Discriminative Training with Tied Covariance Matrices

Wolfgang Macherey, Ralf Schlüter, and Hermann Ney

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

Abstract

Discriminative training techniques have proved to be a powerful method for improving large vocabulary speech recognition systems based on Gaussian mixture hidden Markov models. Typically, the optimization of discriminative objective functions is done using the extended Baum algorithm. Since for continuous distributions no proof of fast and stable convergence is known up to now, parameter re-estimation depends on setting the iteration constants in the update rules heuristically, ensuring that the new variances are positive definite. In case of density specific variances this leads to a system of quadratic inequalities. However, if tied variances are used, the inequalities become more complicated and often the resulting constants are too large to be appropriate for discriminative training. In this paper we present an alternative approach to setting the iteration constants to alleviate this problem. First experimental results show that the new method leads to improved convergence speed and test set performance.

1. Introduction

Discriminative training techniques are a powerful method for improving large vocabulary speech recognition systems based on Gaussian mixture *hidden Markov models* (HMM) [1, 2]. Typically, the optimization of discriminative objective functions and hence the parameter re-estimation is done using the *extended Baum* algorithm [3]. As for continuous distributions no proof of fast and stable convergence is known up to now, the iteration constants occurring in the re-estimation formulae are set based on a heuristic which is usually twice the value necessary to ensure positive variances. Although successfully applied in practice this heuristic works best in combination with density specific variances only. The reason is that only for this setting the lower bounds on the iteration constants are given as the roots of a system of quadratic inequalities which can be computed easily. However, if tied variances are used instead, the inequalities become more complicated and the numeric values of the iteration constants ensuring positive variances increase with the number of densities sharing a common covariance. In this paper we present a novel approach to setting the iteration constants to compensate for this effect by incorporating further maximum likelihood statistics into the auxiliary function used for discriminative training. First experiments performed on a small vocabulary speech corpus show that the new method outperforms the former heuristic in terms of both convergence speed and test set performance.

The remainder of this paper is organized as follows: in Section 2 we introduce a modified version of an auxiliary function that is suitable for discriminative training with tied covariances and briefly review the theory of parameter re-estimation. Additionally we will discuss two variants on setting the iteration constants that were proposed in [1] and [4], respectively, and argue their pros and cons if tied variances are used. In Section 3 the new method for adjusting the iteration constants is derived. Experiments conducted on a small vocabulary speech recognition task are presented in Section 4, showing that the new method is more effective in optimizing discriminative objective functions with different variance tying schemes. The paper concludes with a summary and outlook in Section 5.

2. Discriminative Training

In this Section we review discriminative training under the *Maximum Mutual Information* (MMI) criterion with respect to the tying scheme for the covariance matrices. This tying scheme shall be specified via a set of equivalence classes K with each $k \in K$ comprising the set of indices of all Gaussians that share a common covariance matrix. The following considerations apply to the most general case of using full covariance matrices.

Let $X_r = x_{r1}, x_{r2}, ..., x_{rT_r}$ and $W_r = w_{r1}, w_{r2}, ..., w_{rN_r}$ denote the sequence of acoustic observation vectors and corresponding spoken words of utterances $r = 1, ..., R$ of the training data. The acoustic emission probability for a word sequence W shall be denoted by $p_\vartheta(X_r|W)$ with ϑ as the set of all parameters of the acoustic model. The language model probabilities $p(W)$ are supposed to be given and therefore do not depend on ϑ. Finally let $\mathcal{M}_r$ denote a set of competing word sequences which are considered for discrimination in utterance r. Then the objective function for the MMI criterion which is defined as the sum over the logarithms of the sentence posterior probabilities for all training utterances, can be decomposed as follows:

$$F_{\text{MMI}}(\vartheta) = \sum_{r=1}^{R} \log \frac{p_\vartheta(X_r|W_r)p(W_r)}{\sum_{W \in \mathcal{M}_r} p_\vartheta(X_r|W)p(W)} \tag{1}$$

Given an HMM state s, a mixture distribution for an acoustic vector x is denoted by $p(x|\vartheta_s)$. The according parameters ϑ_s of the mixture distribution are the weights c_{sl} and the parameters ϑ_{sl} of densities l of the mixture. Since each density is modeled as a Gaussian distribution, ϑ_{sl} is given by a density specific mean μ_{sl} and a tied covariance matrix Σ_k. Thus the derivative of $F_{\text{MMI}}(\vartheta)$ with respect to the parameters $\{c_{sl}, \vartheta_{sl}\}$ is given by:

$$\frac{\partial F_{\text{MMI}}(\vartheta)}{\partial\{c_{sl}, \vartheta_{sl}\}} = \Gamma_{sl} \quad \frac{\partial \log c_{sl} p(x|\vartheta_{sl})}{\partial\{c_{sl}, \vartheta_{sl}\}} \tag{2}$$

where the discriminative averages $\Gamma_{sl} \; g\{x\}$ are defined as:

$$\Gamma_{sl} \; g\{x\} = \Gamma_{sl}^{\text{num}} \; g\{x\} - \Gamma_{sl}^{\text{den}} \; g\{x\} \tag{3}$$

The numerator and denominator statistics are given by:

$$\Gamma_{sl}^{\text{num}}\ g\{x\} \ = \sum_{r=1}^{R}\sum_{t=1}^{T_r}\gamma_{rt}(s,l|W_r)\cdot g(x_{rt}) \tag{4}$$

$$\Gamma_{sl}^{\text{den}}\ g\{x\} \ = \sum_{r=1}^{R}\sum_{t=1}^{T_r}\gamma_{rt}(s,l)\cdot g(x_{rt}) \tag{5}$$

The discriminative averages make use of the *conditional forward-backward* (FB) probabilities of the spoken word sequences W_r (Eq. 6) and the *generalized* FB probabilities (Eq. 7). The latter is defined as the sum over the conditional FB probabilities of all competing word sequences weighted with their respective sentence posterior probability $p(W|X_r)$:

$$\gamma_{rt}(s,l|W_r) = p_\vartheta(s_t{=}s|X_r,W_r)\cdot\frac{c_{sl}p(x_{rt}|\vartheta_{sl})}{\sum_{\tilde{l}}c_{s\tilde{l}}p(x_{rt}|\vartheta_{s\tilde{l}})} \tag{6}$$

$$\gamma_{rt}(s,l) = \sum_{W\in\mathcal{M}_r}p_\vartheta(W|X_r)\cdot\gamma_{rt}(s,l|W) \tag{7}$$

2.1. Auxiliary Function and Parameter Re-Estimation

Discriminative training with the MMI criterion usually applies an extended version of *Baum-Welch* training, the EB algorithm [3]. The following auxiliary function can be used to maximize the MMI criterion using tied covariance matrices:

$$\begin{aligned}\mathcal{S}(\vartheta,\overline{\vartheta}) = \sum_{k\in K}\sum_{(s,l)\in k}\Bigg(&\sum_{r=1}^{R}\sum_{t=1}^{T_r}\Big[\gamma_{rt}(s,l|W_r)-\gamma_{rt}(s,l)\Big]\\ &\cdot\log\ \overline{c}_{sl}\cdot p(x_{rt}|\overline{\vartheta}_{sl})\\ &+D_k\cdot\int\Big\{p(x|\vartheta_{sl})\cdot\log\ \overline{c}_{sl}\cdot p(x|\overline{\vartheta}_{sl})\ \Big\}\,dx\Bigg)\end{aligned} \tag{8}$$

The quantities D_k denote the iteration constants which are specific to each equivalence class. Maximizing $\mathcal{S}(\vartheta,\overline{\vartheta})$ with respect to ϑ_{sl} gives the re-estimation formulae for the means $\overline{\mu}_{sl}$ and the tied variances $\overline{\Sigma}_k$:

$$\overline{\mu}_{sl} = \frac{\Gamma_{sl}(x)+D_k\mu_{sl}}{\Gamma_{sl}(1)+D_k} \tag{9}$$

$$\begin{aligned}\overline{\Sigma}_k = \Bigg(&\sum_{(s,l)\in k}\Big[\Gamma_{sl}(x\cdot x^\top)+D_k\cdot\ \Sigma_k+\mu_{sl}\cdot\mu_{sl}^\top\\ &-\ \Gamma_{sl}(1)+D_k\ \cdot\overline{\mu}_{sl}\cdot\overline{\mu}_{sl}^\top\Big]\Bigg)\ \sum_{(s,l)\in k}\Gamma_{sl}(1)+D_k\end{aligned} \tag{10}$$

Since the original EB update rules for mixture weights turned out to be extremely sensitive to small values of $\Gamma_{sl}(1)$, the update scheme proposed in [5] was used instead. This approach is reported to result in a faster increase in the objective function. As it is free from smoothing constants the mixture weight updates are not affected by the choice of the iteration constants D_k.

2.2. Convergence Control

A key issue in discriminative training is the choice for the iteration constants D_k. If the constants are set too large a value, convergence will be slow. On the other hand, if they are set too low a value, they might not increase the objective function. As no proof of fast and stable convergence has been found up to now, D_k is set heuristically. A useful lower bound on D_k was found to be the value which ensures that all variances in the update rules remain positive definite [6]. In case of using full covariances this leads to the following system of inequalities:

$$\forall(s,l)\in k:\ D_k > -\Gamma_{sl}(1) \tag{11}$$

$$\wedge\quad\sum_{(s,l)\in k}\frac{D_k^2\cdot\mathbf{A}_k\ +\ D_k\cdot\mathbf{B}_{sl}\ +\ \mathbf{C}_{sl}}{\Gamma_{sl}(1)+D_k} > 0 \tag{12}$$

with

$$\mathbf{A}_k = \Sigma_k,\quad \mathbf{C}_{sl} = \frac{1}{|k|}\Gamma_{sl}(1)\Gamma_k(x\cdot x^\top)-\Gamma_{sl}(x)\Gamma_{sl}(x^\top),$$

$$\mathbf{B}_{sl} = \frac{1}{|k|}\Gamma_k(x\cdot x^\top)+\Gamma_{sl}(1)\ \Sigma_k+\mu_{sl}\mu_{sl}^\top\ -2\Gamma_{sl}(x)\mu_{sl}^\top$$

2.2.1. Density Specific Variances

If density specific *full* covariance matrices are used there is a one-to-one correspondence between the equivalence classes k and the pairs (s,l), and Eq. (12) has the form of a quadratic eigenvalue problem. This eigenvalue problem can be turned into a linear problem by introducing an additional unknown eigenvector y and solving the resulting non symmetric eigensystem [7, p. 467]:

$$\begin{matrix}\mathbf{0} & \mathbf{1}\\ -\mathbf{A}_{sl}^{-1}\cdot\mathbf{C}_{sl} & -\mathbf{A}_{sl}^{-1}\cdot\mathbf{B}_{sl}\end{matrix}\cdot\begin{matrix}x\\ y\end{matrix} = D_{sl}^{\min}\cdot\begin{matrix}x\\ y\end{matrix}$$

If the variances are estimated under a *diagonal* modeling constraint, the quadratic eigenvalue problem decomposes into a system of quadratic inequalities (one inequality for each dimension d), and its largest positive real root is the lower bound on the sought iteration constant:

$$D_{sl}^{\min} = \max_d\Big\{\ -b_d+\sqrt{b_d^2-4a_dc_d}\ \ /2a_d\Big\} \tag{13}$$

with

$$a_d = \text{diag}(\mathbf{A}_{sl})_d,\quad b_d = \text{diag}(\mathbf{B}_{sl})_d,\quad c_d = \text{diag}(\mathbf{C}_{sl})_d$$

2.2.2. Former Methods on Setting the Iteration Constants

In [1] density specific diagonal variances were used with the iteration constants D_{sl} set on a per-Gaussian level to the maximum of (1) twice the value necessary to ensure positive variance updates for all dimensions of the Gaussian $k\equiv(s,l)$ and (2) a further constant E multiplied with the denominator occupancy $\Gamma_{sl}^{\text{den}}(1)$. This constant E was either set to $E=1$, or $E=2$, or a value called E_{halfmax} with $E_{\text{halfmax}} = \max_{sl}\ D_{sl}^{\min}\ /\min_{sl}\ \Gamma_{sl}^{\text{den}}(1)$. Thus the final iteration constants were given by

$$D_{sl} = \max\ \ h\cdot D_{sl}^{\min},\ E\cdot\Gamma_{sl}^{\text{den}}(1)\ \ +\epsilon,\quad h=2.0 \tag{14}$$

Setting the iteration constants according to Eq. (14) works best in combination with density specific variances only. However, if tied variances are used, the root finding problem becomes much more difficult which is due to the denominators in Eq. (12) that prevent a decomposition into an analytically easier expression. Thus, by multiplying the fractions with all denominators the resulting polynomial would have degree $|k|+1$ and even for moderate tying schemes (e.g. $10<|k|$) finding the roots would analytically be impractical. A putative remedy is to abandon the goal of finding the smallest value ensuring positive variances. Thus the fractions in Eq. (12) can be considered as independent

quadratic inequalities and the maximum over all roots would form a valid choice for D_k. However, this method holds the problem that the magnitude of D_k would be dominated by densities with nearly equal numerator and denominator statistics, i.e. for which $\Gamma_{sl}^{\text{num}}\ g\{x\} \approx \Gamma_{sl}^{\text{den}}\ g\{x\}$ holds (in that case the respective densities were hardly ever misrecognized). Such values cause very small gradients in the update rules and thus result in very large iteration constants with low convergence speed.

In [4] discriminative training using a globally pooled variance or state specific variances, respectively, was investigated. The according state specific iteration constants were determined under the additional constraint that not only the variances have to remain positive, but also the denominators in *all* re-estimation equations, including the update rules for the means. This leads to the following inequalities:

$$\overline{\sigma}_s^2 \geq \alpha > 0, \qquad \Gamma_{sl}(1) + D_s \geq \frac{1}{\beta_s} > 0 \tag{15}$$

with a positive constant α that provides a lower limit for the variances. In [4] α was set to 1. The value of the lower limit to the denominators, β_s, was determined according to the following heuristic formula:

$$\frac{1}{\beta_s} = 1 + \ |\Gamma_{s\eta_s}(1)| - 1 \ \cdot \frac{\Gamma_{s\eta_s}(1)}{\Gamma_{s\eta_s}^{\text{max}}} \tag{16}$$

with

$$\eta_s = \underset{l}{\text{argmax}}\ |\Gamma_{sl}(1)| \tag{17}$$

$$\Gamma_{s\eta_s}^{\text{max}} = \max\ \Gamma_{s\eta_s}^{\text{num}}(1), \Gamma_{s\eta_s}^{\text{den}}(1) \tag{18}$$

The idea behind this formula is to choose $1/\beta_s$ according to the magnitude of $\Gamma_{s\eta_s}(1)$, as far as the ratio $|\Gamma_{s\eta_s}(1)|/\Gamma_{s\eta_s}^{\text{max}}$ is not too low. Otherwise, if the ratio is low, the contributions of $\Gamma_{s\eta_s}^{\text{num}}(1)$ and $\Gamma_{s\eta_s}^{\text{den}}(1)$ nearly cancel and β_s approaches a fixed limit. Based on these quantities the minimal iteration constants fulfilling the constraint of positive variances were given by:

$$D_s^{\text{min}} = \max_d \left(\sum_{l=1}^{L_s} -\Gamma_{sl}(x_d^2) + 2\Gamma_{sl}(x_d)\mu_{sld}^2 - \Gamma_{sl}(1)\mu_{sld}^2 \right.$$
$$\left. + \beta_s\ \Gamma_{sl}(x_d) - \Gamma_{sl}(1)\mu_{sld}\ ^2 + \alpha\Gamma_{sl}(1) \right) \ \sum_{l=1}^{L_s}\ \sigma_{sd}^2 - \alpha$$

Thus the final iteration constants were given by:

$$D_s = h \cdot \max\ \ D_s^{\text{min}},\ \max_l \frac{1}{c_{sl}}\ \ \frac{1}{\beta_s} - \Gamma_{sl}(1) \tag{19}$$

Even though this method proved to be very effective in combination with tied variances, the iteration constants turn out to be too large if density specific variances are used (cf. Section 4).

3. Iteration Constants for Tied Covariances

A useful method for setting the iteration constants with arbitrary tying schemes should meet the following properties: (1) the method should not depend on additional parameters; (2) it should not need further constraints that go beyond the approved requirement of positive definiteness of the variances; (3) in case of using density specific variances the magnitude of the iteration constants should be in the same range as the roots of the respective system of quadratic inequalities; (4) increasing the number of Gaussians within an equivalence class should only cause a small increase in the iteration constants. Both methods described above meet these requirements only in parts. The reason for this becomes apparent when inspecting the re-estimation equations and the effect of the iteration constants. According to Eq. (9), D_k provides some kind of smoothing between the former means μ_{sl} and the discriminative statistics $\Gamma_{sl}(x)$. However, while μ_{sl} is a normalized quantity, the discriminative statistics $\Gamma_{sl}(x)$ are not. As $\Gamma_{sl}(x)$ is composed of the difference between two unnormalized quantities, $\Gamma_{sl}^{\text{num}}(x)$ and $\Gamma_{sl}^{\text{den}}(x)$, D_k has to scale μ_{sl} in such a way that the magnitudes of $D_k \cdot \mu_{sl}$ and $\Gamma_{sl}^{\text{num}}(x), \Gamma_{sl}^{\text{den}}(x)$ are within the same range. Nevertheless, this scaling should be separated from the actual iteration constants as it rather depends on the number of observations assigned with the Gaussian (s, l) than on discrimination. Therefore, D_k should be split into two terms: one term Δ_k that depends on the equivalence class k, and a further density specific term $\Lambda_{sl}(1)$ that accounts for the different magnitudes of μ_{sl} and $\Gamma_{sl}^{\text{num}}(x), \Gamma_{sl}^{\text{den}}(x)$. With the assumption that both quantities, $\Gamma_{sl}^{\text{num}}(x)$ and $\Gamma_{sl}^{\text{num}}(x)$, should be proportional to μ_{sl}, $\Lambda_{sl}(1)$ is set to the maximum likelihood estimates of the state occupancy probabilities of the last preceding training iteration. Thus by replacing the iteration constants in Eq. (8) with

$$D_k \rightarrow \Delta_k \cdot \Lambda_{sl}(1) \tag{20}$$

we obtain a modified expression of the auxiliary function:

$$\mathscr{S}(\vartheta, \widehat{\vartheta}) = \sum_{k \in K} \sum_{(s,l) \in k} \left(\sum_{r=1}^{R} \sum_{t=1}^{T_r} \Big[\gamma_{rt}(s, l|W_r) - \gamma_{rt}(s, l) \Big] \right.$$
$$\cdot \log\ \widehat{c}_{sl} \cdot p(x_{rt}|\widehat{\vartheta}_{sl})$$
$$\left. + \Delta_k \cdot \Lambda_{sl}(1) \int \Big\{ p(x|\vartheta_{sl}) \cdot \log\ \widehat{c}_{sl} \cdot p(x|\widehat{\vartheta}_{sl}) \ \Big\}\, dx \right) \tag{21}$$

Optimizing Eq. (21) with respect to $\widehat{\vartheta}$ yields the following re-estimation equations for the means and tied variances:

$$\widehat{\mu}_{sl} = \frac{\Gamma_{sl}(x) + \Delta_k \Lambda_{sl}(1)\mu_{sl}}{\Gamma_{sl}(1) + \Delta_k \Lambda_{sl}(1)} \tag{22}$$

$$\widehat{\Sigma}_k = \left(\sum_{(s,l) \in k} \Big[\Gamma_{sl}(x \cdot x^\top) + \Delta_k \Lambda_{sl}(1)\ \Sigma_k + \mu_{sl}\mu_{sl}^\top \right. \tag{23}$$
$$\left. -\ \Gamma_{sl}(1) + \Delta_k \Lambda_{sl}(1)\ \widehat{\mu}_{sl}\widehat{\mu}_{sl}^\top \Big] \right) \sum_{(s,l) \in k} \Gamma_{sl}(1) + \Delta_k \Lambda_{sl}(1)$$

As before the new mixture weights are re-estimated according to the update rule proposed in [5]. Finally the statistics from Eq. (12) have to be replaced by

$$\widehat{\mathbf{A}}_{sl} = \Lambda_{sl}^2(1) \cdot \mathbf{A}_k, \quad \widehat{\mathbf{B}}_{sl} = \Lambda_{sl}(1) \cdot \mathbf{B}_{sl} \tag{24}$$

Thus the iteration constants are given by:

$$\Delta_k = \max \Big\{ h \cdot \Delta_k^{\text{min}},\ \max_{(s,l) \in k}\ -\Gamma_{sl}(1)/\Lambda_{sl}(1)\ \Big\} + \epsilon \tag{25}$$

where Δ_k^{min} corresponds with the roots or eigenvalues of the according system of quadratic inequalities.

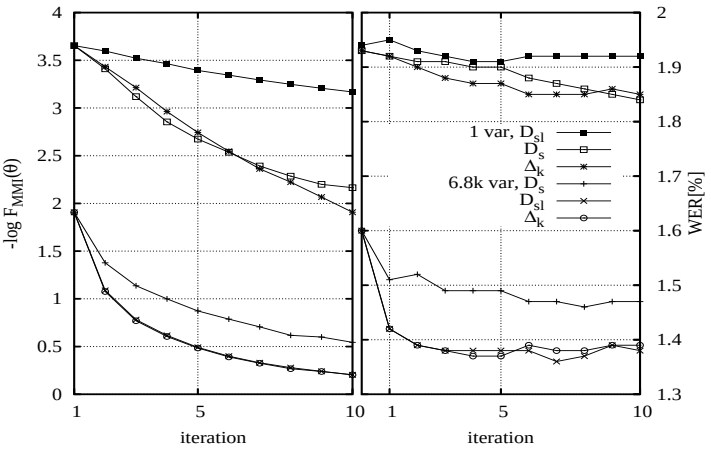

Figure 1: *MMI criterion as a function of the iteration index on the male portion of the SieTill training corpus (left) and word error rates on the SieTill test corpus (right) using either one globally pooled variance (1 var) or density specific variances (6.8k var). The iteration constants D_{sl}, D_s and Δ_k are set according to Eq. (14), (19), and (25).*

4. Experimental Results

First experimental results were produced 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. A detailed corpus description can be found in [4]. 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. Each three contiguous feature vectors are concatenated and projected via an LDA transformation matrix onto a 25 dimensional feature vector. For density specific variances the LDA matrix is combined with a MLLT matrix. The baseline recognizer applies ML training using the Viterbi approximation and achieves a word error rate (WER) of 2.04% using one globally pooled covariance matrix, and 1.56% WER using density specific variances (cf. Table 1). Both ML trained systems serve as starting points for discriminative training using the MMI criterion. Numerator and denominator lattices were re-generated in each training iteration based on an unconstrained recognition. Setting the iteration constants according to Eq. (19) results in a faster increase of the objective function and a lower WER on test data compared with setting the iteration constants according to Eq. (14) if one globally pooled covariance matrix is used. The reason is that the value of $D_{sl}^{\min}$ is dominated by nearly equally occupied numerator and denominator statistics which leads to numerically very large iteration constants. However, if density specific variances are used the effect is reversed. Thus setting D_{sl} according to Eq. (14) results in a WER of 1.36% which has to be compared with 1.46% when setting the iteration constants according to Eq. (19). In contrast to this the new approach achieves faster convergence speed without deteriorating test set performance for both tying schemes. Thus the achieved test set performance is always equal to those systems where the iteration constants were optimally set with respect to the tying scheme.

Table 1: *Word error rates (WER) on the SieTill test corpus for different tying schemes.*

male + female		WER[%]			
#var	#dns	ML	D_{sl}	D_s	Δ_k
1+1	6.8k+6.8k	2.04	1.95	1.78	**1.74**
6.2k+6.2k	6.8k+6.8k	1.56	1.36	1.46	**1.36**

5. Conclusion

In this paper a new method on setting the iteration constants for discriminative training with tied covariances was investigated. Usually different tying schemes require special methods for setting the iteration constants that account for the number of Gaussians that share a common covariance matrix. The new approach circumvents this problem by splitting the iteration constants into two parts: a density specific part that accounts for the number of observations and a variance specific part that controls the positive definiteness of the variances. This factorization turned out to be robust towards varying the tying scheme. Since the new method achieved in all cases the same error rate as the best of the former methods (which were specialized for the respective tying scheme) the new approach is appropriate to replace the former methods. Preliminary experiments currently conducted on larger speech corpora seem to confirm these results also for more challenging tasks.

Acknowledgments: This work was partially funded by the DFG (Deutsche Forschungsgemeinschaft) under the program "Structured Acoustic Models for Speech Recognition".

6. References

[1] P. C. Woodland and D. Povey, "Large scale discriminative training of hidden Markov models for speech recognition," *Computer Speech & Language*, vol. 16, no. 1, pp. 25–48, 2002.

[2] A. Gunawardana and W. Byrne, "Discriminative speaker adaptation with conditional maximum likelihood linear regression," in *Europ. Conf. on Speech Communication and Technology*, Aalborg, Denmark, Sep. 2001, vol. 2, pp. 1203–1206.

[3] P. S. Gopalakrishnan, D. Kanevsky, A. Nadas, and D. Nahamoo, "Generalization of the Baum algorithm to rational objective functions," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Glasgow, UK, May 1989, vol. 2, pp. 631–634.

[4] R. Schlüter, W. Macherey, B. Müller, and H. Ney, "Comparison of discriminative training criteria and optimization methods for speech recognition," *Speech Communication*, vol. 34, pp. 287–310, 2001.

[5] D. Povey and P. C. Woodland, "An investigation of frame discrimination for continuous speech recognition," Tech. Rep. CUED/F-INFENG/TR332, Cambridge University Engineering Department, May 2000.

[6] Y. Normandin, "Maximum mutual information estimation of hidden Markov models," in *Automatic Speech and Speaker Recognition*, C.-H. Lee, F. Soong, and K. Paliwal, Eds., pp. 57–81. Kluwer Academic Publishers, Norwell, MA, 1996.

[7] W. H. Press, W. T. Vetterling, S. A. Teukolsky, and B. P. Flannery, *Numerical Recipes in C++*, Cambridge University Press, Cambridge, UK, 2nd edition, 2002.