Modified MMI/MPE: A Direct Evaluation of the Margin in Speech Recognition

Georg Heigold HEIGOLD@CS.RWTH-AACHEN.DE
Thomas Deselaers DESELAERS@CS.RWTH-AACHEN.DE
Ralf Schlüter SCHLUETER@CS.RWTH-AACHEN.DE
Hermann Ney NEY@CS.RWTH-AACHEN.DE
RWTH Aachen University Chair of Computer Science 6 - Computer Science Department D-52056 Aachen, Germany

Abstract

In this paper we show how common speech recognition training criteria such as the Minimum Phone Error criterion or the Maximum Mutual Information criterion can be extended to incorporate a margin term. Different margin-based training algorithms have been proposed to refine existing training algorithms for general machine learning problems. However, for speech recognition, some special problems have to be addressed and all approaches proposed either lack practical applicability or the inclusion of a margin term enforces significant changes to the underlying model, e.g. the optimization algorithm, the loss function, or the parameterization of the model. In our approach, the conventional training criteria are modified to incorporate a margin term. This allows us to do large-margin training in speech recognition using the same efficient algorithms for accumulation and optimization and to use the same software as for conventional discriminative training. We show that the proposed criteria are equivalent to Support Vector Machines with suitable smooth loss functions, approximating the non-smooth hinge loss function or the hard error (e.g. phone error). Experimental results are given for two different tasks: the rather simple digit string recognition task Sietill which severely suffers from overfitting and the large vocabulary European Parliament Plenary Sessions English task which is supposed to be dominated by the risk and the generalization does not seem to be such an issue.

Appearing in *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, 2008.

1. Introduction

A central issue in machine learning is the robust estimation of the model parameters Λ with good generalization ability, based on a finite number of observations. An interesting result from information theory is the PAC bound on the expected risk (Vapnik, 1995). The VC dimension plays an important role in this inequality and is a direct measure for the generalization ability. This bound is general in the sense that it does neither depend on the underlying probability distribution nor on the specific risk function. Furthermore, the bound implies that in general, the consideration of the empirical risk alone is suboptimal (Vapnik, 1995), see Tab. 1. Assuming that the features are in a sphere, the VC dimension of gap-tolerant classifiers is bounded above by an expression which is inversely proportional to the margin, leading to large-margin classifiers (Jebara, 2002).

These theoretical results are the main motivation for Support Vector Machines (SVMs) (Vapnik, 1995), M-SVMs (Weston & Watkins, 1999), or Hidden Markov SVMs (Altun et al., 2003) which have been successfully used for many applications in pattern recognition. The direct application of SVMs in Automatic Speech Recognition (ASR) has not been successful so far. This might be because they are not sufficiently flexible regarding: 1) the choice of the loss function, conventional criteria in ASR are Maximum Mutual Information (MMI), Minimum Classification Error (MCE), or Minimum Phone Error (MPE) which is probably the criterion of choice in ASR; 2) they are unable to cope with the immense amount of data used to train state-of-the-art ASR systems, which are commonly trained on more than 100 hours of speech (>30,000,000 observation vectors). Another problem might be the combinatorial number of classes (number of possible word sequences). Stimulated by the success of SVMs, different margin-based training algorithms have been proposed for ASR, e.g. (Yu et al., 2007; Yin & Jiang, 2007; Sha & Saul, 2007; Li et al., 2007). Although the reported results for these approaches are very promising, the approaches have some shortcomings in particular for large-scale appli-

Table 1. Relative importance of loss and margin terms under different conditions.

Loss	vs.	Margin
infinite data	↔	sparse data
many training errors	↔	few training errors

cations. The approach proposed in (Yu et al., 2007) comes closest to ours but uses MCE on N-best lists without regularization. In most state-of-the-art large-scale ASR systems, however, MPE in combination with strong regularization, i.e., i-smoothing has been established to be the criterion of choice (Povey & Woodland, 2002). In (Yin & Jiang, 2007; Sha & Saul, 2007; Li et al., 2007) not only the margin term is introduced, but the approaches use different optimization algorithms, different loss functions, or different model parameterizations which makes it difficult to evaluate the effect of the margin term in these approaches. Furthermore, none of these papers reports experimental results for competitive large vocabulary systems whose behavior in terms of generalization ability and relative improvements of performance often is different to systems using suboptimal models or for "simple" small vocabulary tasks (e.g. TIDIGITS and TIMIT). A large amount of training data and a relatively large number of training errors are typical of such large vocabulary systems. From this observation, we expect that the margin term has only little impact on the performance of such systems, cf. Tab. 1. In this work, we pursue a similar approach as in (Zhang et al., 2003) where the standard M-SVM with the hinge loss function is approximated by modified logistic regression. To the best of our knowledge, this approach, is computationally unfeasible in ASR because of the pairwise treatment of the correct and all the competing word sequences. To avoid the exponential complexity, our approximations are based on the Hidden Markov SVM proposed in (Altun et al., 2003). Formally similar results can be found in (Jebara, 2002), which are derived from probabilistic reasoning. Using the smoothed segment error of MCE in combination with $N-$best lists and without regularization, the margin-based MCE criterion proposed in (Yu et al., 2007) is recovered as a special instance of our approach.

The remainder of this paper is organized as follows: Sec. 2 reviews SVMs in a notation suitable for our discussion. Approximations to the SVMs with different loss functions, resembling the MMI and MPE criterion are proposed in Sec. 3 and extended to ASR in Sec. 4. Experimental results using these modified criteria are presented for the Sietill and the European Parliament Plenary Sessions (EPPS) English ASR tasks, cf. Sec.6. The results of the latter task give an idea of the importance of the margin in a state-of-the-art large vocabulary system. Finally, Sec. 5 shows that the transducer-based implementation of MMI and MPE differs merely in the choice of the semiring. This section may be skipped at the first reading.

2. Support Vector Machines (SVMs)

According to (Altun et al., 2003), the optimization problem of SVMs for C classes, N observation pairs (x_n, c_n), and feature functions $f_i(x, c)$ can be formulated as follows

$$\hat{\Lambda} = \arg\min_{\Lambda} \left\{ \frac{1}{2}\|\Lambda\|^2 + \frac{J}{N}\sum_{n=1}^{N} l(c_n; d_{n1}, \ldots, d_{nC}) \right\} \quad (1)$$

with $d_{nc} = \sum_i \lambda_i (f_i(x_n, c_n) - f_i(x_n, c))$, or more compactly in vector notation $d_{nc} = \lambda^\top (f(x_n, c_n) - f(x_n, c))$. The empirical constant $J > 0$ is used to balance the margin and the loss terms. The typical loss function of SVMs is the hinge loss function

$$l^{(hinge)}(c_n; d_{n1}, \ldots, d_{nC}) = \max_{c \neq c_n} \{\max\{-d_{nc} + 1, 0\}\}. \quad (2)$$

This effectively reduces the multiclass problem to a two-class problem ("correct" vs. "recognized"). Ideally, the loss function is the margin error

$$l^{(error)}(c_n; d_{n1}, \ldots, d_{nC}) = E[\hat{c}_n | c_n], \quad (3)$$

which in the simplest case counts the errors of the observations, $1 - \delta(\hat{c}_n, c_n)$. For ASR, however, we choose string-based error measures like the phone error. In this loss function, $\hat{c}_n$ is in fact a function of $(c_n; d_{n1}, \ldots, d_{nC})$ and denotes the recognized class (with margin)

$$\hat{c}_n = \begin{cases} \arg\min_{c \neq c_n} \{d_{nc}\} & \text{if } \exists c \neq c_n : d_{nc} < 1 \\ c_n & \text{otherwise.} \end{cases} \quad (4)$$

Due to the definition of the loss function and in contrast to (Altun et al., 2003), this formulation of SVM does not require the introduction of slack variables ξ_n^c subject to $d_{nc} \geq \xi_n^c + 1$ and $\xi_n^c \geq 0$ for all $c \neq c_n$ and n. The resulting optimization problem is non-smooth, but it is only used for theoretical purposes whereas the experiments are carried out with smoothed loss functions as it is common in ASR. In contrast to the multiclass SVM proposed by (Weston & Watkins, 1999), this definition allows for efficient calculation of the sum over the classes in ASR (cf. Sec. 5).

In (Taskar et al., 2003), the size of the margin is set to be proportional to the length of the sequence, e.g. the number of correct symbols. For ASR, due to the additional alignment problem, this is extended such that the margin between two sequences is set to the associated sequence/string accuracy. Note that this extension is reasonable because it guarantees consistency with the above SVM in case of i.i.d. sequences, see Sec. 4 for further details.

Finally, the task of testing consists of finding the class with the highest score

$$\hat{c}(x) = \arg\max_{c} \left\{ \lambda^\top f(x, c) \right\}, \quad (5)$$

which should not be confused with $\hat{c}_n$ in Eq. (4).

3. SVMs with Smooth Loss Functions

This section provides smooth approximations to the SVM in Eq. (1) for different loss functions. More precisely, the loss function is replaced with a smoothed loss function without breaking the large margin nature of the original SVM. These approximations are identical to modified formulations of the well-known training criteria MMI and MPE for Hidden Conditional Random Fields (HCRFs), which are introduced in the next two subsections. Analogously, a similar result can be derived for (lattice-based) MCE. In contrast to (most) other margin-based approaches, these approximations have the advantage that the effect of the margin can be evaluated directly without changing the parameterization of the model, the loss function, or the optimization algorithm.

Keep in mind that the modifications concern only the training, i.e., the calculation of the probabilities in the search remains unchanged:

$$p_\Lambda(c|x) = \frac{\exp(\lambda^\top f(x,c))}{\sum_{c'} \exp(\lambda^\top f(x,c'))}.$$

The resulting decision rule is equivalent to the decision rule in Eq. (5) for SVMs because monotone transformations of the discriminant function do not change the decision rule.

In the next two subsections, we define modified criteria based on the conventional MMI and MPE criteria and show the relationship with SVMs.

3.1. Modified Maximum Mutual Information (MMI)

In ASR, MMI commonly refers to the maximum likelihood (ML) for the class posteriors. We define a modified MMI criterion for log-linear HCRFs [1]

$$\mathcal{F}_\gamma^{(MMI)}(\Lambda) = \frac{1}{2}\|\Lambda\|^2 - \frac{J}{N}\sum_{n=1}^{N}\frac{1}{\gamma}\log\left(\frac{\exp(\gamma\,(\lambda^\top f(x_n,c_n)-1))}{\sum_c \exp(\gamma\,(\lambda^\top f(x_n,c)-\delta(c,c_n)))}\right). \quad (6)$$

See Fig. 1 for a comparison of the hinge loss function, MMI, and modified MMI. The approximation level γ is an additional parameter to control the smoothness of the criterion. The regularization constant is proportional to $\frac{1}{J}$. The major difference to the standard MMI formulation (including L_2-norm regularization) is the additional margin parameter which is non-zero only for the correct class c_n. This margin term can be interpreted as an additional observation dependent prior, weakening the true prior (Jebara, 2002).

It can be shown that the objective function $\mathcal{F}_\gamma^{(MMI)}(\Lambda)$ converges pointwise to the SVM optimization problem using the hinge loss function in Eq. (2) for $\gamma \to \infty$, similar to (Zhang et al., 2003). In other words, $\mathcal{F}_\gamma^{(MMI)}(\Lambda)$ is a smooth approximation to an SVM with hinge loss function, which can be optimized with standard gradient-based optimization techniques. The proof mainly consists of building the limit of the logarithm in Eq. (6):

$$\begin{aligned}
&-\frac{1}{\gamma}\log\left(\frac{\exp(\gamma\,(\lambda^\top f(x_n,c_n)-1))}{\sum_c \exp(\gamma\,(\lambda^\top f(x_n,c)-\delta(c,c_n)))}\right)\\
&=\frac{1}{\gamma}\log\left(1+\sum_{c\neq c_n}\exp(\gamma\,(-d_{nc}+1))\right)\\
&\overset{\gamma\to\infty}{\to}\begin{cases}\max_{c\neq c_n}\{-d_{nc}+1\} & \text{if } \exists c\neq c_n : d_{nc}<1\\ 0 & \text{otherwise.}\end{cases}
\end{aligned}$$

This function can be identified with the hinge loss function in Eq. (2).

We feel that the weak point about the hinge loss in pattern recognition is that it is not the measure used to evaluate the recognition systems eventually. This means that there is some guarantee regarding the generalization for the hinge loss, but *not* the recognition error. Furthermore, it is often unclear how these two quantities are related.

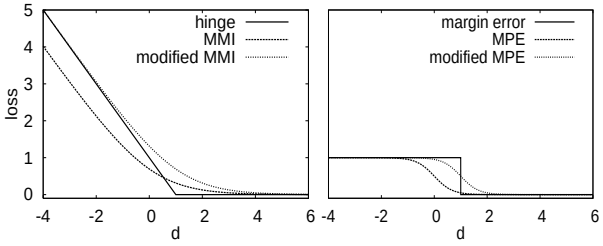

Figure 1. Left: comparison of hinge loss, MMI, and modified MMI, $\gamma = 1$. Right: comparison of margin error loss, MPE, and modified MPE, $\gamma = 3$. In either case $C = 2$, and $d = d_{nc_n}$.

3.2. Modified Minimum Phone Error (MPE)

In contrast to the hinge loss, the recognition error is bounded as illustrated in Fig. 1. Hence, a single observation cannot dominate the objective function. In particular, do not mix up a weighted margin with a weighted error.

We shall show that the modified MPE-like objective function representing a smoothed margin error with L_2-norm regularization,

$$\mathcal{F}_\gamma^{(MPE)}(\Lambda) = \frac{1}{2}\|\Lambda\|^2 + \frac{J}{N}\sum_{n=1}^{N}\sum_c E[c|c_n]\frac{\exp(\gamma\,(\lambda^\top f(x_n,c)-\delta(c,c_n)))}{\sum_{c'}\exp(\gamma\,(\lambda^\top f(x_n,c')-\delta(c',c_n)))}$$

converges to the above SVM optimization problem with a hard and weighted loss function $E[\cdot|\cdot]$ as in Eq. (3), e.g. the phone error. The proof is analogous to the proof for MMI.

[1] The first order features in (Zhang et al., 2003) are a special case of the more general feature functions used here.

The main step is to show that the "posterior probabilities" in $\mathcal{F}_\gamma^{(MPE)}(\Lambda)$ converge to a Kronecker delta such that only a single term contributes to the sum of the empirical risk

$$\begin{aligned}
&\frac{\exp(\gamma\,(\lambda^\top f(x_n,c)-\delta(c,c_n)))}{\sum_{c'}\exp(\gamma\,(\lambda^\top f(x_n,c')-\delta(c',c_n)))}\\
&= \begin{cases}\frac{1}{1+\sum_{c'\neq c_n}\exp(\gamma(-d_{nc'}+1))} & \text{if } c=c_n,\\ \frac{\exp(\gamma(-d_{nc}+1))}{1+\sum_{c'\neq c_n}\exp(\gamma(-d_{nc'}+1))} & \text{otherwise}\end{cases}\\
&\overset{\gamma\to\infty}{\to} \begin{cases}\delta(c,\arg\min_{c\neq c_n}\{d_{nc}\}) & \text{if } \exists c\neq c_n: d_{nc}<1\\ \delta(c,c_n) & \text{if } \forall c\neq c_n: d_{nc}>1\end{cases}\\
&= \delta(c,\hat{c}_n).
\end{aligned}$$

Note that now we have pointwise convergence almost surely (i.e., everywhere except for points on the decision boundary where the loss function is not continuous). As before, $\hat{c}_n$ denotes the recognized class with margin defined in Eq. (4). In summary, we have $\mathcal{F}_\gamma^{(MPE)}(\Lambda) \overset{\gamma\to\infty}{\to} \frac{1}{2}\|\Lambda\|^2 + \frac{J}{N}\sum_{n=1}^N E[\hat{c}_n|c_n]$ which is identical to the SVM optimization problem using the loss function in Eq. (3).

3.3. Optimization

In general, the resulting optimization problems are no longer convex and thus, the optimization might get stuck in local optima. We believe that this problem is inherent in ASR, e.g. due to the time alignment from HMMs. Although it is possible to make the objective function convex by keeping the alignment fixed, the best results on large-scale tasks that are reported in the literature have been obtained by using non-convex objective functions. Finally, the problem of local optima is alleviated by combining the suggested approach with stochastic annealing techniques where the approximation level acts as the temperature.

In fact, the optimization strategy suggested in (Zhang et al., 2003) can be adopted, i.e., find the optimum for a given approximation level and carry out this step iteratively for increasingly finer levels. The optimization can be done with general optimization algorithms, e.g. RProp. The idea of incrementally regulated discriminative margins suggested by (Yu et al., 2007) is along the same lines.

In this work, the approximation level and the margin are chosen beforehand and then kept fixed during the complete optimization. This single step optimization scheme has the advantage that the loss function remains unchanged and that thus, the criterion differs only in the margin term. This approach is reasonable as long as the changes in the initial model are small, e.g. if the discriminative training is initialized with a good ML baseline. This is the typical situation in ASR. Further details and specifics of ASR are discussed in the next section.

4. Automatic Speech Recognition (ASR)

The smooth variants of SVMs introduced in Sec. 3.1 and 3.2 can directly be incorporated into the ASR framework. In this case, the HMM state sequences s_1^T correspond to the classes c. Similar to (Taskar et al., 2003) and (Sha & Saul, 2007), we would like the margin to scale with the length of the speech segments (cf. discussion in Sec. 2). In ASR, a reasonable choice is to set the margin of a sentence to the number of correct phones. More precisely, the simple accuracy $\delta(c,c_n)$ used to represent the margin so far is replaced with the phone accuracy. These approximations directly combine learning theory, HCRFs, and risk-based training of HMMs. Note that Gaussian HMMs (GHMMs) are HCRFs (possibly) with parameter constraints (Heigold et al., 2007).

Typically, MPE is used in combination with the more refined Gaussian regularization centered around Λ_0' (e.g. the maximum likelihood estimate of the generative model), which is comparable with the i-smoothing for GHMMs (Povey & Woodland, 2002). This regularization is combined with the L_2-norm regularization from the SVM

$$J_0^{-1}\|\Lambda\|^2 + J_1^{-1}\|\Lambda-\Lambda_0'\|^2 = J^{-1}\|\Lambda-\Lambda_0\|^2 + \text{const}(\Lambda)$$

with $J^{-1} = J_0^{-1} + J_1^{-1}$ and $\Lambda_0 = \frac{1}{1+\frac{J_1}{J_0}}\Lambda_0'$. Thus, the Gaussian regularization with a properly scaled center Λ_0 (scaling does not change the classification in the maximum approximation) covers the weaker L_2-norm regularization.

Similar to (Heigold et al., 2007), we use n-th order features, e.g. first order features are defined to be $f_{tsd}^{(1st)}(x_1^T, s_1^T) = \delta(s,s_t)x_{td}$. Zeroth and higher order features are defined in a similar fashion. This choice of feature function has the advantage that HCRFs and GHMMs are directly related.

The relationship between SVMs and common training criteria like MMI and MPE allows us to justify some important heuristics typically employed in discriminatively trained ASR systems to achieve good performance: the approximation level γ corresponds to the scaling of the probabilities, i-smoothing is the (refined) regularization term, and the weak unigram language model might be considered an approximation of the margin concept as explained in Sec. 3.1 ("weak prior"). We believe that the frame-based approach proposed to improve the generalization ability is also an attempt to approximate the margin by replacing the context priors (Heigold et al., 2007) by the global relative frequencies.

To apply the existing efficient algorithms, it is important that the margins of the different competing hypotheses can be represented as a weighted transducer sharing the topology with the common lattices, and thus can be integrated into most state-of-the-art systems. This is not always possible in an efficient way for the exact accuracy. Therefore,

approximate accuracies are used. For MPE, an intuitive margin is the approximate phone accuracy (Povey & Woodland, 2002), which is basically the same quantity also used for the loss function[2]. In this case, no additional quantities have to be calculated. The combined acoustic and language model scores are then augmented with these margins by composition. The subsequent steps of the accumulation and estimation remain unchanged. Thus, it is not necessary to modify our transducer-based implementation of the (discriminative) training because the margin can be incorporated by simply configuring an additional composition. The transducer-based implementation also has the advantage that the quantities used for the MMI and MPE accumulation can be represented in terms of *generalized* FB probabilities calculated in *different* semirings. This approach results in the same recursion formulae as used in (Povey & Woodland, 2002), but leads to a unified implementation of the different training criteria. The details on this issue are worked out in the next section.

5. Covariance & Expectation Semiring

In this section, we present an abstraction and generalization of the recursion formulae used for MMI and MPE (Povey & Woodland, 2002). The efficient calculation of the gradient of the objective function is an issue in ASR (and for HCRFs as well) because of the combinatorial number of possible word sequences. The proposed approach unifies these two recursion formulae and extends the speech-specific recursion formula for MPE to HCRFs. As mentioned above, this abstraction is not essential for this work. However, this formalism might be a nice feature of any (probabilistic) transducer library. As an example, it might facilitate the development of more refined training algorithms, e.g. it provides an efficient solution to the unified criterion in (He et al., 2008). The calculation of the gradient under consideration (as probably several other problems in pattern recognition) can be reduced to the calculation of the covariance of two suitably defined random variables, as discussed at the end of this section.

The expectation of the random variable $\mathcal{X}$ w.r.t. the probabilistic transducer $\mathcal{P}$ is defined to be

$$E_{\mathcal{P}}[\mathcal{X}] := \sum_{\pi \in \mathcal{P}} w_{\mathcal{P}}[\pi] w_{\mathcal{X}}[\pi]$$

where $w_{\cdot}[\pi]$ denotes the weight of path π in the respective transducer. The covariance of two (additive) random variables $\mathcal{X}$ and $\mathcal{Y}$ w.r.t. $\mathcal{P}$ is defined to be (with $E_{\mathcal{P}}[\cdot] \equiv E[\cdot]$)

$$\mathrm{Cov}_{\mathcal{P}}(\mathcal{X}, \mathcal{Y}) := \sum_{\pi \in \mathcal{Y}} w_{\mathcal{P}}[\pi] \left(w_{\mathcal{X}}[\pi] - E[\mathcal{X}]\right) \left(w_{\mathcal{Y}}[\pi] - E[\mathcal{Y}]\right).$$

[2] Assume the distance $E[w_1^N, v_1^M]$ between strings w_1^N and v_1^M. Then, the accuracy of string v_1^M given string w_1^N is $A[v_1^M | w_1^N] = N - E[w_1^N, v_1^M]$.

Here, we assume that $\mathcal{P}$, $\mathcal{X}$, and $\mathcal{Y}$ can be represented by acyclic transducers which share the topology, i.e., differ only in the weights. Using these assumptions, we shall show that the covariance can be efficiently calculated by simply exchanging the probability semiring by the expectation semiring in the standard FB algorithm. So, the probability semiring can be used to compute the first order statistics whereas the expectation semiring can be used to compute the second order statistics. It is rather straightforward to define a covariance semiring to calculate third order statistics etc.

We start with introducing the expectation semiring and the abstract definitions which are needed to formulate the propositions.

Expectation semiring. The *expectation semiring* (Eisner, 2001) is a multiplex semiring with weights $(p, v) \in \mathbb{R}^+ \times \mathbb{R}$, and

- $(p_1, v_1) \oplus (p_2, v_2) = (p_1 + p_2, v_1 + v_2)$;
- $(p_1, v_1) \otimes (p_2, v_2) = (p_1 p_2, p_1 v_2 + v_1 p_2)$;
- $\overline{1} = (1, 0)$, $\overline{0} = (0, 0)$.

In addition, the inverse is defined to be $\mathsf{inv}(p, v) = (p^{-1}, -p^{-2}v)$. Observe that the first component corresponds to the probability semiring whereas the second component accounts for the additivity of the random variable. The (partial) path weight of path π is the "product" of the corresponding arc weights $w_{\mathcal{P}}[a]$, $w_{\mathcal{P}}[\pi] = \bigotimes_{a \in \pi} w_{\mathcal{P}}[a]$.

FB potentials. The *forward potential* α_q at the state q of the transducer $\mathcal{P}$ is the sum of the weights of all partial paths π going from the initial state *init* to the state q

$$\alpha_q := \bigoplus_{\pi = (init, q) \in \mathcal{P}} w_{\mathcal{P}}[\pi].$$

These quantities are efficiently calculated by recursion

$$\alpha_{init} = \overline{1} \qquad \alpha_q = \bigoplus_{a = (p,q) \in \mathcal{P}} \alpha_p \otimes w_{\mathcal{P}}[a].$$

The "sum" is over all arcs a of the transducer $\mathcal{P}$ connecting the state p with q. The backward potentials β_q are defined similarly on the transposed $\mathcal{P}$.

Posteriors. The *posterior* transducer $\mathcal{Q}(\mathcal{P})$ associated with the transducer $\mathcal{P}$ has the arc weights

$$w_{\mathcal{Q}(\mathcal{P})}[a] := \left(\bigoplus_{\pi \in \mathcal{P} : a \in \pi} w_{\mathcal{P}}[\pi] \right) \otimes \mathsf{inv}\left(\bigoplus_{\pi \in \mathcal{P}} w_{\mathcal{P}}[\pi] \right).$$

The weight of arc $a = (p, q)$ can be expressed in terms of the above defined forward and backward potentials

$$w_{\mathcal{Q}(\mathcal{P})}[a] = \left(\alpha_p \otimes w_{\mathcal{P}}[a] \otimes \beta_q \right) \otimes \mathsf{inv}(\beta_{init}).$$

Here, we used the fact that β_{init} equals the "normalization constant" in the case of a unique initial state *init*. To make the analogy of the calculation of the expectation and the covariance more clear, we first state the well-known proposition based on the probability semiring.

Proposition 1. *Assume an acyclic transducer $\mathcal{P}$ with probability semiring, and a weighted transducer $\mathcal{X}$ with log semiring. $\mathcal{P}$ and $\mathcal{X}$ share the topology. Then,*

$$E_{\mathcal{P}}[\mathcal{X}] = \sum_{a \in \mathcal{P}} w_{\mathcal{X}}[a] w_{Q(\mathcal{P})}[a].$$

This proposition is then extended to the expectation semiring. Note that for the p-component, we recover the previous proposition.

Proposition 2. *Assume an acyclic transducer $\mathcal{P}$ with probability semiring, and transducers $\mathcal{X}$ and $\mathcal{Y}$ with log semiring. $\mathcal{P}$, $\mathcal{X}$, $\mathcal{Y}$ share the topology. Define the transducer $\mathcal{Z}$ with* expectation semiring *and assign the weights $w_{\mathcal{Z}}[a] = (w_{\mathcal{P}}[a], w_{\mathcal{P}}[a] w_{\mathcal{X}}[a])$ to the arcs. Then,*

$$Cov_{\mathcal{P}}(\mathcal{X}, \mathcal{Y}) = \sum_{a \in \mathcal{Y}} w_{\mathcal{Y}}[a] w_{Q(\mathcal{Z})}[a][v].$$

We conclude this section by showing how the calculation of the gradient of the objective function fits into this framework.

Gradient of objective function. To simplify the discussion, we restrict our consideration to objective functions of the type $\mathcal{F}(\Lambda) = f(E_{\mathcal{P}}[\mathcal{A}])$ rather than using the unified objective function in (He et al., 2008). Here, $\mathcal{P}$ stands for the word lattice with the joint probabilities $p_\Lambda(s_1^T, v_1^M | x_1^T)$ and $\mathcal{A}$ denotes some additive risk (e.g. phone error). In addition, a non-linearity f can be applied to the expectation. Then, building the derivative of this objective function leads to $\nabla \mathcal{F}(\Lambda) = \mathrm{Cov}_{\mathcal{P}}(\mathcal{L}, \nabla \log \mathcal{P})$ with $\mathcal{L} := f'(E_{\mathcal{P}}[\mathcal{A}])\mathcal{A}$. Examples: $\mathcal{A}$ = phone accuracy, $f(x) = x$ (MPE); $\mathcal{A} = \chi_{\mathrm{spk}}$ (characteristic function of spoken sequence, i.e., one for the spoken sequence and zero otherwise), $f(x) = \log x$ (MMI); or $\mathcal{A} = \chi_{\mathrm{spk}}$, $f(x)$ =sigmoid function (MCE).

6. Experimental Results

The presented approaches were evaluated on two different tasks. First, we tested the proposed criterion on the German digit string recognition task Sietill (Heigold et al., 2007), which due to its small size allows for a thorough experimental evaluation. Second, experiments were carried out on the large vocabulary EPPS English task, which represents a realistic ASR task. The baseline MPE result was part of our 2007 TC-STAR evaluation system, which performed best in the restricted and public evaluation conditions for both English and Spanish (Lööf et al., 2007). For completeness, we provide some description of the speech recognition systems. Non-experts, however, can skip these technical parts, keeping in mind that highly competitive systems are used for the discriminative training.

Table 2. Corpus statistics.

Task	Corpus	Data [h]	#run. words [k]	#frames [k]
Sietill	Train	5.5	43	1,980
	Test	5.5	43	1,980
EPPS En	Train	92.0	661	33,120
	Dev06	3.2	27	1,152
	Eval06	3.2	30	1,152
	Eval07	2.9	27	1,044

Our modified MMI criterion is identical with the recently proposed boosted MMI (Povey et al., 2008). These results, however, should be interpreted with some care because in most experiments, the boosting factor is not the only change. Probably, there is a single experiment which is directly comparable with our results on the EPPS task, i.e., which modifies only the boosting factor and which is set upon a state-of-the-art baseline. Very much like our results on the EPPS task, this result supports the hypothesis that the effect of the margin on such systems is marginal.

6.1. Sietill

The recognition system is based on gender-dependent whole-word HMMs. For each gender, 214 distinct states plus one for silence are used. The vocabulary consists of the 11 German digits (including the pronunciation variant 'zwo'). The observation vectors consist of 12 cepstral features without derivatives. The gender-independent Linear Discriminant Analysis (LDA) is applied to 5 consecutive frames and projects the resulting feature vector to 25 dimensions (Heigold et al., 2007). The corpus statistics is summarized in Tab. 2. The ML baseline system uses Gaussian mixtures with globally pooled variances and serves as initialization of the log-linear HMMs. The margin is represented by the approximate word accuracy and has been chosen to be the point where the word error rate (WER) on the training corpus begins to increase rapidly. The final performance turned out to be rather insensitive to the exact value. The optimization was carried out using RProp. Fig. 2 shows the progress of the word error rate (WER) vs. the iteration index on the test corpus. Margin-based MMI was validated on log-linear mixture models of different complexity (16 and 64 densities per HMM state with first order features only) and on a purely log-linear model with second and third order features (instead of using only first order features). The discriminative training was initialized with the respective ML baseline model except for the experiments including third order features. These were initialized with the model from frame-based training (Heigold et al., 2007). The discriminative results were all obtained

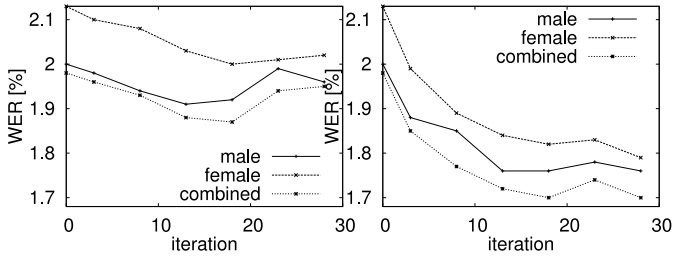

Figure 2. Effect of margin: progress of word error rate (WER) on Sietill test corpus, MMI (left) vs. modified MMI (right) (16 densities/mixture).

Table 3. Word error rates (WER) for Sietill test corpus.

Dns/Mix	Criterion	Margin	WER [%]
16	ML	-	1.98
	MMI	-	1.88
		word	1.72
64	ML	-	1.81
	MMI	-	1.77
		word	1.59
1+f2+3	Frame	word	1.75
	MMI	-	1.68
		word	1.53

using a regularization term. Tab. 3 summarizes the results. The results clearly benefit from the additional margin term, both regarding the performance and the robustness. This might be because the training data are separable for the given configurations. For the experiments using second and third order features ('1+f2+3') the training was initialized with the models from frame-based MMI training which benefits from the margin only slightly (cf. Sec. 4).

6.2. EPPS English

This task contains recordings from the European Parliament Plenary Sessions (EPPS). The corpus statistics of the different EPPS corpora can be found in Tab. 2. The acoustic front end comprises MFCC features augmented by a voicing feature. 9 consecutive frames are concatenated and the resulting vector is projected to 45 dimensions by means of LDA. The MFCC features are warped using a fast variant of the Vocal Tract Length Normalization (VTLN). On top of this, Speaker Adaptive Training (SAT) is applied. The triphones are clustered using CART, resulting in 4,501 generalized triphone states. The HMM states are modeled by Gaussian mixtures with globally pooled variances. The ML baseline system is made up of approximately 900,000 densities. For recognition, a lexicon with 50,000 entries in combination with a 4-gram language model was used (Lööf et al., 2007). The development (Dev06) and evaluation (Eval06) data from the evaluation campaign 2006 as described in Tab. 2 were used to tune the different parameters (e.g. language model scale or the number of MPE iterations). The evaluation data from the evaluation campaign 2007 (Eval07) were used only for testing.

Table 4. Word error rates (WER) for EPPS English corpus, MPE with different margins.

LM (train)	Margin	WER [%] Dev06	Eval06	Eval07
1g	-	13.4	10.1	11.5
	word	13.4	10.2	11.3
	phone	13.3	10.2	11.3
2g	-	13.3	10.3	11.6
	word	13.2	10.2	11.3
	phone	13.2	10.2	11.3

Table 5. Word error rates (WER) for EPPS English corpus, interdependence of weak language model and phone margin.

Crit.	Margin	LM (train)	WER [%] Dev06	Eval06	Eval07
ML	-	-	14.4	10.8	12.0
MPE	no	1g	13.4	10.1	11.5
		2g	13.3	10.3	11.6
	yes	1g	13.3	10.2	11.3
		2g	13.2	10.2	11.3

The word-conditioned lattices used in MPE training were generated with the VTLN/voicedness system in combination with a bigram language model. Since the lattices are dominated by silence and noise arcs, the lattices were filtered. The idea behind this filtering is to correct the posteriors for accumulation of discriminative statistics. For the acoustic rescoring during discriminative training, the exact match approach is used, i.e., the word boundary times are kept fixed.

The margins are tuned on a small fraction of the training corpus such that the margin-based approach in combination with a bigram language model and the standard MPE setup with a unigram language model have the same WER. Independent control experiments imply that no further tuning of the margin parameter is required. In the first experiment we have tested the impact of different margins on the performance, more specifically we have tested the approximate word and phone accuracies according to (Povey & Woodland, 2002). Tab. 4 shows that the differences are marginal. For convenience we decided to use the approximate phone accuracy-based margin for the remaining experiments. In Tab. 5 the interdependence of the weak unigram language model and the margin was investigated. There is ongoing work to clarify the interdependence of the language model used for the optimization and the margin. Using the acoustic model from the standard MPE training, the same 4-gram language model and only each tenth segment, the relative improvement of WER is 5.6% on the training data. This probably indicates that the generalization performance on the test data (Eval07) is not optimal with a relative improvement of 4.2% (and does not appear to be an issue on the development data, i.e., Dev06 and Eval06). The experimental results show the expected tendency, see Tab. 1.

7. Conclusions

We proposed modified formulations of MMI and MPE to include a margin term into the discriminative training of models for ASR. Furthermore, we showed that these modified criteria can directly be used in existing state-of-the-art ASR frameworks, since they can be represented as an additional transducer composition. The modified criteria are directly related to SVMs using a suitable loss function, which allows us to justify some important heuristics used in the discriminative training of acoustic models. The experimental results are consistent with our expectations. For the German digit string recognition task Sietill, where overfitting is achieved after a few iterations, the margin is essential for the robust estimation of the model parameters and allows to achieve significant improvements over the ML baseline. In contrast, on the large vocabulary EPPS English task the observed improvements are transferred well to the test data and the effect under consideration is marginal. So far, we have investigated the effect of the margin for the discriminative re-estimation based on generatively estimated and strongly tuned acoustic models. The benefits due to the margin might be better visible, when the discriminative, margin-based training builds on top of a suboptimal ML baseline. However, models building on top of better baseline models might still have a better absolute performance.

Acknowledgments

This material is partly based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001-06-C-0023, and was partly funded by the European Union under the integrated project TC-STAR (FP6-506738). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the DARPA.

References

Altun, Y., Tsochantaridis, I., & Hofmann, T. (2003). Hidden markov support vector machines. *Int. Conf. on Machine Learning (ICML)*.

Eisner, J. (2001). Expectation semirings: Flexible EM for finite-state transducers. *Finite-State Methods and Natural Language Processing (FSMNLP)*. Helsinki, Finland.

He, X., Deng, L., & Chou, W. (2008). Discriminative learning in sequential pattern recognition – A unifying review for optimization-oriented speech recognition. *IEEE Signal Processing Magazine*.

Heigold, G., Schlüter, R., & Ney, H. (2007). On the equivalence of Gaussian HMM and Gaussian HMM-like hidden conditional random fields. *Proc. of the Int. Conf. on Spoken Language Processing (Interspeech)*. Antwerp, Belgium.

Jebara, T. (2002). *Discriminative, generative, and imitative learning*. Doctoral dissertation, Massachusetts Institute of Technology.

Li, J., Yan, Z., Lee, C., & Wang, R. (2007). A study on soft margin estimation for LVCSR. *Proc. of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. Kyoto, Japan.

Lööf, J., Gollan, C., Hahn, S., Heigold, G., Hoffmeister, B., Plahl, C., Rybach, D., Schlüter, R., & Ney, H. (2007). The RWTH 2007 TC-STAR evaluation system for European English and Spanish. *Proc. of the Int. Conf. on Spoken Language Processing (Interspeech)*. Antwerp, Belgium.

Povey, D., Kanevsky, D., Kingsbury, B., Ramabhadran, B., Saon, G., & Visweswariah, K. (2008). Boosted MMI for model and feature-space discriminative training. *Proc. of the Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*. Las Vegas, NV.

Povey, D., & Woodland, P. C. (2002). Minimum phone error and I-smoothing for improved discriminative training. *Proc. of the Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*. Orlando, FL.

Sha, F., & Saul, L. (2007). Comparison of large margin training to other discriminative methods for phonetic recognition by hidden Markov models. *Proc. of the Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*. Honolulu, Hawaii.

Taskar, B., Guestrin, C., & Koller, D. (2003). Max-margin markov networks. *Neural Information Processing Systems Conference (NIPS)*.

Vapnik, V. (1995). *The nature of statistical learning theory*. Springer.

Weston, J., & Watkins, C. (1999). Support vector machines for multi-class pattern classification. *Proc. of the Seventh European Symposium on Artificial Neural Networks*.

Yin, Y., & Jiang, H. (2007). A fast optimization method for large margin estimation of HMMs based on second order cone programming. *Proc. of the Int. Conf. on Spoken Language Processing (Interspeech)*. Antwerp, Belgium.

Yu, D., Deng, L., He, X., & Acero, A. (2007). Large-margin minimum classification error training for large-scale speech recognition tasks. *Proc. of the Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*. Honolulu, Hawaii, USA.

Zhang, J., Jin, R., Yang, Y., & Hauptmann, A. (2003). Modified logistic regression: An approximation to SVM and its applications in large-scale text categorization. *Int. Conference on Machine Learning (ICML)*.