ELSEVIER

Speech Communication 34 (2001) 287–310

SPEECH COMMUNICATION

www.elsevier.nl/locate/specom

Comparison of discriminative training criteria and optimization methods for speech recognition

Ralf Schlüter *, Wolfgang Macherey, Boris Müller, Hermann Ney

Lehrstuhl für Informatik VI, RWTH Aachen, University of Technology, Ahornstraße 55, D-52056 Aachen, Germany

Received 12 October 1999; received in revised form 17 March 2000; accepted 5 April 2000

Abstract

The aim of this work is to build up a common framework for a class of discriminative training criteria and optimization methods for continuous speech recognition. A unified discriminative criterion based on likelihood ratios of correct and competing models with optional smoothing is presented. The unified criterion leads to particular criteria through the choice of competing word sequences and the choice of smoothing. Analytic and experimental comparisons are presented for both the *maximum mutual information* (MMI) and the *minimum classification error* (MCE) criterion together with the optimization methods gradient descent (GD) and extended *Baum* (EB) algorithm. A tree search-based restricted recognition method using word graphs is presented, so as to reduce the computational complexity of large vocabulary discriminative training. Moreover, for MCE training, a method using word graphs for efficient calculation of discriminative statistics is introduced. Experiments were performed for continuous speech recognition using the ARPA wall street journal (WSJ) corpus with a vocabulary of 5k words and for the recognition of continuously spoken digit strings using both the *TI digit string* corpus for American English digits, and the *SieTill* corpus for telephone line recorded German digits. For the MMI criterion, neither analytical nor experimental results do indicate significant differences between EB and GD optimization. For acoustic models of low complexity, MCE training gave significantly better results than MMI training. The recognition results for large vocabulary MMI training on the WSJ corpus show a significant dependence on the context length of the language model used for training. Best results were obtained using a unigram language model for MMI training. No significant correlation has been observed between the language models chosen for training and recognition. © 2001 Elsevier Science B.V. All rights reserved.

Zusammenfassung

Ziel dieser Arbeit ist die Schaffung eines einheitlichen Rahmens für eine Klasse von diskriminativen Trainingskriterien und Optimierungsmethoden für die kontinuierliche Spracherkennung. Dazu wird ein einheitliches Kriterium definiert, das auf Wahrscheinlichkeitsverhältnissen von korrekten und konkurrierenden Modellen basiert. Spezielle Kriterien ergeben sich daraus durch die Wahl der konkurrierenden Wortfolgen sowie der Glättung. Für die Kriterien *maximum mutual information* (MMI) und *minimum classification error* (MCE), sowie deren Optimierung mittels Gradientenabstieg (GD) und erweitertem *Baum* (EB) Algorithmus werden analytische und experimentelle Vergleiche durchgeführt. Die Zeitkomplexität des diskriminativen Trainings bei großem Vokabular wurde durch eine Methode zur Einschränkung der Baumsuche reduziert. Für MCE-Training wird zudem eine effiziente Methode zur Berechnung

* Corresponding author. Tel.: +49-241-80-21612; fax: +49-241-8888-219.
E-mail addresses: schlueter@informatik.rwth-aachen.de (R. Schlüter), ney@informatik.rwth-aachen.de (H. Ney).

0167-6393/01/$ - see front matter © 2001 Elsevier Science B.V. All rights reserved.
PII: S 0 1 6 7 - 6 3 9 3 (0 0) 0 0 0 3 5 - 2

diskriminativer Statistiken auf Wortgraphen eingeführt. Es wurden Experimente für kontinuierliche Spracherkennung unter Verwendung des ARPA Wall Street Journal (WSJ) Korpus (Vokabular: 5k Wörter), sowie für die kontinuierliche Ziffernkettenerkennung durchgeführt (Korpora: *TI digit string*, amerikanisches Englisch; *SieTill*, Deutsch, Telefonsprache). Analytische und experimentelle Ergebnisse gaben keine Hinweise auf signifikante Unterschiede zwischen EB und GD Optimierung des MMI-Kriteriums. MCE-Training lieferte deutlich bessere Ergebnisse als MMI-Training für suboptimale akustische Modelle. MMI Training bei großem Vokabular (WSJ Korpus) zeigte eine signifikante Abhängigkeit von der Kontextlänge des Trainingssprachmodells. Beste Ergebnisse wurden mit einem Unigramm Sprachmodell im MMI-Training erzielt. Es konnte keine signifikante Korrelation zwischen der Wahl der Sprachmodelle für Training und Erkennung beobachtet werden. © 2001 Elsevier Science B.V. All rights reserved.

Résumé

Le but de ce travail est de définir un cadre commun incluant un ensemble de critères d'apprentissage discriminant et de méthodes d'optimisation pour la reconnaissance de la parole continue. Nous introduisons un critère discriminant fondé sur le rapport entre la vraissemblance des modèles corrects et concurrents. Ce critère général conduit à définir des critères spécifiques par le choix des séquences de mots en concurrence et par celui de la méthode de lissage. Des comparaisons analytiques et expérimentales sont menées pour les critères *d'information mutuelle maximale* (MMI) et *d'erreur de classification minimum* (MCE) ainsi que pour leur optimisation par la déscente de gradient (GD) et l'algorithme *Baum* étendu (EB). Une méthode de reconnaissance restrictive fondée sur une recherche arborescente est proposée pour réduire la complexité de l'apprentissage discriminant pour les grands vocabulaires. De plus une méthode efficace a été introduite dans l'apprentissage MCE, utilisant des graphes de mots pour le calcul des statistiques discriminantes. Des expériences de reconnaissance de parole continue ont été menées sur le corpus *ARPA Wall Street Journal* (WSJ) (vocabulaire de 5k mots) ainsi que pour la reconnaissance de chiffres connectés sur les corpus *TI digit string* (anglais américain) et *Sie Till* (allemand par téléphone). Les résultats analytiques et expérimentaux n'ont pas mis en évidence des différences significatives entre les méthodes d'optimisation EB et GD pour le critère MMI. Pour des modèles acoustiques de faible complexité, l'apprentissage MCE a fourni des résultats significativement meilleurs que l'apprentissage MMI. Les résultats de reconnaissance pour l'apprentissage MMI avec un grand vocabulaire sur le corpus WSJ montrent une forte dépendance à la taille du contexte pour le modèle de langage utilisé pendant l'apprentissage. Les meilleurs résultats ont été obtenus pour un modèle de langage unigramme avec l'apprentissage MMI. Aucune corrélation significative n'a été observée entre le choix du modèle de langage pour l'apprentissage et celui pour la reconnaissance. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Discriminative training; Maximum mutual information; Minimum classification error; Corrective training; Speech recognition

1. Introduction

It has been shown that discriminative training methods are able to produce consistent, in the case of small parameter sets even large improvements in performance in comparison to the conventional *maximum likelihood* (ML) training criterion. Most applications of discriminative training methods for speech recognition use either the *maximum mutual information* (MMI) (Bahl et al., 1986; Brown, 1987; Cardin et al., 1993; Chow, 1990; Kapadia et al., 1993; Normandin, 1996; Normandin et al., 1994a,b; Normandin and Morgera, 1991; Reichl and Ruske, 1995; Valtchev et al., 1996, 1997) or the *minimum classification error* (MCE) (Chou et al., 1992, 1993, 1994; Paliwal et al., 1995; Reichl and Ruske, 1995) criterion. In MCE training, an approximation to the error rate on the training data is optimized, whereas MMI training optimizes the a posteriori probability of the training utterances and hence the class separability.

Since there does not exist any discriminative training method guaranteed to converge under all practical conditions, much effort has been made to develop parameter optimization techniques with fast and reliable convergence. The commonly used parameter optimization techniques for discriminative training are the extended *Baum* (EB) algorithm and the gradient descent (GD) method. EB is an extension to the standard *Baum–Welch*

algorithm designed for optimization of the MMI criterion. EB was first developed for discriminative training of discrete probabilities (Cardin et al., 1993; Gopalakrishnan et al., 1991; Normandin et al., 1994a; Normandin and Morgera, 1991), but was later extended to continuous densities (Normandin, 1991, 1996). Optimization of the MCE criterion is usually performed in combination with GD. In (Schlüter et al., 1997), we presented a special choice of step sizes for GD optimization of single Gaussian density parameters, showing that the EB algorithm and GD are in fact very similar and give similar recognition results in the case of the MMI criterion.

In discriminative training for speech recognition, an important point is the choice of competing word sequences and the accumulation of statistics for the discriminative model. A number of experiments have been performed using only the best recognized (for MMI) or the best incorrectly recognized (for MCE) word sequence for discrimination. For the MMI criterion this is also known as corrective training (CT) (Normandin, 1996); for MCE, we will call this approximation falsifying training (FT), since it optimizes the spoken word sequence at the expense of the best competing word sequence. To reduce complexity, in (Brown, 1987) it is assumed that the training data is segmented according to the spoken word sequence, and that a unigram language model is used for MMI training. Thereby, competing word sequences could be reduced to independent lists of competing words for each word position. Beyond these approximations, N-best lists of competing word hypotheses could be used, for which experiments have been reported both for MMI (Chow, 1990; Reichl and Ruske, 1995) and MCE training (Chou et al., 1992, 1993, 1994; Reichl and Ruske, 1995). Especially for large vocabulary applications, a much more efficient way of collecting statistics of competing word sequences is the use of word graphs. This was first presented for the MMI criterion (Normandin et al., 1994b; Valtchev et al., 1996, 1997).

Especially for large vocabulary applications, the determination of the set of competing word sequences, i.e. the recognition on the training data, takes most of the computational load needed for discriminative training. In (Valtchev et al., 1997), recognition was done once, i.e. word graphs were initially obtained for the training data, which were used for acoustic rescoring within each discriminative training iteration step.

As a further aspect, discriminative training of large vocabulary speech recognizers introduces language models to training in several views. Firstly, the language model for the – at least initial – recognition of competing word sequences for training has to be chosen. Secondly, the choice of language models for discriminative training itself will have impact on the resulting acoustic models. Finally, the question arises to what extent recognition results using a particular language model depend on the language models chosen for training. In an (MCE) training approach with a vocabulary of 1000 words, using no language model for training at all has been reported to give better results than using a word pair grammar, where in both cases a word pair grammar was used for evaluation (Chou et al., 1993). In (Valtchev et al., 1997) a bigram language model was used for MMI training of a speech recognizer with 65k vocabulary. Clearly, improvements in comparison to the baseline ML results diminished with increasing context length of the language model for recognition.

The goal of this work is to present a unified approach for efficient discriminative training of both small and large vocabulary continuous speech recognizers, using a class of discriminative training criteria (Schlüter and Macherey, 1998) and optimization methods (Schlüter et al., 1997). The approach is based on a formulation for MMI training given in (Normandin, 1996) and includes both the MMI and the MCE criterion and the corresponding corrective (CT) and falsifying training (FT) approximations, respectively. We present experiments comparing these criteria for varying degrees of model complexity, thus extending a comparison presented in (Reichl and Ruske, 1995). The parameter optimization technique chosen is based on an extension of EB (Kanevsky, 1995). The approach is formally independent of the particular criterion in question; the dependence on particular criteria is solely contained in the accumulators for discriminative statistics, which we call *discriminative averages*.

The comparison of EB and GD presented in (Schlüter et al., 1997) was completed to include mixture densities, again showing strong similarities between EB and GD optimization. It is shown, that the EB algorithm could also be interpreted as a means of finding more optimal step sizes for GD optimization of discriminative criteria.

The original method to use word graphs for the accumulation of discriminative statistics so far only applies to MMI training. Since the computational efficiency of word-based discriminative training methods, as they are discussed here, highly depends on the use of word graphs, we will extend this method, so as to apply it to criteria like MCE. For MCE training, the spoken word sequence needs to be excluded from the set of competing word sequences. In general, if the spoken word sequence would be removed from the word graph itself, other word sequences would be removed at the same time. Here, we propose an algorithm to use word graphs correctly for efficient MCE training.

For the reduction of the computational requirements of large vocabulary discriminative training, we present an approach for constrained recognition using word graphs for the determination of competing word sequences. The constrained recognition approach both preserves the recognition accuracy and significantly reduces the training times. In addition, it is shown that discriminative training using constrained recognition performs better than using word graph rescoring with fixed word boundary times.

The investigations on discriminative training are completed by systematic investigations on the interdependence between language model choice for large vocabulary MMI training and recognition. It is shown that the recognition performance of the MMI trained models significantly depend on the choice of the language model context length used for *training*. Moreover, results are presented that do not indicate considerable correlation between the choice of language models for training and recognition.

The remaining part of the paper is organized as follows. In Section 2, we present a unifying approach to discriminative training including a comparison of EB and GD optimization. Section 3 focuses on efficient ways of accumulation of discriminative statistics. In Section 4, a set of comparative experiments for small vocabulary speech recognition is presented, and in Section 5, experiments using constrained recognition for MMI training and comparative experiments using language models of varying context length for MMI training of large vocabulary speech recognition systems are discussed. Conclusions are given in Section 6.

2. Discriminative criteria

In this section, we will present a unifying approach for a class of discriminative training criteria, including the MCE, MMI and related criteria as special cases. Furthermore, a close relation between the parameter optimization methods GD and EB will be shown analytically. In addition, the interdependence of discriminative training and language models for training will be discussed (see Table 1).

2.1. Unifying view of discriminative training

The training data shall be given by training utterances r with $r = 1, \ldots, R$, each consisting of a sequence X_r of acoustic observation vectors $x_{r1}, \ldots, x_{rt}, \ldots, x_{rT_r}$ and the corresponding sequence $W_r = w_{r1}, \ldots, w_{rt}, \ldots, w_{rN_r}$ of N_r spoken words. The emission probability for an acoustic observation sequence X_r given a word sequence W_r shall be denoted by $p_\theta(X_r|W_r)$. The parameter θ represents the set of all parameters of the acoustic model. The language model probability for a word sequence W_r is defined by $p(W_r)$. In the following, the language model probabilities are assumed to be given. We now define the following unified discriminative training criterion:

$$F(\theta) = \sum_{r=1}^{R} f\left(\log \frac{p_\theta^\alpha(X_r|W_r)p^\alpha(W_r)}{\sum_{W\in\mathcal{M}_r} p_\theta^\alpha(X_r|W)p^\alpha(W)} \right) = \sum_{r=1}^{R} f\left(\log \frac{p_\theta^\alpha(X_r|W_r)}{\sum_{W\in\mathcal{M}_r} \left[\frac{p(W)}{p(W_r)}\right]^\alpha p_\theta^\alpha(X_r|W)} \right). \tag{1}$$

Table 1
List of symbols

F	unified discriminative criterion
θ	set of all parameters of the acoustic model
r	index of a speech utterance
T_r	number of time frames of utterance r
X_r	sequence of acoustic observation vectors $x_{r1}, \ldots, x_{rT_r}$ of utterance r
N_r	number of spoken words of utterance r
W_r	sequence of spoken words $w_{r1}, \ldots, w_{rN_r}$ of utterance r
$\mathscr{M}_r$	set of alternative word sequences of utterance r
$p_\theta(X_r\|W_r)$	acoustic emission probability density for utterance r given the spoken word sequence W_r
$p_\theta(W_r\|X_r)$	a posteriori probability for the spoken word sequence of utterance r
$p(W_r)$	language model probability of the spoken word sequence of utterance r
α	weighting exponent
f, f', f_r	smoothing function, the corresponding derivative, and the value of the derivative for utterance r respectively
t	time frame index
s	state of a hidden Markov model (HMM)
l	density index of a mixture density component
c_{sl}	mixture weight for density l in state s
μ_{sl}	mean vector parameter of a single Gaussian density l in state s
Σ_{sl}	covariance matrix of a single Gaussian density l in state s
σ^2_{sl}	variance vector of a single Gaussian density l in state s (diagonal covariance)
θ_{sl}	all parameters $\{c_{sl}, \mu_{sl}, \Sigma_{sl}\}$ of a single Gaussian probability density
$p(x_{rt}\|\mu_{sl}, \Sigma_{sl})$	single Gaussian emission probability density
$p_\theta(x_{rt}\|s)$	mixture Gaussian emission probability density conditioned by state s
$\delta(i,j)$	Kronecker delta, equals 1 for $i = j$, and 0 otherwise
$s_t(X_r, W)$	state of the optimal Viterbi alignment path at time t for utterance r given a word sequence W
$l(x,s)$	index of the mixture density component that maximizes the emission probability for state s given the acoustic observation x
$\gamma_{rt}(s;W)$	forward–backward (FB) probability to observe state s at time t for utterance r given a word sequence W
$\gamma_{rt}(s)$	generalized FB probability to observe state s at time t for utterance r given all alternative word sequences
$g(x)$	any (usually polynomial) function of the acoustic observations
$\Gamma_{sl}(g(x))$	discriminative averages over function g of the acoustic observations with respect to density l in state s
$\Gamma^{\text{spk}}_{sl}(g(x))$	averages over function g of the acoustic observations with respect to density l in state s for the spoken word sequences
$\Gamma^{\text{gen}}_{sl}(g(x))$	averages over function g of the acoustic observations with respect to density l in state s for all alternative word sequences
$\mathscr{S}(\theta, \hat{\theta})$	auxiliary function of the extended *Baum* (EB) algorithm
$\Delta\mu_{sl}$	step size for gradient descent (GD) optimization of mean parameter μ_{sl}
$\Delta\sigma^{\pm 2}_s$	step size for gradient descent (GD) optimization of variance parameter σ^2_{sl}
Δc_{sl}	step size for gradient descent (GD) optimization of mixture weight parameter c_{sl}
D_s	iteration constants of the EB algorithm
$q(w\|t_b, t_e, X_r)$	posterior probability to observe word w with word boundaries t_b, t_e for utterance r given acoustic observations X_r
ϱ	slope of a sigmoidal smoothing function

Here $\mathscr{M}_r$ denotes the set of discriminated or competing word sequences, over which the sum in the denominator is evaluated. The choice of the set of competing word sequences, together with the optional smoothing function f and the optional weighting exponent α determine the choice of the particular criterion. In Table 2, examples for the choice of $\mathscr{M}_r, f$ and α are listed for the MMI and MCE criterion as well as the CT and the FT criterion. The ML criterion is also contained in the unified approach and therefore is listed, too. Ideally, the MCE and FT criterion would represent the sentence error rate on the training data. Especially for FT, the argument of

Table 2
Choice of the set of competing words for discrimination, $\mathscr{M}_r$, the smoothing function f, and the weighting exponent α for several criteria included in the unified criterion

Criterion	Smoothing function $f(z)$	Word sequences included in $\mathscr{M}_r$	Exponent α
ML	Identity	–	–
MMI	Identity	All (recognized), *including* spoken	1
CT	Identity	Best recognized	(∞)
MCE	$-1/(1+e^{2\varrho z})$	All (recognized), *excluding* spoken	$\geqslant 1$
FT	$-1/(1+e^{2\varrho z})$	Best recognized, *excluding* spoken	(∞)

the smoothing function is the score difference of the spoken word sequence and the best recognized word sequence different from the spoken word sequence. Therefore, if f were a step function, the FT criterion would represent the sentence error rate on the training data, since the score difference is lower than zero for correctly recognized utterances and greater zero otherwise. In order to obtain a criterion, which is differentiable with respect to the acoustic parameters θ, a smoothed version of the step function is chosen for f instead, i.e. f is usually given by a sigmoid function. Moreover, for the MCE criterion not only the best recognized, but all (recognized) word sequences excluding the spoken word sequence are chosen for training, in order to obtain a further smoothing effect between the competing word sequences with scores near to the best recognized word sequence.

As another criterion included in the unified approach, the MMI criterion is given by the sum over the a posteriori probabilities of the spoken word sequences W_r on the training data, given the corresponding acoustic observations X_r. No smoothing function is needed for the MMI criterion, i.e. the smoothing function is given by the identity function.

For MMI and MCE training, the sets of competing word sequences, $\mathscr{M}_r$, are usually approximated by those word sequences determined by a recognition pass on the training data, as indicated in Table 2. In the cases of the CT and the FT criterion, either the determination of the set of competing word sequences, $\mathscr{M}_r$, or the definition of the weighting exponent $\alpha = \infty$ is redundant, since their choice for both CT and FT is mutually dependent. Therefore the choice of the weighting exponent is given in brackets in these cases.

All discriminative training criteria included in the unified approach represent sums over logarithmic, optionally smoothed, likelihood ratios. In other words, the objective of all discriminative training methods discussed here is to optimize the likelihood of the spoken word sequence at the expense of some competing model, which here is defined by sums over competing word sequences. If the sums over competing word sequences were extended to include sums over any models, which are not restricted to represent word sequences, then even methods like frame discriminative training (Bahl et al., 1996; Povey and Woodland, 1999) could be represented by the unified approach discussed here.

Under the supposition that the smoothing function f is increasing, the unified discriminative criterion is to be maximized according to the acoustic parameters θ. An optimization of the unified criterion therefore tries to simultaneously maximize the emission probabilities of the spoken word sequence and minimize a weighted sum over the emission probabilities for each competing word sequence given the acoustic observation sequence for each training utterance. The weights in the sum over the competing word sequences are given by the language model probabilities relative to the spoken word sequence. Thus the unified discriminative criterion optimizes the class separability according to the words under consideration of the language model.

2.1.1. Optimization of discriminative criteria

In our experiments, we apply continuous mixture density hidden *Markov* models (HMM) for acoustic modelling. The probability density for a state s is defined by

$$p_\theta(x_{rt}|s) = \sum_l c_{sl} \cdot p(x_{rt}|\mu_{sl}, \Sigma_{sl}),$$

where in the following state, indices are identified with their corresponding mixtures indices. Each index l represents a Gaussian mixture probability density $p(x_{rt}|\mu_{sl}, \Sigma_{sl})$ with parameters $\theta_{sl} = \{c_{sl}, \mu_{sl}, \Sigma_{sl}\}$, i.e. the mixture weights c_{sl}, the mean vectors μ_{sl} and the pooled, state or density specific variances Σ_{sl}. In addition, we define the forward–backward (FB) probability $\gamma_{rt}(s; W)$ for being in mixture s at time t, given a word sequence W and the acoustic observation sequence X_r of a training utterance r. In the *Viterbi* approximation (Ney, 1990), the FB probability equals one for the states of the best alignment path $s_t(X_r, W)$ and zero otherwise,

$$\gamma_{rt}(s; W) = p_\theta(s_t = s|X_r, W) = \frac{p_\theta(s_t = s, X_r|W)}{p_\theta(X_r|W)} \overset{\text{Viterbi}}{\approx} \delta(s, s_t(X_r, W)),$$

with the *Kronecker* delta function δ. Similarly, we define the FB probability $\gamma_{rt}(s)$ for being in mixture s at time t, given the acoustic observation sequence X_r of a training utterance r accumulated over the set of competing word sequences,

$$\gamma_{rt}(s) = \sum_W \frac{p_\theta^\alpha(X_r|W)p^\alpha(W)}{\sum_V p_\theta^\alpha(X_r|V)p^\alpha(V)} \cdot \gamma_{rt}(s; W) \overset{\text{Viterbi}}{\approx} \sum_W \frac{p_\theta^\alpha(X_r|W)p^\alpha(W)}{\sum_V p_\theta^\alpha(X_r|V)p^\alpha(V)} \cdot \delta(s, s_t(X_r, W)). \tag{2}$$

Formal differentiation of the unified discriminative criterion with respect to parameters θ_{sl} of the acoustic emission probabilities leads to the following expression:

$$\frac{\partial F}{\partial \theta_{sl}} = \sum_{r=1}^{R} \alpha f'\left(\log \frac{p_\theta^\alpha(X_r|W_r)p^\alpha(W_r)}{\sum_W p_\theta^\alpha(X_r|W)p^\alpha(W)}\right) \cdot \sum_{t=1}^{T_r} [\gamma_{tr}(s; W_r) - \gamma_{tr}(s)] \frac{c_{sl}p(x_{rt}|\mu_{sl}, \Sigma_{sl})}{\sum_k c_{sk}p(x_{rt}|\mu_{sk}, \Sigma_{sk})} \cdot \frac{\partial}{\partial \theta_{sl}} \log c_{sl}p(x_{rt}|\mu_{sl}, \Sigma_{sl}).$$

Since expressions like these occur frequently in the following, we define discriminative averages for functions $g(x)$ of individual acoustic observations x, separated by the contributions from the spoken (spk) and the competing word sequences (gen),

$$\Gamma_{sl}(g(x)) = \Gamma_{sl}^{\text{spk}}(g(x)) - \Gamma_{sl}^{\text{gen}}(g(x)), \tag{3}$$

with

$$\Gamma_{sl}^{\text{spk}}(g(x)) = \sum_{r=1}^{R} f_r \sum_{t=1}^{T_r} \gamma_{rt}(s; W_r) \cdot \frac{c_{sl}p(x_{rt}|\mu_{sl}, \Sigma_{sl})}{\sum_k c_{sk}p(x_{rt}|\mu_{sk}, \Sigma_{sk})} \cdot g(x_{rt}),$$

$$\Gamma_{sl}^{\text{gen}}(g(x)) = \sum_{r=1}^{R} f_r \sum_{t=1}^{T_r} \gamma_{rt}(s) \cdot \frac{c_{sl}p(x_{rt}|\mu_{sl}, \Sigma_{sl})}{\sum_k c_{sk}p(x_{rt}|\mu_{sk}, \Sigma_{sk})} \cdot g(x_{rt})$$

and

$$f_r = f'\left(\log \frac{p_\theta^\alpha(X_r|W_r)p^\alpha(W_r)}{\sum_W p_\theta^\alpha(X_r|W)p^\alpha(W)}\right).$$

Clearly, the smoothing function f leads to an utterance-wise weighting given by its derivatives f_r. Using the above definitions, the formal differentiation of the unified discriminative criterion could be reduced to

$$\frac{\partial F}{\partial \theta_{sl}} = \alpha \cdot \Gamma_{sl}\left(\frac{\partial}{\partial \theta_{sl}} \log c_{sl}p(x_{rt}|\mu_{sl}, \Sigma_{sl})\right).$$

A similar expression holds for the case of state specific parameters θ_s. In order to simplify later expressions, we define state specific discriminative averages,

$$\Gamma_s(g(x)) = \sum_l \Gamma_{sl}(g(x)).$$

Like the Viterbi approximation for the case of state time-alignment, we also apply the maximum approximation to the calculation of mixture densities. The best density index given observation x and state s shall be denoted by $l(x, s)$. Using the Viterbi approximation and maximum approximation at the mixture level, the discriminative averages can be simplified to

$$\Gamma_{sl}(g(x)) \approx \sum_{r=1}^{R} f_r \sum_{t=1}^{T_r} [\delta(s, s_t(X_r, W_r)) - \gamma_{rt}(s)] \cdot \delta(l, l(x_{rt}, s)) \cdot g(x_{rt}). \quad (4)$$

2.1.2. Comparison of different criteria

It should be noted, that characteristics of different training criteria could be discussed completely on the basis of the discriminative averages, cf. Eq. (4), since parameter optimization depends on discriminative averages only. There are three basic features, which distinguish different criteria by the discriminative averages:

- the derivative f' of the smoothing function, including its parameters;
- the set $\mathscr{M}_r$ of word sequences used for discrimination, and
- the weighting exponent α.

Here, we will discuss the MMI criterion, the CT criterion, which is an approximation to MMI, the MCE criterion, and the FT approximation to MCE. These criteria are those applied and reported most frequently in literature. Table 2 summarizes the characteristics of these criteria.

If an acoustic observation gives a significant emission probability for the state of the best alignment path of the spoken word sequence only, then the generalized FB probability for this state will be near or equal to 1. Since the discriminative averages are weighted by the difference of the FB probabilities, the contribution of the corresponding observation to the discriminative averages will cancel out or be small. This process of cancelling can be found for all criteria, but it should be remembered, that in the case of MCE training, the spoken word sequence is removed from the set $\mathscr{M}_r$. Hence, for MCE training, frame-wise cancelling could not occur for every state of the spoken word sequence, since the contributions from the spoken word sequences will be excluded from the generalized FB probabilities.

In the case of MCE training, an additional method of weighting is applied at the utterance level, given by the derivatives of the smoothing function f. These give high weights only, if the spoken word sequence could be found near to the decision boundary. Thus, the contribution of securely recognized utterances cancels for MCE also, but in contrast to MMI the contribution cancels simultaneously for all of the corresponding acoustic observations of an utterance. In the same way, very badly recognized utterances also do not contribute to reestimation for MCE.

2.2. Parameter optimization

For the case of Gaussian mixture emission densities and a given criterion, we show that EB and GD optimization are nearly equal for a special choice of step sizes for GD. Explicit reestimation formulae will be derived for the case of state specific diagonal covariances. Similar formulae apply for the case of density specific and pooled diagonal as well as full covariances.

2.2.1. Extended Baum (EB) algorithm

Discriminative training with the MMI criterion usually applies an extended version of Baum–Welch training, the EB algorithm (Gopalakrishnan et al., 1991; Normandin, 1991, 1996; Normandin et al., 1994b). For the case of continuous emission probabilities, the unified criterion is maximized via the following auxiliary function (derived from Normandin, 1991, p. 100):

$$S(\theta, \hat{\theta}) = \sum_s \sum_{r=1}^{R} f'\left(\frac{p_\theta^\alpha(X_r, W_r)}{\sum_W p_\theta^\alpha(X_r, W)}\right) \cdot \sum_{t=1}^{T_r} [\gamma_{rt}(s; W_r) - \gamma_{rt}(s)] \cdot \log p(x_{rt}|\hat{\theta}_s) + \sum_s D_s \int \mathrm{d}x\, p(x|\theta_s) \cdot \log p(x|\hat{\theta}_s),$$

which is to be optimized iteratively. It should be noted, that this auxiliary function originally was derived for the MMI criterion, for which convergence has been proved in case of discrete probability models (Gopalakrishnan et al., 1991). Later, the approach has even been generalized to cover objective functions, which are not necessarily rational with respect to the probability models (Kanevsky, 1995), like the unified criterion presented here. Nevertheless, we will discuss discriminative training with respect to continuous probability models. For the MMI criterion, it has

been shown how to extend the EB algorithm to the continuous case (Normandin, 1991), which could equally well be done for the generalization presented in (Kanevsky, 1995). However, the corresponding iteration constants D_s needed to guarantee convergence are infinite in the continuous case (Normandin, 1991), which means that convergence could not be guaranteed under realistic conditions. However, we performed this extension in order to transfer the method for choosing step sizes from the EB algorithm to gradient descent.

One motivation for this was the fact that EB has been reported to perform better than GD (Kapadia et al., 1993). Another motivation was to provide a common optimization framework which could equally well be applied to all criteria included in the unified approach. It should be noted that proportionalities found for the step sizes for GD optimization by comparison to the EB algorithm are in agreement with the results from independent theoretical considerations for GD step sizes for MCE training presented in (Chou et al., 1992).

Applying the maximum approximation for mixture density calculation, the differentiation of the above auxiliary function with respect to the new iterated parameters $\hat{\theta}_{sl}$ leads to the following expression, from which reestimation formulae can be derived by setting the corresponding derivatives equal to zero,

$$\frac{\partial S(\theta,\hat{\theta})}{\partial \hat{\theta}_{sl}} = \Gamma_{sl}\left(\frac{\partial}{\partial \hat{\theta}_{sl}} \log \hat{c}_{sl} p(x_{rt}|\hat{\theta}_{sl})\right) + D_s c_{sl} \cdot \int \mathrm{d}x\, p(x|\theta_{sl}) \frac{\partial \log \hat{c}_{sl} p(x|\hat{\theta}_{sl})}{\partial \hat{\theta}_{sl}}.$$

Analogous to the case of reestimating discrete probabilities (Normandin, 1991), here the integral term enables convergence, by smoothing the discriminative averages with the corresponding parameters of the previous iteration. The constants D_s control the convergence rate. For reestimation of Gaussian mixture densities with state-specific diagonal variance, the parameter θ_{sl} represents the initial parameter set of a density consisting of the mixture weight c_{sl}, the Gaussian mean vectors μ_{sl} and variance σ_s^2. Using the EB algorithm, we obtain the following reestimation equations for the mean vectors $\hat{\mu}_{sl,\mathrm{EB}}$ of mixture s and density l:

$$\hat{\mu}_{sl,\mathrm{EB}} = \frac{\Gamma_{sl}(x) + D_s c_{sl} \mu_{sl}}{\Gamma_{sl}(1) + D_s c_{sl}}, \tag{5}$$

the variance vector $\hat{\sigma}^2_{s,\mathrm{EB}}$:

$$\hat{\sigma}^2_{s,\mathrm{EB}} = \frac{\Gamma_s(x^2) + D_s(\sigma_s^2 + \sum_l c_{sl} \cdot \mu_{sl}^2)}{\Gamma_s(1) + D_s} - \sum_l \frac{\Gamma_{sl}(1) + D_s c_{sl}}{\Gamma_s(1) + D_s} \hat{\mu}_{sl}^2, \tag{6}$$

and the corresponding mixture weight $\hat{c}_{sl,\mathrm{EB}}$:

$$\hat{c}_{sl,\mathrm{EB}} = \frac{\partial F(\theta)/\partial c_{sl} + D_s}{\sum_k c_{sk} \cdot \partial F(\theta)/\partial c_{sk} + D_s} \cdot c_{sl}. \tag{7}$$

There also exists a more intuitive explanation for the EB reestimation formula. The reestimation equations could equally well be obtained by setting the derivatives of the unified criterion with respect to the parameters to zero, while assuming that all discriminative averages $\Gamma_{sl}(g(x))$ occurring in the resulting equations are independent of the new parameters, and assuming that each discriminative average is smoothed by D_s multiplied by the according previous parameters, i.e. smoothing $\Gamma_{sl}(1)$ by c_{sl}, $\Gamma_{sl}(x)$ by μ_{sl}, and $\Gamma_s(x^2)$ by $\sigma_s^2 + \sum_l c_{sl}\mu_{sl}^2$. It should be noted, that the reestimation (Eq. (7)) for the mixture weights is not used with the exact derivatives of the criterion F, but with smoothed versions as proposed by Normandin (1996),

$$\partial F(\theta)/\partial c_{sl} \approx \frac{\Gamma_{sl}^{\mathrm{spk}}(1)}{\Gamma_s^{\mathrm{spk}}(1)} - \frac{\Gamma_{sl}^{\mathrm{gen}}(1)}{\Gamma_s^{\mathrm{gen}}(1)}.$$

This leads to the following smoothed reestimation equation:

$$\hat{c}_{sl} = \frac{\frac{\Gamma_{sl}^{\mathrm{spk}}(1)}{\Gamma_s^{\mathrm{spk}}(1)} - \frac{\Gamma_{sl}^{\mathrm{gen}}(1)}{\Gamma_s^{\mathrm{gen}}(1)} + C_s}{\sum_{l'} c_{sl'}\left[\frac{\Gamma_{sl'}^{\mathrm{spk}}(1)}{\Gamma_s^{\mathrm{spk}}(1)} - \frac{\Gamma_{sl'}^{\mathrm{gen}}(1)}{\Gamma_s^{\mathrm{gen}}(1)}\right] + C_s} \cdot c_{sl},$$

which also requires new iteration constants C_s, since the magnitude of the smoothed terms differs from the corresponding terms for the means and variances.

2.2.2. Gradient descent

Performing gradient descent for parameter optimization, the following iterative reestimation equation is applied for parameters θ_{sl}:

$$\hat{\theta}_{sl} = \theta_{sl} + \Delta\theta_{sl} \cdot \frac{\partial F(\theta)}{\partial \theta_{sl}}.$$

For gradient descent, we obtain the following reestimation equations:

$$\hat{\mu}_{sl,\mathrm{GD}} = \mu_{sl} + \Delta\mu_{sl} \frac{\partial F(\theta)}{\partial \mu_{sl}}, \tag{8}$$

$$\hat{\sigma}^2_{s,\mathrm{GD}} = \sigma^2_s + \Delta\sigma^2_s \frac{\partial F(\theta)}{\partial \sigma^2_s} = \sigma^2_s - \Delta\sigma^{-2}_s \frac{\partial F(\theta)}{\partial \sigma^{-2}_s}, \tag{9}$$

$$\hat{c}_{sl,\mathrm{GD}} = \frac{c_{sl} + \Delta c_{sl}(\partial F(\theta)/\partial c_{sl})}{1 + \sum_k \Delta c_{sk}(\partial F(\theta)/\partial c_{sk})}. \tag{10}$$

As for the case of the EB algorithm, the derivatives for reestimation of the mixture weights are replaced by smoothed versions according to (Normandin, 1996).

2.2.3. Comparison of GD and EB

In (Schlüter et al., 1997), we derived step sizes for the case of gradient descent leading to reestimation formulae for the parameters of single densities, which resemble those of the EB algorithm. Here, this comparison is extended to mixture density modelling. The special step sizes we obtained for GD are

$$\Delta\mu_{sl} = \frac{\sigma^2_{sl}}{\Gamma_{sl}(1) + D_s \cdot c_{sl}}, \tag{11}$$

$$\Delta\sigma^{-2}_s = \frac{2}{\Gamma_s(1) + D_s}, \tag{12}$$

$$\Delta c_{sl} = \frac{\partial F(\theta)/\partial c_{sl} + D_s - 1}{\partial F(\theta)/\partial c_{sl}} \cdot c_{sl}. \tag{13}$$

Using the above step sizes for gradient descent, we obtain the following relations between the reestimated parameters of GD and EB, provided the initial parameters are equal:

$$\begin{aligned} \hat{\mu}_{sl,\mathrm{GD}} &= \hat{\mu}_{sl,\mathrm{EB}}, \\ \hat{\sigma}^2_{s,\mathrm{GD}} &= \hat{\sigma}^2_{s,\mathrm{EB}} + \sum_l \frac{\Gamma_{sl}(1) + D_s c_{sl}}{\Gamma_s(1) + D_s} \left(\mu_{sl} - \hat{\mu}_{sl,\mathrm{EB}}\right)^2, \\ \hat{c}_{sl,\mathrm{GD}} &= \hat{c}_{sl,\mathrm{EB}}. \end{aligned} \tag{14}$$

Clearly, the means and mixture weights are equally reestimated by GD and EB, whereas the variances differ only in the sum over the weighted squared step sizes of the corresponding means. Since the only dependence on the particular criterion applied is contained in the discriminative averages, the above resemblance of GD and EB holds for all criteria contained in the unified approach discussed here.

Looking at the reestimation formula for the mean vectors, we find the step sizes for GD being proportional to the corresponding variance. This result is inherited in the EB algorithm and by the above comparison it is transferred to GD without additional assumptions. The variance factor in the step sizes for GD reestimation of Gaussian mean vectors, cf. Eq. (14) was introduced independently in (Chou et al., 1992) with theoretical arguments.

2.3. Iteration control

Proofs of convergence do exist for both GD (Chou et al., 1992) and (in case of discrete probabilities) for EB (Baum and Eagon, 1967; Gopalakrishnan et al., 1991). In the case of EB reestimation of continuous emission probabilities, convergence is only proven for infinitesimal step sizes (Normandin, 1991). In practice, reasonable fast convergence is achieved in the EB case, if the iteration constants D_s are chosen in such a way that the variances remain positive (Normandin, 1996). In addition, we ensure that all denominators in the reestimation equations remain nonsingular. For density specific variances, the condition of positive variances leads to inequalities which are quadratic in the iteration constants and could be solved explicitly to give the lowest iteration constant ensuring positive variance (Valtchev et al., 1997). On the other hand, it is not possible to find an explicit formula for the lowest iteration constant ensuring the condition of positive variances for the case of pooled or state specific

variances. This is due to the second term in Eq. (6), which prevents an explicit solution, since, through $\hat{\mu}_{sl}^2$, D_s occurs in the denominator within the summation over the densities. In order to find the smallest iteration constants ensuring positive variances in the case of state specific or pooled variances we require

$$\sigma_{s,\mathrm{EB}}^2 \geqslant \sigma_{\min}, \tag{15}$$

$$\Gamma_{sl}(1) + c_{sl} D_s \geqslant \frac{1}{\beta_s}, \tag{16}$$

with positive constants $\sigma_{\min} > 0$. The value of $\sigma_{\min}$ provides a lower limit for the variances and therefore depends on the magnitude of the acoustic features. We have found a value of 1 to be appropriate, which has been approximately 10^4 times lower than the usual magnitude of the variances observed in our experiments. The value of the lower limit to the denominators, β_s, is determined according to the magnitude of the counts $\Gamma_{sl}^{\mathrm{spk}}(1)$ and $\Gamma_{sl}^{\mathrm{gen}}(1)$ and the corresponding difference $\Gamma_{sl}(1) = \Gamma_{sl}^{\mathrm{spk}}(1) - \Gamma_{sl}^{\mathrm{gen}}(1)$, cf. Eq. (3). In preliminary experiments we developed the following heuristic formula to calculate β_s to obtain optimal training convergence:

$$\frac{1}{\beta_s} = 1 + \left(|\Gamma_{s\eta_s}(1)| - 1\right) \frac{|\Gamma_{s\eta_s}(1)|}{\Gamma_{s\eta_s}^{\max}}, \tag{17}$$

with

$$\eta_s = \operatorname*{argmax}_l |\Gamma_{sl}(1)|,$$

$$\Gamma_{s\eta_s}^{\max} = \max\left\{\Gamma_{s\eta_s}^{\mathrm{spk}}(1), \Gamma_{s\eta_s}^{\mathrm{gen}}(1)\right\}.$$

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 $|\Gamma_{s\eta_s}(1)|/\Gamma_{s\eta_s}^{\max}$ is not too low. Otherwise, if the ratio is low, the contributions of $\Gamma_{s\eta_s}^{\mathrm{spk}}(1)$ and $\Gamma_{s\eta_s}^{\mathrm{gen}}(1)$ nearly cancel, which requires that β_s approaches a fixed limit. Otherwise the iteration constants would become very large, which leads to low convergence rates, as follows from Eq. (18). Using the reestimation (Eq. (6)) for the variances in the EB case, we calculated the following estimation of the minimal iteration constant $D_{s,\min}$ fulfilling the constraint of positive variances for each acoustic feature component:

$$D_{s,\min} = \frac{1}{\sigma_s^2 - \sigma_{\min}} \cdot \left\{ \beta_s \sum_l \left[\Gamma_{sl}(x) - \Gamma_{sl}(1)\mu_{sl}\right]^2 - \Gamma_s(x^2) + \sigma_{\min}\Gamma_s(1) + \sum_l \left[2\Gamma_{sl}(x) - \Gamma_{sl}(1)\mu_{sl}\right]\mu_{sl} \right\}. \tag{18}$$

Finally, for reestimation we choose

$$D_s = h \cdot \max\left\{ D_{s,\min}, \max_l \frac{1}{c_{sl}} \left[\frac{1}{\beta_s} - \Gamma_s(1)\right] \right\}. \tag{19}$$

The terms in the maximization make sure that both the constraint on the denominators and on the variances are fulfilled, cf. Eqs. (15) and (16), respectively. The global factor $h > 1$ controls the convergence of the iteration process, high values leading to low step sizes. Substituting the above choice of the iteration constant D_s into Eq. (11), we realized that the constraint on the denominators in the EB case implies an upper bound of $\beta_s \sigma_s^2 / h$ to the resulting step size of GD. This upper bound for the step sizes is reached only if the step size estimated from the constraint of positive variance becomes too high.

For the iteration constants of the mixture weights, C_s, a similar expression is applied, which makes sure, that the denominators of the reestimation equations of the mixture weights are positive, non-singular (case $\epsilon_s = 0$), and their magnitude is near to the differences of the relative counts of the corresponding reestimation equations,

$$C_s = h \cdot \left[\max\left\{ -\max_l \left[\frac{\Gamma_{sl}^{\mathrm{spk}}(1)}{\Gamma_s^{\mathrm{spk}}(1)} - \frac{\Gamma_{sl}^{\mathrm{gen}}(1)}{\Gamma_s^{\mathrm{gen}}(1)}\right], 0 \right\} + \left\{ \begin{matrix} 1 & \text{iff } \epsilon_s = 0 \\ \epsilon_s & \text{else,} \end{matrix} \right\} \right],$$

with

$$\epsilon_s = \max_l \left| \frac{\Gamma_{sl}^{\mathrm{spk}}(1)}{\Gamma_s^{\mathrm{spk}}(1)} - \frac{\Gamma_{sl}^{\mathrm{gen}}(1)}{\Gamma_s^{\mathrm{gen}}(1)} \right|.$$

In our experiments we found factors of $h = 2$ for the *TI digit string* corpus and $h = 1.1$ for the *Sie-Till* and WSJ0 corpus to be optimal in terms of convergence rate and recognition results on the training corpus.

In general, convergence can only be guaranteed, if the step sizes used are sufficiently small. Therefore, in addition to the above estimation of step sizes, the training procedure was based on the following method. In each iteration step we checked, whether the word error rate on the training data was more than doubled. If this occured, the corresponding iteration step was removed by a fall-back to the previous reestimation, for which statistics were temporarily saved, and the reestimation was repeated with an iteration constant of $h = 5$. For our large vocabulary applications this special step size control never became active – the error rate on the training data showed strictly monotonic behaviour using $h = 1.1$ throughout all iterations.

2.4. Choice of language models

For discriminative training of large vocabulary speech recognition systems, language models are introduced as a new aspect, compared to small vocabulary applications. From the definition of the discriminative criteria discussed here, it is not at all clear, what the best choice of language models for training would be. Firstly, there are three levels, at which the choice of language models might be important:

1. the recognition of competing word sequences;
2. the discriminative criterion itself; and
3. the correlation between training and recognition.

The first aspect should not have any considerable effect. In the worst case, a non-matching language model for the recognition of competing word sequences would lead to missing word sequences in the word graphs, which should not cause any problem, if the word graph densities are high enough. The second point should be significant, since the acoustic parameters obtained by MMI training directly depend on the language model. It is not clear, what effect different language models will have on discriminative training; and if there are any correlations between the language models used for training and those used for recognition on unseen test data. For MMI training, it could easily be shown that the contributions of parts of training utterances decrease with increasing score difference to corresponding competing parts. This applies for whole sentences, as well as words or even single HMM-states. Therefore, two diametrical hypotheses are conceivable.

Correlation hypothesis. With respect to the recognition situation, one would expect that only those acoustic models need optimization, which do not sufficiently discriminate between correct and incorrect word sequences. If this argument holds, a strong correlation between the language models chosen for training and evaluation has to be concluded.

Covering hypothesis. With respect to the quality of the acoustic model, the language model usually largely improves the recognition accuracy and might cover or lead away from deficiencies of the acoustic models. Such an effect would call for suboptimal language models for training. Moreover, the choice of language models for training should not considerably correlate with those chosen for evaluation.

3. Estimation of discriminative statistics

In this section, an algorithm using word graphs for MCE training and an efficient constrained recognition approach using word graphs for the determination of competing word sequences are presented.

3.1. Discriminative training using word graphs

One possibility to include more than the best recognized word sequence into the set of competing word sequences is the application of N-best lists (Chow, 1990; Chou et al., 1993). Using N-best lists, time alignment and reestimation have to be done for every word sequence contained in the N-best list. Since different word sequences of an N-best list usually only differ in few words or even states, much of the calculations done using N-best lists are redundant. This redundancy could be

prevented by using word graphs for discriminative training as introduced in (Valtchev et al., 1996, 1997).

A word graph shall be given by a set of word hypotheses w with boundary times t_b and t_e for the beginning and end of a word, respectively, which also defines the set of predecessor and successor word hypotheses. Now, using the Viterbi approximation, and taking into account that the boundary times of each word in the word graph are known, the Viterbi alignment of a word sequence used in Eq. (2) could be divided into the Viterbi alignments of each individual word of the sequence

$$s_t(X_r, W) = s_t(X_r, w, t_b, t_e),$$

where word w is part of the word sequence W and spans time t, i.e. $t_b \leqslant t \leqslant t_e$. Thus the Viterbi alignment could be done for each word of the word graph independently and separates from the sum over all competing word sequences. We now define the reweighted [1] posterior probability $q(w|t_b, t_e, X_r)$ for hypothesizing the word w with word boundary times t_b, t_e given the complete set of acoustic observations X_r of an utterance,

$$q(w|t_b, t_e, X_r) = \sum_{\substack{\{W \in \mathscr{M}_r | \\ (w|t_b, t_e) \in W\}}} \frac{p_\theta^\alpha(X_r, W)}{\sum_{V \in \mathscr{M}_r} p_\theta^\alpha(X_r, V)}. \tag{20}$$

Here, the sum in the numerator runs over all word sequences W covered by the word graph, which contain word w with boundary times t_b, t_e. The method to calculate word posterior probabilities, as presented in (Valtchev et al., 1996, 1997) is similar to the calculation of forward–backward probabilities for HMM states, where the graph of state transitions is replaced by word transitions on a word graph. A detailed description of the algorithm including the use of N-gram language models is given in (Wessel et al., 1998). Depending on the pruning characteristics chosen while producing the word graphs, the spoken word sequence might not be included into the word graph. Hence, for discriminative training, the Viterbi alignment of the spoken word sequence is forced into the word graph before constrained recognition. Even if the spoken word sequence is pruned, its time alignment will be included, before FB-word-probabilities are calculated. Both in small and large vocabulary applications, word graphs were produced using the word pair approximation (Schwartz and Austin, 1991; Ortmanns et al., 1997).

Using word probabilities and Viterbi alignment, the generalized FB probability simplifies to the following expression:

$$\gamma_{rt}(s) \overset{\text{Viterbi}}{=} \sum_{\substack{\{w \in \mathscr{M}_r | \\ t_b \leqslant t \leqslant t_e\}}} q(w|t_b, t_e, X_r) \cdot \delta(s, s_t(X_r, w, t_b, t_e)). \tag{21}$$

The sum runs over all words contained in the set $\mathscr{M}_r$ of competing word sequences represented by the word graph, which pass through time t. Using word probabilities from word graphs, the complexity of the calculation of generalized FB probabilities becomes linear to the number of words processed in the word graph while covering every possible word sequence resulting from the word graph.

It should be noted that, beyond discriminative training, the same type of word posterior probabilities has been applied successfully to improve confidence measures for several large vocabulary speech recognition tasks and corresponding languages (Wessel et al., 1998).

3.2. MCE Training using word graphs

Performing MCE training, the spoken word sequence W_r of an utterance r has to be excluded from the calculation of the reweighted word probabilities in Eq. (20). In order to be still able to perform this calculation efficiently using word graphs, the spoken word sequence would have to be excluded from the word graph. In general it is not possible to exclude the spoken word sequence from the word graph without excluding other word sequences at the same time, since particular word hypotheses of the spoken word sequence

[1] The term 'reweighted' refers to the exponent α. For $\alpha = 1$, $q(w|t_b, t_e, X_r)$ represents a true posterior probability, if all significant word sequences are included into the set $\mathscr{M}_r$.

might be part of other sequences. Therefore the sum over all word sequences in the word graph (represented by $\mathscr{M}_r$) *including* the spoken word sequence is performed first, which afterwards is subtracted by the probability of the spoken word sequences if necessary,

$$\begin{aligned} q_{\mathrm{MCE}}(w|t_{\mathrm{b}}, t_{\mathrm{e}}, X_r) &= \frac{\sum_{\substack{\{W\in\mathscr{M}_r|W\neq W_r\\ \wedge(w|t_{\mathrm{b}},t_{\mathrm{e}})\in W\}}} p_\theta^\alpha(X_r, W)}{\sum_{\{V\in\mathscr{M}_r|V\neq W_r\}} p_\theta^\alpha(X_r, V)} \\ &= \frac{\sum_{\substack{\{W\in\mathscr{M}_r|\\ (w|t_{\mathrm{b}},t_{\mathrm{e}})\in W\}}} p_\theta^\alpha(X_r, W) - \sum_{\substack{\{W\in\mathscr{M}_r|W=W_r\\ \wedge(w|t_{\mathrm{b}},t_{\mathrm{e}})\in W\}}} p_\theta^\alpha(X_r, W)}{\sum_{V\in\mathscr{M}_r} p_\theta^\alpha(X_r, V) - \sum_{\{V\in\mathscr{M}_r|V=W_r\}} p_\theta^\alpha(X_r, V)} \\ &= \frac{q(w|t_{\mathrm{b}}, t_{\mathrm{e}}, X_r) - \sum_{\substack{\{W\in\mathscr{M}_r|W=W_r\\ \wedge(w|t_{\mathrm{b}},t_{\mathrm{e}})\in W\}}} p_\theta^\alpha(X_r, W)}{\sum_w \sum_t q(w|t, T_r, X_r) - \sum_{\{V\in\mathscr{M}_r|V=W_r\}} p_\theta^\alpha(X_r, V)}, \end{aligned} \tag{22}$$

where the $q(w|t_{\mathrm{b}}, t_{\mathrm{e}}, X_r)$ (cf. Eq. (20)) are calculated using word graphs as discussed in Section 3.1. Note that a word graph could contain multiple copies of the spoken word sequence having different word boundary times. This is reflected in the sums subtracted from numerator and denominator of Eq. (22).

3.3. Constrained recognition using word graphs

For our small vocabulary applications of discriminative training, we performed unconstrained recognition every iteration step. For large vocabulary applications, unconstrained recognition for whole training corpora in every iteration of discriminative training would clearly be unrealistic in terms of computation time.

In (Valtchev et al., 1997), discriminative training using the WSJ SI-284 training corpus is reported, where unconstrained recognition was performed only once in order to produce an initial word lattice, which was then used for constrained recognition in each iteration step of discriminative training. Preliminary experiments for discriminative training applying acoustic and language model rescoring on word graphs with fixed boundary times showed little effect or even degradations in performance. As a consequence we developed a method of constrained recognition, where the boundary times are relaxed to intervals around the boundary times given by the word graph. At each time frame τ, where new word hypotheses are to be started, not only the word hypotheses starting at exactly this time frame in the word graph are allowed in this approach, but also those words starting at time frames in the vicinity of time frame τ defined by the interval $[\tau - \Delta\tau, \tau + \Delta\tau]$, as shown by a section of a word graph in Fig. 1. The successor word candidates thus obtained from the word graph are then used to reduce the possible search space by constraining the lexical tree, as illustrated in Fig. 2. This method of extended constrained recognition even enables to recognize new word sequences not originally represented by the word graph, which would not be produced by simple acoustic or language model rescoring on the word graph, because boundary times of subsequent word hypotheses might not match. In addition the approach still makes use of the advantage of a tree lexicon. In our experiments a time interval of 11 frames was used, i.e. $\Delta\tau = 5$. In order to reduce computation time, the Viterbi state alignment paths from constrained recognition were saved on disk, such that they need not be estimated again word-wise for accumulation of statistics.

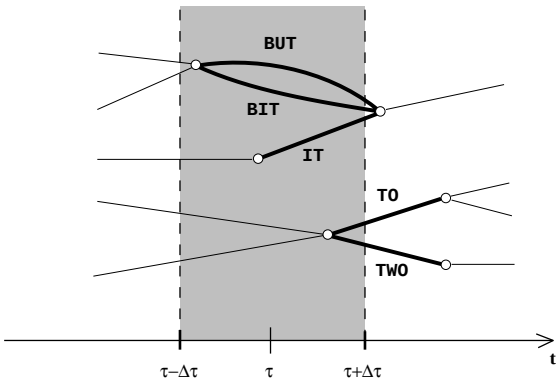

Fig. 1. Section of a word graph showing word hypotheses having beginning times in the time interval $[\tau - \Delta\tau, \tau + \Delta\tau]$. In an approach for constrained recognition, these word hypotheses serve as successor word candidates of word hypotheses ending at time frame $\tau - 1$.

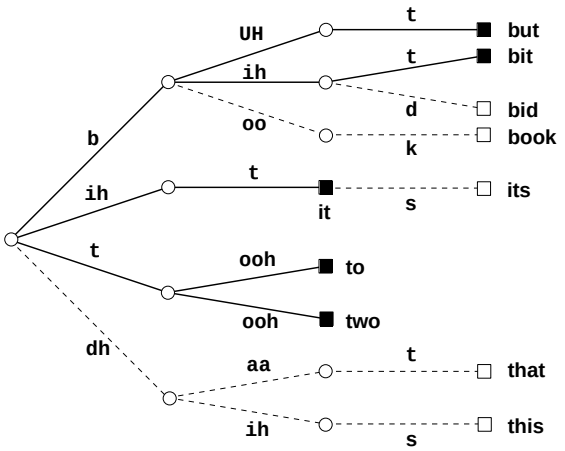

Fig. 2. Schematic view of a lexical tree for constraining the recognition to those successor word candidates determined by the word graph shown in Fig. 1. Words allowed to be hypothesized are represented by closed squares and deactivated arcs of the lexical tree are indicated by dithered lines.

4. Experiments for small vocabulary

Experiments for the comparison of discriminative training methods and optimization criteria were carried out both for the *TI digit string* (Leonard, 1984) corpus for American English digit strings and the *SieTill* (Eisele et al., 1996) corpus for telephone line recorded German digit strings. In Table 3 some information on corpus statistics is summarized. The recognition systems for both digit string corpora are based on whole word HMMs using continuous emission distributions. They are characterized as follows:

SieTill recognition system:

- 11 German digits including 'zwo';
- gender-dependent whole word and silence HMMs with 214 states plus one state for silence per gender;
- mixture Gaussian densities with global pooled or state specific diagonal covariances;
- 12 cepstral features with first derivatives and the second derivative of the energy, 10 ms frame shift.

TI digit string recognition system:

- 11 English digits including 'oh';
- gender-dependent whole word and silence HMMs with 357 states plus one state for silence per gender;
- single Gaussian densities with state specific diagonal covariances;
- 16 cepstral features with first and second derivatives, 10 ms frame shift.

Both baseline recognizers for digit-string recognition apply ML training using the Viterbi approximation (Ney, 1990) and their results serve as starting points for additional discriminative training. A detailed description of the baseline system for small vocabulary speech recognition could be found in (Welling et al., 1995).

4.1. Training procedure and complexity

All discriminative trainings for small vocabulary were initialized with a standard ML training. A standard ML training consists of a number of up to ten expectation–maximization (EM) iterations in Viterbi approximation followed by one mixture density splitting step. This procedure has been found to give optimal results for ML training. For the highest number of 64 densities per mixture presented here, one iteration of Viterbi training took about 2.5 h for each gender, resulting in a real time factor (RTF) of about 0.4 on an ALPHA 5000 PC.

The training procedures following ML training were as follows. In order to speed up training

Table 3
Corpus statistics for the SieTill and the TI digit string corpus

Corpus	Recording/language	Test/train	Female		Male		Total	
			Sent.	Digits	Sent.	Digits	Sent.	Digits
SieTill	Telephone/German	Test	6176	20 205	6938	22 881	13 114	43 086
		Train	6150	20 226	6886	22 631	13 036	42 857
TI	Microphone/English	Test	4389	14 424	4311	14 159	8700	28 583
		Train	4388	14 414	4235	13 915	8623	28 329

times, several of the experiments were performed in an additive fashion, this is CT followed by MMI, MCE or further CT, as indicated in detail below. For single densities, each discriminative training was initialized with 20 iterations of CT, which served as common starting point for MMI and MCE training as well as further CT with 10 iterations each. For FT training of single density models, 30 iterations were performed following ML training. For mixture Gaussian densities with 32 densities per state, each discriminative training was initialized with 10 iterations of CT, which served as common starting point for MMI and MCE training, respectively, as well as further CT with 10 iterations each. For FT training of models with 32 densities per mixture 10 iterations were performed following ML training. For MMI and MCE training of models with 64 densities per mixture 15 iterations were performed following the initialization with ML training.

The acoustic models used for discriminative training were exactly the same as those used for ML training, i.e., for a given number of densities per mixture, the number of trained parameters are all the same for each training method considered. In terms of computational complexity, the discriminative training methods discussed here are dominated by the recognition on the training data. Therefore the training times for MMI, CT, MCE and FT show only minor variations. One iteration of discriminative training took slightly more than 9 h resulting in an RTF of about 1.5 on an ALPHA 5000 PC, which is about 3–4 times the time needed for one ML iteration.

4.2. Convergence

Since no proof of convergence exists for EB training of the parameters of continuous density HMMs for non-infinitesimal step sizes, we first investigated the convergence behaviour of the discriminative criteria applied in this work.

In our first experiments, we applied CT for both the EB and GD optimization methods. Using iteration factors $h = 1.1$ for pooled variances (SieTill) and $h = 2$ for state-specific variances (*TI digit string*, cf. Figs. 3 and 5), we found relatively steady convergence for both GD and EB. Similar results could be observed for the word error rates on test and training data, as shown in Fig. 4 for the male portion of the *TI digit string* corpus. Clearly, convergence on test and training data is comparable, the same also holds for the female portion of the *TI digit string* corpus. Thus the convergence of the error rate on the training data was used as criterion to stop an iteration. Although overall convergence could be observed, the CT criterion (Fig. 5) and the word error rates (Fig. 6) on the SieTill training corpus show jumps in the course of

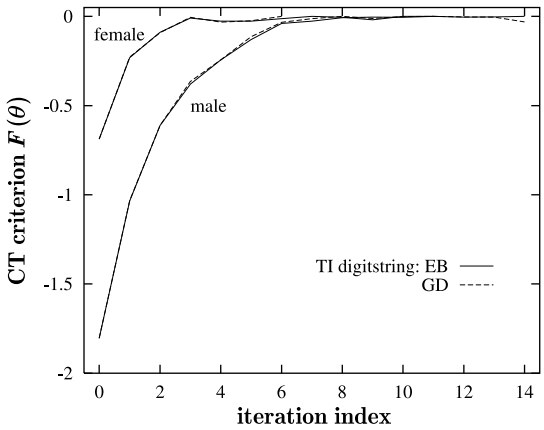

Fig. 3. CT criterion as a function of the iteration index for single Gaussian densities (TI digit string training corpus).

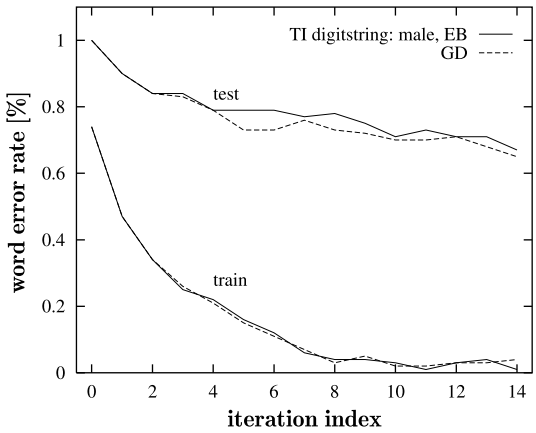

Fig. 4. Word error rate as a function of the iteration index for CT using single Gaussian densities (male portion of the TI digit string corpus).

the training iterations for both single and mixture Gaussian densities. Preliminary experiments with varying iteration factors showed that despite the jumps the choice of $h = 1.1$ was optimal according to the convergence rate. The same was observed when using the MMI, the MCE and the FT criterion. As shown in Figs. 5 and 6, the discriminative training shows non-monotonic behaviour of the criteria and error rates on the training data, which is due to the fact that convergence could not be guaranteed. Therefore, the parameter sets from discriminative training to be used for digit string recognition were chosen according to the best recognition results on the training corpus.

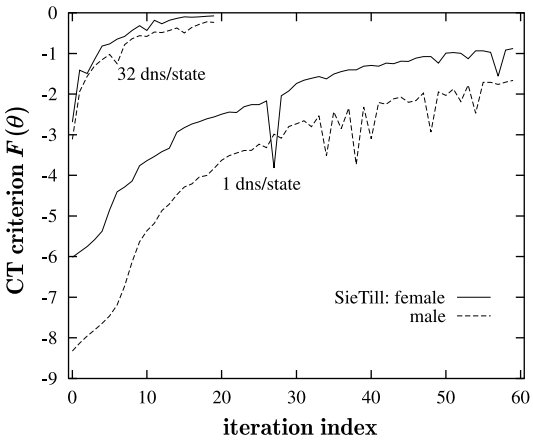

Fig. 5. CT criterion as a function of the iteration index for single and mixture Gaussian acoustic models (SieTill training corpus).

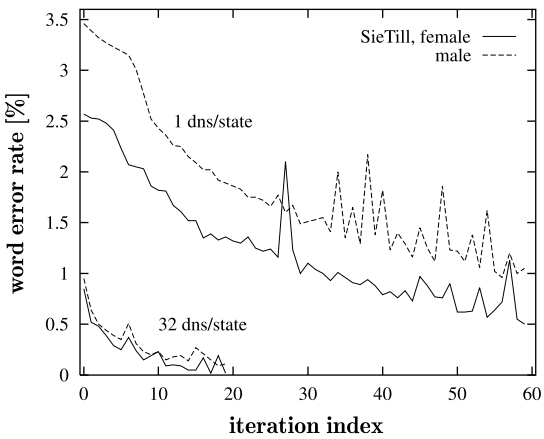

Fig. 6. Word error rates on the training corpus as a function of the iteration index for corrective training (CT) using single and mixture Gaussian acoustic models (SieTill training corpus).

4.3. Recognition results

Especially for single densities on the SieTill corpus relative improvements in word error rate of up to 1/3 compared to ML training were obtained. It should be noted that the baseline system using ML training needed about 4–8 times more parameters in order to equal the results from discriminative training using single densities. In our experiments for the SieTill corpus using mixture densities, ML training always needed more than twice the number of parameters to equal the corresponding discriminative results. The results obtained for the SieTill corpus using mixture densities are the best known to the authors.

In the case of the *TI digit string* corpus, an interesting fact is the reduction to no errors on the training data (Table 4 and Fig. 6). On the one hand this shows the strong homogeneity of the TI digit string corpus and that single densities should at least have the ability to model such a corpus completely without significant numbers of errors. On the other hand it clearly brings up the limitation of corrective training, since having no or very few errors on the training data prevents any progress in the iteration process. In the case of the SieTill corpus, a word error rate of nearly zero errors during discriminative training occurs only for a high number of 32 densities per state (cf. Table 6), indicating that in contrast to the TI digit string corpus much more detailed acoustic modelling is needed to describe the SieTill corpus properly.

4.3.1. Comparison of parameter optimization

In Tables 4 and 5, results for discriminative training comparing EB and GD optimization are given. As was expected analytically, no consistent differences between results using GD and EB reestimation could be observed for either the TI digit string or the SieTill corpus employing several kinds of acoustic modelling. Note that these comparative results on the SieTill corpus are not

Table 4
Recognition results for the TI digit string corpus. Word (WER) and sentence error rates (SER) for maximum likelihood (ML) and corrective training (CT) using extended Baum (EB) and gradient descent (GD) optimization. Single Gaussian densities with state specific diagonal covariance

Corpus	Criterion	Method	Del/ins (%)	WER (%)	SER (%)
Train	ML	–	0.28/0.04	0.56	1.69
	CT	EB	0.00/0.00	0.00	0.00
		GD	0.01/0.01	0.02	0.06
Test	ML	–	0.20/0.11	0.72	2.00
	CT	EB	0.12/0.08	0.50	1.38
		GD	0.13/0.08	0.47	1.32

Table 5
Recognition results for the SieTill corpus. Word error rates (WER) for maximum likelihood (ML) and corrective training (CT) using extended Baum (EB) and gradient descent (GD) optimization. Gaussian mixture densities with four densities per state, one pooled diagonal covariance and LDA

Corpus	Criterion	Method	Del/ins (%)	WER (%)
Train	ML	–	0.22/0.67	2.29
	CT	EB	0.21/0.16	0.67
		GD	0.21/0.15	0.69
Test	ML	–	0.35/1.04	3.05
	CT	EB	0.62/0.52	2.59
		GD	0.63/0.48	2.53

the best reported in this paper, since they were produced while still optimizing the baseline recognition system. In (Kapadia et al., 1993) it has been reported that EB optimization performed better than several variants of GD optimization. Our results show that, compared to EB, the performance of GD is a matter of appropriate step sizes, cf. Eqs. (11)–(13).

As a consequence of the comparison of EB and GD optimization, we arbitrarily chose the GD algorithm for all following experiments. For MCE training we also applied GD for parameter optimization, and used the formalism for finding optimal step sizes as obtained from the comparison of GD and EB in the case of the CT and MMI criteria.

4.3.2. Comparison of discriminative criteria

Table 6 shows recognition results for the SieTill corpus obtained for ML, CT, MMI, FT and MCE training. For single densities the best result of 2.6% word error rate was obtained using MCE training, whereas MMI training as well as CT and FT only gave word error rates around 2.8%. One reason for the good performance of MCE training for low model complexity is that outliers are ignored, since these do not have good chances to be modelled by coarse models. In contrast to this, MMI and more so CT do try to correct outliers. Furthermore, by using more than a single competing word sequence, MCE – in contrast to FT – introduces a smoothing, which facilitates the process of finding and optimizing most of those parts of a coarse model, which are possible to lead to improvements.

For mixture densities the MCE, MMI and FT criteria give consistently better results than the CT criterion. Using corrective training for mixture densities the error rate on the training corpus rapidly reduces the number of word errors on the training corpus nearly to zero. Since for corrective training only misrecognized word sequences contribute to reestimation, no further improvement could be obtained. The same was observed on the TI digit string corpus for single densities, where no recognition errors occurred on the training data. For 32 densities per mixture, no significant difference between MCE and MMI could be detected. For 64 densities MCE further improves, whereas MMI does not. Considering the best results,

Table 6
Recognition results for the SieTill corpus. Word error rates (WER) for minimum classification error (MCE), maximum mutual information (MMI) and corrective training (CT) using GD optimization and for maximum likelihood (ML) training. Gaussian mixture densities with one pooled diagonal covariance and LDA

Densities per mixture	Training criterion	Error rates (%)			
		Training		Test	
		Del–ins	WER	Del-ins	WER
1	ML	0.55–0.38	3.04	0.71–0.63	3.78
	CT	0.42–0.17	1.26	0.76–0.47	2.85
	MMI	0.45–0.16	1.28	0.81–0.41	2.81
	FT	0.54–0.19	1.48	0.65–0.64	2.80
	MCE	0.56–0.13	1.32	0.730–0.41	2.60
32	ML	0.25–0.28	0.90	0.46–0.47	1.97
	CT	0.03–0.02	0.06	0.52–0.30	1.82
	MMI	0.03–0.02	0.05	0.42–0.37	1.74
	FT	0.17–0.10	0.39	0.41–0.37	1.67
	MCE	0.04–0.04	0.11	0.41–0.37	1.75
64	ML	0.13–0.28	0.58	0.46–0.38	1.81
	MMI	0.02–0.01	0.02	0.44–0.44	1.79
	MCE	0.09–0.04	0.12	0.42–0.34	1.69

independent of the number of densities per mixture, the error rate-based discriminative training methods, MCE and FT, give better results than MMI and CT. Only marginally better than MCE, FT produces the best word error rate of 1.67% on the SieTill corpus, which means a relative improvement of nearly 8% in comparison to the best ML result. For higher numbers of densities per mixture, all methods deteriorated on the SieTill data. Overall, the MCE criterion the comparative experiments suggest, that the MCE criterion is the best choice for training of models with arbitrary complexity.

5. Experiments for large vocabulary

Experiments for large vocabulary continuous speech recognition were performed both in order to evaluate the constrained recognition approach presented and to investigate the interdependence between MMI training and the choice of language models for training. The experiments were carried out using the ARPA wall street journal (WSJ) corpus. Table 7 gives some information on corpus statistics. The recognition system used for the WSJ0 corpus is characterized as follows:

- recognition lexicon containing 4986 words plus 668 pronounciation variants;
- 2000 decision tree-based triphone states plus one state for silence;
- 96 150 gender independent Gaussian densities with global pooled diagonal covariance;
- 16 cepstral features with first derivatives and the second derivative of the energy, 10 ms frame shift;
- bigram and trigram language model.

As for the small vocabulary applications, the baseline recognizer applies ML training using the *Viterbi* approximation (Ney, 1990) and its results serve as starting point for additional discriminative training. A further description of the RWTH large vocabulary continuous speech recognition system is presented in (Ney et al., 1998).

The number of different words observed in the training corpus is more than twice the number of words contained in the recognition lexicon. Therefore these words had to be added to the recognition lexicon *for discriminative training*,

Table 7
Corpus statistics for the ARPA WSJ0 Nov. '92 development and evaluation test and training set

Corpus	Speakers	Sent.	Words
WSJ0			
Nov. '92 eval	8	330	5353
Nov. '92 dev	10	410	6779
Train	84	7240	131 395

which contains 10 108 words plus 668 pronunciation variants. This presented an additional problem: about half of the words of the training recognition lexicon are unknown to the language models for recognition. Preliminary tests with special language models for discriminative training did not produce improvements using the original language models on the test corpora. Therefore, all words, which were unknown to the language model for recognition, were mapped to the unknown word class, which was renormalized according to the number of words included into it. As a consequence, the language model perplexities on the training corpus were significantly higher than those on the test corpora. The perplexities of all language models used for the corresponding corpora are summarized in Table 8.

5.1. Training procedure, constrained recognition and complexity

Discriminative training for large vocabulary was initialized with a standard ML training, which consists of a number of six EM iterations in Viterbi approximation followed by one mixture density splitting step. This procedure has been found to give optimal results for ML training. For the optimal number of 96k mixture densities presented here, one iteration of Viterbi training took about 3.5 h on the WSJ0 training corpus, resulting in a RTF of about 0.2 on an ALPHA 5000 PC. Discriminative training used exactly the same acoustic models as ML training with the same number of 96k mixture densities, resulting in a number of about 3.2 million free parameters.

For large vocabulary tasks, discriminative training methods become computationally very extensive. Most of the training time is needed for determination and calculation of the discriminative part of the criterion and the discriminative averages. Performing unconstrained recognition, we obtained word graphs for the approximately 15 h of WSJ0 training data with a word graph density of 29. The word graphs took about 150 MB of disk space without compression. The completion of the unconstrained recognition pass on the training data took a bit less than a week on an ALPHA 5000 PC, resulting in a RTF of 10.4. This recognition time was then reduced to RTF 2.3 using the extended constrained recognition on the resulting word graph as described in Section 3.3.

Table 9 shows recognition results for MMI training with rescoring and constrained recognition in comparison to the initial ML results. Clearly, the determination of competing word

Table 8
Language model perplexities: ARPA WSJ0 training and testing corpora[a]

Corpus	Perplexity					
	Zero	Uni	Bi	Bi-phr	Tri	Tri-phr
Training	10 110	1372	398	–	289	–
Nov. '92 Dev.	–	–	107	94	58	54
Nov. '92 Eval.	–	–	107	91	53	48

[a] The notations 'bi-phr' and 'tri-phr' refer to language models containing phrases/multiwords.

Table 9
Comparison of rescoring and constrained recognition using word graphs for the determination of competing word sequences during discriminative training. Results on ARPA WSJ0 Nov. '92 corpus, training and recognition with bigram language model

Training criterion	Determination of alternative word sequences	Word error rates (%)		
		Dev	Eval	Dev & eval
ML	–	6.91	6.78	6.86
MMI	Rescoring	6.96	6.41	6.72
	Constrained recogn.	6.71	6.20	6.48

Table 10
Comparison of full (unrestricted) recognition and constrained recognition using word graphs with $\Delta\tau = 5$. Recognition with bigram language model[a]

Recognition method	Search space: number of				WER (%)	RTF
	States	Arcs	Trees	Words		
Full	6472	1835	36	106	6.86	10.5
Constrained	989	239	17	67	6.86	1.9

[a] The search space is indicated by the numbers of state, arc, tree and word hypotheses. The real time factors (RTF) correspond to an ALPHA 5000 PC. Results on the ARPA WSJ0 Nov. '92 corpus.

sequences using constrained recognition performs better than word graph rescoring, since the word boundaries from the initial word graphs are left unchanged by rescoring. Therefore, constrained recognition was chosen in all subsequent experiments on MMI training presented here.

As shown in Table 10, without any deterioration in recognition performance, the constrained recognition algorithm reduced the corresponding recognition time by a factor of more than 5, resulting in an RTF of 1.9 on an ALPHA 5000 PC. Note that these experiments were performed on unseen data. Therefore the corresponding RTFs differ from those given for training above. Including the calculation of word probabilities and the reestimation process, a single iteration step of MMI training on the ARPA WSJ0 training corpus took about 1.5 days resulting in an RTF of about 2.3 on an ALPHA 5000 PC.

5.2. Convergence

When changing to discriminative training on large vocabulary tasks the optimization methods developed for small vocabulary whole word recognizers were transferred. Especially the method to obtain estimations of the iteration constants for pooled variance had to be checked for large vocabulary. Similar to the small vocabulary applications, for MMI training on the WSJ0 training corpus good overall convergence could be observed for both the MMI criterion itself (Fig. 7) and more smoothly for the word error rate on the training data (Fig. 8). Note that, in contrast to the small vocabulary application (Fig. 4), convergence of the word error rates on the training corpus and on the development test set is not similar (Fig. 8). Consequently the choice of references for evaluation was made according to the best recognition results on the development test set.

5.3. Interdependence of language models and MMI training

In order to check the hypotheses on the interdependence of language models and discriminative training stated in Section 2.4, experiments using language models of varying context length for training and recognition were performed on the WSJ0 corpus, as shown in Table 11. The initial recognition and the constrained recognition for the trigram training has been performed using the trigram language model, and the constrained recognitions for the zerogram, unigram and bigram

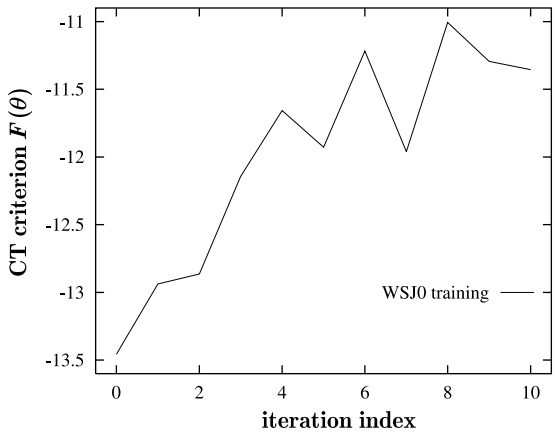

Fig. 7. MMI criterion as a function of the iteration index for the WSJ0 training corpus.

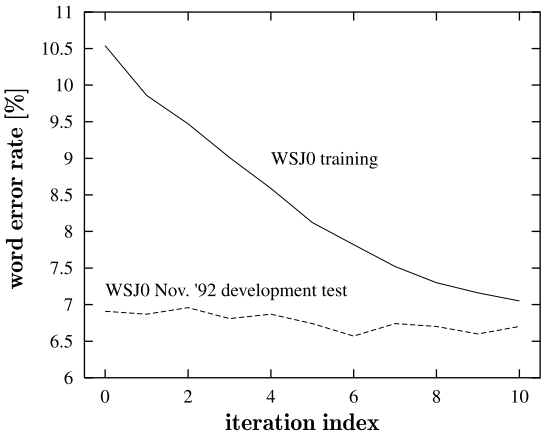

Fig. 8. Word error rates as a function of the iteration index of an MMI training for the WSJ0 training corpus and the WSJ0 Nov. '92 development test set.

training were performed using the bigram. In order to distinguish the recognition for MMI training from the test set recognition, the latter will be referred to as 'test' in the following.

For testing with either bigram or trigram language models, clearly the best results are obtained using a unigram language model for MMI training resulting in relative improvements of up to 11% in word error rate. Moreover, for testing with the bigram, the results for training with the trigram language model are even worse than those for training with the zerogram. Even for testing with the trigram, the results for training with the trigram language model are only slightly better than those for training with the zerogram. Best results were obtained using a unigram language model for MMI training, which resulted in a word error rate of 4.01% using a trigram language model for testing.

In another experiment, the correlation between the language models chosen for training and testing was examined. As shown in Table 11, in comparison to ML training the improvements obtained by MMI training using a bigram language model for training remained approximately the same for testing with a bigram, trigram, phrase-bigram and phrase-trigram language model. For these cases, the relative improvements in word error rate in comparison to ML training ranged between 5% and 6%.

It should be noted that both sets of experiments clearly support the covering hypothesis as stated in Section 2.4. It suggests that language models which are too accurate are in fact able to cover deficiencies of acoustic models by weighting down their contributions from MMI training. Moreover, the experiments presented here indicate that the improvements obtained by discriminative training using a particular language model are fairly independent of the choice of language model for evaluation.

Table 11
Comparison of several language models for MMI training and recognition. Results on ARPA WSJ0 Nov. '92 corpus

Language models		Criterion	Word error rates (%)		
Test	Training		Dev	Eval	Dev & eval
Bi	–	ML	6.91	6.78	6.86
	Zero	MMI	6.71	6.03	6.41
	Uni		**6.59**	**6.00**	**6.33**
	Bi		6.71	6.20	6.48
	Tri		6.87	6.54	6.72
Tri	–	ML	4.82	4.11	4.51
	Zero	MMI	4.63	4.05	4.38
	Uni		**4.30**	**3.64**	**4.01**
	Bi		4.48	3.94	4.24
	Tri		4.58	4.00	4.33
Bi-phrase	–	ML	6.40	5.79	6.13
	Bi	MMI	5.91	5.60	5.78
Tri-phrase	–	ML	4.76	4.26	4.54
	Bi	MMI	4.48	4.07	4.30

6. Conclusion

In this paper we presented a unifying approach to discriminative training for both small and large vocabulary speech recognition. Based on this approach a comparison of the frequently applied minimum classification error (MCE) and maximum mutual information (MMI) criteria was performed. For acoustic models of low complexity, the MCE criterion was found to give better performance than the MMI criterion, whereas for sufficient model complexity no significant differences were observed.

For parameter optimization, both gradient descent (GD) and the extended Baum (EB) algorithm were investigated. In the case of the MMI criterion, special step sizes were found for GD optimization showing strong similarities between EB and GD optimization. Consequently, experiments did not show significant differences between GD and EB. Based on the unifying criterion presented, the similarity of GD and EB was used to find optimal step sizes for GD optimization from the EB algorithm. For the case of the MCE criterion, this approach lead to good overall convergence as was the case for MMI training with EB optimization.

For large vocabulary applications of discriminative training an extended constrained recognition method using word graphs was developed. This approach was found to give better performance than acoustic and language model rescoring alone. In combination with word graph-based methods for the accumulation of discriminative statistics, it presents an improved method for efficient realization of discriminative training for large vocabulary speech recognition.

Experiments were performed both for the recognition of continuous digit strings and for large vocabulary speech recognition. For digit string recognition both the TI digit string corpus for American English digits and the SieTill corpus for telephone line recorded German digits were used. For these tasks relative improvements in word error rate of up to 1/3 were observed in comparison to ML training. Largest improvements were obtained for low complexity of acoustic models, where ML trained acoustic models needed up to 8 times more parameters to outperform discriminatively trained models. The results obtained for the SieTill corpus are the best known to the authors.

Finally, MMI training for large vocabulary speech recognition has been investigated with special reference to its interdependence with the choice of language models for training and recognition. Experiments were performed on the ARPA WSJ0 corpus. Best results were obtained using a unigram language model for MMI training. Using a trigram language model for recognition, a relative improvement of 11% was obtained in comparison to ML training leading to a word error rate of 4% on the test data. No significant correlation between the choice of language models for training and recognition has been observed.

Acknowledgements

This work was partly supported by Siemens AG, Munich.

References

Bahl, L.R., Brown, P.F., de Souza, P.V., Mercer, R.L., 1986. Maximum mutual information estimation of hidden Markov model parameters for speech recognition. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing, Tokyo, May 1986, Vol. 1, pp. 49–52.

Bahl, L.R., Padmanabhan, M., Nahamoo, D., Gopalakrishnan, P.S., 1996. Discriminative training of Gaussian mixture models for large vocabulary speech recognition systems. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing, Atlanta, GA, May 1996, Vol. 2, pp. 613–616.

Baum, L.E., Eagon, J.A., 1967. An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. Bull. Am. Math. Soc. 73, 360–363.

Brown, P.F., 1987. The acoustic-modeling problem in automatic speech recognition. Ph.D. thesis, Department of Computer Science, Carnegie Mellon University, Pittsburgh, p. 119.

Cardin, R., Normandin, Y., Millien, E., 1993. Inter-word coarticulation modeling and MMIE training for improved connected digit recognition. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing, April 1993, Minneapolis, MN, Vol. 2, pp. 243–246.

Chou, W., Juang, B.-H., Lee, C.-H., 1992. Segmental GPD training of HMM-based speech recognizer. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing, March 1992, San Francisco, CA, Vol. 1, pp. 473–476.

Chou, W., Lee, C.-H., Juang, B.-H., 1993. Minimum error rate training based on *N*-best string models. In: Proc. Internat.

Conf. on Acoustics, Speech and Signal Processing, April 1993, Minneapolis, MN, Vol. 2, pp. 652–655.

Chou, W., Lee, C.-H., Juang, B.-H., 1994. Minimum error rate training of inter-word context dependent acoustic model units in speech recognition. In: Proc. Internat. Conf. on Speech and Language Processing, September 1994, Yokohama, Japan, Vol. 2, pp. 439–442.

Chow, Y.-L., 1990. Maximum mutual information estimation of HMM parameters for continuous speech recognition using the *N*-best algorithm. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing, April 1990, Albuquerque, NM, pp. 701–704.

Eisele, T., Haeb-Umbach, R., Langmann, D., 1996. A comparative study of linear feature transformation techniques for automatic speech recognition. In: Proc. Internat. Conf. on Spoken Language Processing, October 1996, Philadelphia, PA, Vol. I, pp. 252–255.

Gopalakrishnan, P.S., Kanevsky, D., Nádas, A., Nahamoo, D., 1991. An inequality for rational functions with applications to some statistical estimation problems. IEEE Trans. Inform. Theory 37 (1), 107–113.

Kanevsky, D., 1995. A generalization of the Baum algorithm to functions on nonlinear manifolds. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing, May 1995, Detroit, MI, Vol. 1, pp. 473–476.

Kapadia, S., Valtchev, V., Young, S.J., 1993. MMI training for continuous phoneme recognition on the TIMIT database. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing, April 1993, Minneapolis, MN, Vol. 2, pp. 491–494.

Leonard, R.G., 1984. A database for speaker-independent digit recognition. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing 1984, March 1984, San Diego, CA, pp. 42.11.1–42.11.4.

Ney, H., 1990. Acoustic modeling of phoneme units for continuous speech recognition. In: Proc. Fifth European Signal Processing Conf., September 1990, Barcelona, pp. 65–72.

Ney, H., Welling, L., Ortmanns, S., Beulen, K., Wessel, F., 1998. The RWTH large vocabulary continuous speech recognition system. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing, April 1993, Seattle, WA, Vol. 2, pp. 853–856.

Normandin, Y., 1991. Hidden Markov models, maximum mutual information estimation, and the speech recognition problem. Ph.D. thesis, Department of Electrical Engineering, McGill University, Montreal, p. 159.

Normandin, Y., 1996. Maximum mutual information estimation of hidden Markov models. In: Lee, C.-H., Soong, F.K., Paliwal, K.K. (Eds.), Automatic Speech and Speaker Recognition. Kluwer Academic Publishers, Norwell, MA, pp. 57–81.

Normandin, Y., Morgera, S.D., 1991. An improved MMIE training algorithm for speaker-independent, small vocabulary, continuous speech recognition. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing, May 1991, Toronto, Canada, Vol. 1, pp. 537–540.

Normandin, Y., Cardin, R., De Mori, R., 1994a. High-performance connected digit recognition using maximum mutual information estimation. IEEE Trans. Speech Audio Process. 2 (2), 299–311.

Normandin, Y., Lacouture, R., Cardin, R., 1994b. MMIE training for large vocabulary continuous speech recognition. In: Proc. Internat. Conf. on Spoken Language Processing, September 1994, Yokohama, Vol. 3, pp. 1367–1370.

Ortmanns, S., Ney, H., Aubert, X., 1997. A word graph algorithm for large vocabulary continuous speech recognition. Comput. Speech Lang. 11 (1), 43–72.

Paliwal, K.K., Bacchiani, M., Sagisaka, Y., 1995. Minimum classification error training algorithm for feature extractor and pattern classifier in speech recognition. In: Proc. European Conf. on Speech Communication and Technology, September 1995, Madrid, Vol. 1, pp. 541–544.

Povey, D., Woodland, P.C., 1999. Frame discrimination training of HMMs for large vocabulary speech recognition. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing, May 1999, Phoenix, AZ, Vol. 1, pp. 333–336.

Reichl, W., Ruske, G., 1995. Discriminative training for continuous speech recognition. In: Proc. European Conf. on Speech Communication and Technology, September 1995, Madrid, Vol. 1, pp. 537–540.

Schlüter, R., Macherey, W., 1998. Comparison of discriminative training criteria. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing, May 1998, Seattle, WA, Vol. 1, pp. 493–496.

Schlüter, R., Macherey, W, Kanthak, S., Ney, H., Welling, L., 1997. Comparison of optimization methods for discriminative training criteria. In: Proc. European Conf. on Speech Communication and Technology, September 1997, Rhodes, Greece, Vol. 1, pp. 15–18.

Schwartz, R., Austin, S., 1991. A comparison of several approximate algorithms for finding multiple (*N*-Best) sentence hypotheses. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing, May 1991, Toronto, pp. 701–704.

Valtchev, V., Odell, J.J., Woodland, P.C., Young, S.J., 1996. Lattice-based discriminative training for large vocabulary speech recognition. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Process., May 1996, Atlanta, GA, Vol. 2, pp. 605–608.

Valtchev, V., Odell, J.J., Woodland, P.C., Young, S.J., 1997. MMIE training of large vocabulary recognition systems. Speech Communication 22 (4), 303–314.

Welling, L., Ney, H., Eiden, A., Forbrig, C., 1995. Connected digit recognition using statistical template matching. In: Proc. European Conf. on Speech Communication and Technology, September 1995, Madrid, Vol. 2, pp. 1483–1486.

Wessel, F., Macherey, K., Schlüter, R., 1998. Using word probabilities as confidence measures. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing, May 1998, Seattle, WA, Vol. 1, pp. 225–228.