a sequence of labeled

training samples (x_n, c_n) , the objective function of the ME criterion is given by

$$G(\Lambda) := \sum_{n=1}^N \log p_\Lambda(c_n|x_n) = \sum_{x,c} N(x, c) \log p_\Lambda(c|x) \quad (4)$$

where $N(x, c)$ is the number of occurrences of the pair (x, c) in the training corpus. The parameter setting Λ that maximizes Eq. 4 can be obtained by derivating G wrt. λ_i :

$$\frac{\partial G}{\partial \lambda_i} = N_i - Q_i(\Lambda) \stackrel{!}{=} 0 \quad (5)$$

where

$$N_i := \sum_{x,c} N(x, c) \cdot f_i(x, c) \quad (6)$$

$$Q_i(\Lambda) := \sum_x N(x) \sum_c p_\Lambda(c|x) \cdot f_i(x, c). \quad (7)$$

Note that the counts N_i are not necessarily integer values as α_c in Eq. 1 might be any positive real number. Since $G(\Lambda)$ is a sum of convex functions it is also convex and there exists exactly one global maximum that can effectively be determined with the GIS algorithm. For each iteration, the GIS algorithm requires a constant number of active features which can always be fulfilled by adding a *correction feature* f_0 :

$$f_0(x, c) := F - \sum_i f_i(x, c), \quad F := \max_{x,c} \sum_i f_i(x, c) \quad (8)$$

F has to be determined beforehand on the training corpus. According to [7], the parameter update $\Delta \lambda_i$ for each feature i results from solving the equation

$$Q_i(\Lambda) \cdot \exp[\Delta \lambda_i F] = N_i \quad (9)$$

which finally leads to the GIS algorithm as shown in Table 1.

2.1. ME in Automatic Speech Recognition

Let $r = 1, \dots, R$ denote a sequence of training utterances. Each utterance shall be given by a sequence of acoustic observations $X_r = x_{r1}, \dots, x_{rT_r}$ together with a reference transcription $W_r = w_{r1}, \dots, w_{rN_r}$. As most speech recognizers employ cepstral features, we will define the feature functions $f_{s',d}$ as a set of mappings that project a D -dimensional observation vector $x \in \mathbb{R}^D$ onto its d th component iff the predicted Markov state s' corresponds with the state s the observation is aligned

with. Moreover, to ensure that the features are positive and sum up to one they are affinely transformed with a diagonal scaling matrix A and a bias b such that

$$y_t := A^\top \cdot x_t + b \quad \text{with } y_{td} > 0 \quad \text{and} \quad \sum_{r=1}^R \sum_{t=1}^{T_r} y_t = \mathbf{1} \quad (10)$$

holds for each training sample x_t . Thus, we can define the feature functions by

$$f_{s',d}(x, s) := \begin{cases} y_d & \text{if } s = s' \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

Note that affine transformations can always be applied to arbitrary, finite sets of training samples to force the positivity and the normalization of the data, as these transformations do not change the optimum of the GIS algorithm [6].

The emission probability of a Markov state s given an acoustic observation vector x is modeled via a Gaussian distribution $p(x|\theta_s) = \mathcal{N}(x|\mu_s, \Sigma)$ with $\theta_s = \{\mu_s, \Sigma\}$ denoting a state dependent mean μ_s and a globally pooled covariance matrix Σ . Using only one covariance matrix has the advantage that the terms quadratic in x in the exponent of $\mathcal{N}$ can be neglected as they cancel due to the explicit renormalization. Thus, we obtain the required log-linear model for the emission probabilities:

$$\begin{aligned} p(W_r|X_r) &\stackrel{\text{Viterbi}}{=} \\ &= \frac{p(W_r) \cdot \max_{s_1^{T_r}|W_r} \left\{ \prod_{t=1}^{T_r} p(s_t|s_{t-1}) \cdot \mathcal{N}(x_t|\mu_{s_t}, \Sigma) \right\}}{\sum_{W \in \mathcal{M}_r} p(W) \cdot \max_{s_1^{T_r}|W} \left\{ \prod_{t=1}^{T_r} p(s_t|s_{t-1}) \cdot \mathcal{N}(x_t|\mu_{s_t}, \Sigma) \right\}} \\ &= \frac{p(W_r) \cdot \max_{s_1^{T_r}|W_r} \left\{ \prod_{t=1}^{T_r} p(s_t|s_{t-1}) e^{\sum_{d=1}^{D+1} \hat{\lambda}_{s_t,d} \cdot f_{s_t,d}(\hat{x}, s_t)} \right\}}{\sum_{W \in \mathcal{M}_r} p(W) \cdot \max_{s_1^{T_r}|W} \left\{ \prod_{t=1}^{T_r} p(s_t|s_{t-1}) e^{\sum_{d=1}^{D+1} \hat{\lambda}_{s_t,d} \cdot f_{s_t,d}(\hat{x}, s_t)} \right\}} \end{aligned} \quad (12)$$

where $\mathcal{M}_r$ denotes a set of competing word sequences, $\hat{\lambda}_s = [\eta_s^\top \Sigma^{-1} A^{-\top}] - \frac{1}{2} (\log \det(2\pi \Sigma) + \eta_s^\top \Sigma^{-1} \eta_s)]^\top \in \mathbb{R}^{D+1}$ with $\eta_s = \mu_s + A^{-\top} b$, and the augmented observation vector $\hat{x} = [x^\top, 1]^\top \in \mathbb{R}^{D+1}$. Ideally, the set $\mathcal{M}_r$ would contain all possible word sequences. In practice, $\mathcal{M}_r$ is obtained through a recognition pass and can be represented by word lattices or N -best lists [8]. Even though the ME objective function corresponds with the Maximum Mutual Information (MMI) criterion, we denote the respective function with $\mathcal{F}_{\text{ME}}$ in order to emphasize the use of the log-linear model:

$$\mathcal{F}_{\text{ME}}(\Lambda) = \sum_{r=1}^R \log \frac{p(W_r) \cdot \rho(X_r|W_r)}{\sum_{W \in \mathcal{M}_r} p(W) \cdot \rho(X_r|W)} \quad (13)$$

where

$$\rho(X_r|W) = \max_{s_1^{T_r}|W} \left\{ \prod_{t=1}^{T_r} p(s_t|s_{t-1}) e^{\sum_{d=1}^{D+1} \hat{\lambda}_{s_t,d} \cdot f_{s_t,d}(\hat{x}, s_t)} \right\}$$

Due to the omitted normalization, $\rho(X_r|W)$ is in contrast to $p(W_r|X_r)$ no real distribution. Nevertheless, it can still be used to decode a spoken utterance since it is sufficient to compare the scores only during a recognition phase. In order to maximize

Table 1: Implementation of the GIS algorithm. The parameter ε is a small positive value that controls the terminating condition of the outer loop.

$\forall i$: compute N_i , init. $\Lambda_0 = \{\lambda_i^0\}$; $j := 0$
loop
$G(\Lambda_j) := 0$, $\forall i$: $Q_i(\Lambda_j) := 0$
for each sample $n = 1, \dots, N$ do
$G(\Lambda_j) = G(\Lambda_j) + \log p_{\Lambda_j}(c_n x_n)$
for each class index $c = 1, \dots, C$ do
for each active feature i do
$Q_i(\Lambda_j) := Q_i(\Lambda_j) + p_{\Lambda_j}(c x_n)$
$\forall i$: $\lambda_i^{j+1} := \lambda_i^j + \frac{1}{F} \log(N_i/Q_i(\Lambda_j))$
if $G(\Lambda_{j+1})/N < \varepsilon$ stop; else $j := j + 1$;

the objective function, we must determine the partial derivative of $\mathcal{F}_{\text{ME}}$ wrt. $\lambda_{s,d}$:

$$\frac{\partial \mathcal{F}_{\text{ME}}(\Lambda)}{\partial \lambda_{s,d}} = \sum_{r=1}^R \sum_{t=1}^{T_r} [\gamma_{r,t}(s|W_r) - \gamma_{r,t}(s)] \cdot y_{r,t,d} \quad (14)$$

with the forward-backward (FB) probability $\gamma_{r,t}(s|W_r)$ of hypothesizing the state s at time frame t given the transcription W_r , and the generalized FB probability $\gamma_{r,t}(s)$ of hypothesizing the state s at time frame t independently of any word sequence. The FB probabilities can be estimated efficiently on word lattices or state graphs [8].

3. ME and Discriminative Training

So far, we have not yet addressed the problem of aligning the acoustic observations with the states of a Markov chain. As mentioned before, the alignment cannot be observed directly and thus has to be modeled via a hidden variable. Moreover, in general an alignment is not fixed and may change over the training passes. However, we can make the assumption that the alignment of a spoken word sequence and hence, the $N_{s,d}$ statistics remains unaltered over some GIS iterations. This assumption is indeed risky and holds a certain wrongness since an update of the parameter set Λ might change the optimal alignment of a spoken sentence. To compensate for this, the training utterances have to be re-aligned at times in order to ensure that the $N_{s,d}$ statistics are still reliable. As a consequence, the $N_{s,d}$ counts now depend on the parameter set Λ and we have to redefine the statistics from Eq. 6 and 7:

$$N_{s,d}(\Lambda) = \sum_{r=1}^R \sum_{t=1}^{T_r} \gamma_{r,t}(s|W_r) \cdot y_{r,t,d} \quad (15)$$

$$Q_{s,d}(\Lambda) = \sum_{r=1}^R \sum_{t=1}^{T_r} \gamma_{r,t}(s) \cdot y_{r,t,d} \quad (16)$$

Due to the necessary re-alignments, the optimization problem is not convex any longer and thus, we cannot expect that the GIS algorithm will converge to a global optimum. However, the gradients of the resulting update rules can still be used to train the parameters of the log-linear model of the emission probabilities:

$$\bar{\lambda}_{s,d} = \lambda_{s,d} + \frac{1}{F} \log [N_{s,d}(\Lambda)/Q_{s,d}(\Lambda)] \quad (17)$$

3.1. Comparison with Discriminative Training

In contrast to Eq. 17, the re-estimation formulae in 'standard' discriminative training lead to update rules that depend on the logarithm of the difference between $N_{s,d}$ and $Q_{s,d}$ instead of their ratio. To demonstrate this, we consider a standard optimization technique for discriminative training that is based on the Extended Baum (EB) algorithm. Although the objective functions are identical for the ME and the MMI criterion, we introduce the MMI objective function as a separate definition in order to distinguish the dependency on the model (i.e. log-linear vs. Gauss) and the optimization method used (i.e. GIS vs. EB):

$$\mathcal{F}_{\text{MMI}}(\theta) = \sum_{r=1}^R \log \frac{p(W_r) \cdot p_\theta(X_r | W_r)}{\sum_{W \in \mathcal{M}_r} p(W) \cdot p_\theta(X_r | W)} \quad (18)$$

The derivation of $\mathcal{F}_{\text{MMI}}(\theta)$ wrt. θ_s yields

$$\frac{\partial \mathcal{F}_{\text{MMI}}(\theta)}{\partial \theta_s} = \Gamma_s \left(\frac{\partial \log p(x | \theta_s)}{\partial \theta_s} \right) \quad (19)$$

where Γ_s denotes the discriminative average for state s which is defined by:

$$\Gamma_s(g(X)) := \sum_{r=1}^R \sum_{t=1}^{T_r} [\gamma_{r,t}(s|W_r) - \gamma_{r,t}(s)] \cdot g(x_{r,t}) \quad (20)$$

The EB method leads to the following re-estimation equations:

$$\bar{\mu}_{s,d} = \mu_{s,d} + \epsilon_{s,d} \cdot [\Gamma_{s,d}(x) - \mu_{s,d} \cdot \Gamma_s(1)] \quad (21)$$

with $\epsilon_{s,d} = 1/(\Gamma_s(1) - D_s)$. Here, D_s is a state specific iteration constant that controls the convergence of the training process. The choice of D_s has to ensure that $\epsilon_{s,d}$ is positive which can be achieved by setting

$$D_s = h \cdot \max \{ D_s^{\min}, 1/\beta - \Gamma_s(1) \}$$

where $D_s^{\min}$ is a constant that guarantees positive definite variances; the parameter $\beta > 0$ is chosen to prevent overflows caused by low-valued denominators and $h > 1$ is a scaling factor that controls the step size of the gradient [8]. With the decomposition of the discriminative averages Γ_s into the correct model Γ'_s and the competitive model Γ''_s ,

$$\Gamma'_s(g(X)) = \sum_{r=1}^R \sum_{t=1}^{T_r} \gamma_{r,t}(s|W_r) \cdot g(x_{r,t}) \quad (22)$$

$$\Gamma''_s(g(X)) = \sum_{r=1}^R \sum_{t=1}^{T_r} \gamma_{r,t}(s) \cdot g(x_{r,t}) \quad (23)$$

we can rewrite Eq. 17 and obtain the following re-estimation formulae for the GIS algorithm and the EB method:

$$\text{GIS: } \bar{\lambda}_{s,d} = \lambda_{s,d} + \frac{1}{F} \cdot \log [\Gamma'_{s,d}(y) / \Gamma''_{s,d}(y)] \quad (24)$$

$$\text{EB: } \bar{\mu}_{s,d} = \mu_{s,d} + \epsilon_{s,d} \cdot [\Gamma_{s,d}(x) - \mu_{s,d} \cdot \Gamma_{s,d}(1)] \quad (25)$$

Comparing both gradients we may expect that the convergence of the GIS algorithm will turn out to be very slow, since the update rules depend on the logarithm of the ratio of the correct model Γ'_s and the competing model Γ''_s , whereas the gradient of the EB algorithm is based on the difference $\Gamma'_s - \Gamma''_s$.

3.2. Integration of Linear Feature Space Transformations

The log-linear form of the emission probabilities allows for initializing the parameter set Λ with a linear feature transformation, e.g. a linear discriminant analysis (LDA). Let H denote a LDA transformation matrix. If H has full rank, then the mean m and the variance S of the scaled features can be computed from the parameters of a Gaussian distribution $\mathcal{N}(H^\top x | \mu_z, \Sigma_z)$ in the unscaled LDA-transformed feature space:

$$m = AH^{-\top} \mu_z + b \quad (26)$$

$$S = AH^{-\top} \Sigma_z H^{-1} A^\top \quad (27)$$

Thus, we obtain the following expression for λ_s , which can be used in order to initialize the free parameters of the log-linear model:

$$\lambda_s = \begin{bmatrix} S^{-1} m_s \\ -\frac{1}{2} (\log \det(2\pi \Sigma_z) + m_s^\top S^{-1} m_s) \end{bmatrix} \quad (28)$$

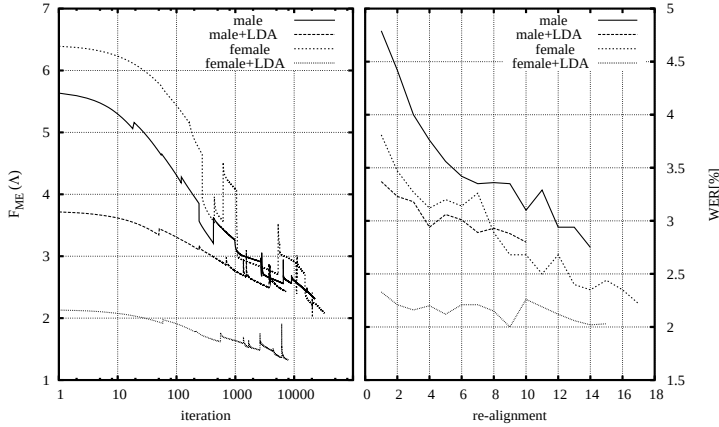


Figure 1: Evolution of the ME objective function in the course of the iteration process (left) and word error rates at the re-alignment points (right) on the SieTill training corpus.

4. Experimental Results

Experiments were performed on the *SieTill* corpus for telephone line recorded German continuous digit strings. The corpus consists of approximately 43k spoken digits in 13k sentences for both training and test set. In Table 2 some information on corpus statistics is summarized.

The recognition system is based on gender-dependent whole-word HMMs using continuous emission densities. For each gender 214 distinct states plus one for silence is used. The observation vectors consist of 12 cepstral features with first derivatives and the second derivative of the first component. The baseline recognizer applies ML training using the Viterbi approximation and achieves a word error rate (WER) of 4.59% (cf. Table 3). Both the ME training and the EB training were initialized with the ML trained parameters. Every time the objective function increases by more than 10% relative, the GIS algorithm is terminated and the training utterances are re-aligned. After 15 re-alignments, the ME based system achieves a WER of 3.52%, which is a relative improvement of 23% compared with the ML based system. Using the same number of re-alignments for the EB algorithm, the standard discriminative approach achieves a WER of 4.11%. However, the GIS algorithm requires several thousand iterations between the re-alignments in order to obtain the reported performance.

Using a Linear Discriminant Analysis (LDA), the relative performance gain decreases, but is still significant. Thus the relative improvement between the baseline result and the ME approach is 14%. In contrast to the ML and the EB trained systems, the parameters of the ME trained system were initialized with the LDA transformation matrix according to Eq. 28. As a consequence, the GIS algorithm operates on the untransformed high-dimensional features and thus, it has the possibility to extract more information from the data, whereas the ML and the EB based systems use the LDA transformed features. However, due to the small number of re-alignments, the ME based approach has not yet reached its optimum and further training iterations will be necessary.

Table 2: Corpus statistics for the *SieTill* corpus.

corpus	female		male	
	sent.	digits	sent.	digits
test	6176	20205	6938	22881
train	6113	20115	6835	22463

Table 3: Word error rates on the *SieTill* test corpus for different optimization criteria.

method	LDA	# re-align	WER[%]	SER[%]
ML	no	30	4.59	11.34
ME		15	3.52	9.17
EB		15	4.11	10.36
ML	yes	30	3.78	9.74
ME		10	3.24	8.44
EB		20	2.95	7.56

5. Conclusion

In this paper the Generalized Iterative Scaling (GIS) algorithm was applied to discriminatively train the parameters of a speech recognizer. For this purpose some constraints of the GIS algorithm were modified. The Maximum Entropy approach was analytically and experimentally compared with a standard approach to discriminative training based on the Extended Baum (EB) algorithm. Experiments conducted on a connected digit string recognition task achieved a relative improvement of up to 23% compared with a Maximum Likelihood trained system. In combination with a linear discriminant analysis, the EB algorithm achieved a performance gain of less than 10% relative compared with the ME approach.

Acknowledgments: This work was funded by the European Commission under the Human Language Technologies Project CORETEX (IST-1999-11876).

6. References

- [1] R. Rosenfeld, "A maximum entropy approach to adaptive statistical language modeling," *Computer, Speech, and Language*, vol. 10, no. 3, pp. 187–228, July 1996.
- [2] O. Bender, K. Macherey, F. J. Och, and H. Ney, "Comparison of alignment templates and maximum entropy models for natural language understanding," in *Proc. of the 10th Conf. of the Europ. Chapter of the Assoc. for Computational Linguistics*, Budapest, Hungary, Apr 2003.
- [3] F. J. Och and H. Ney, "Discriminative training and maximum entropy models for statistical machine translation," in *Proc. Annual Meeting of the Association for Computational Linguistics*, Philadelphia, PA, July 2002, pp. 295–302.
- [4] A. Likhododev and Y. Gao, "Direct models for phoneme recognition," in *Int. Conf. on Acoustics, Speech, and Signal Processing*, Orlando, FL, May 2002, vol. 1, pp. 89–92.
- [5] A. McCallum, D. Freitag, and F. Pereira, "Maximum entropy Markov models for information extraction and segmentation," in *Proc. 17th International Conf. on Machine Learning*, 2000, pp. 591–598, Morgan Kaufmann, San Francisco, CA.
- [6] J. N. Darroch and D. Ratcliff, "Generalized iterative scaling for log-linear models," *Annals of Mathematical Statistics*, vol. 43, no. 5, pp. 1470–1480, 1972.
- [7] S. Della Pietra, V. Della Pietra, and J. Lafferty, "Inducing features of random fields," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 19, no. 4, pp. 380–393, April 1997.
- [8] W. Macherey, "Implementation and comparison of discriminative training methods for automatic speech recognition," Diploma thesis, Lehrstuhl für Informatik VI, RWTH Aachen, Aachen, Nov. 1998.