

# MODIFIED MPE/MMI IN A TRANSDUCER-BASED FRAMEWORK

G. Heigold, R. Schlüter, and H. Ney

Chair of Computer Science 6 - Computer Science Department  
RWTH Aachen University, Aachen  
{heigold,schlueter,ney}@cs.rwth-aachen.de

## ABSTRACT

In this paper we show how common training criteria like for example MPE or MMI can be extended to incorporate a margin term. In addition, a transducer-based training implementation is presented, which covers a large variety of discriminative training criteria for ASR, including the standard MMI, MPE, and MCE criteria, as well as the modifications to these criteria presented here. The modified criteria are directly related with the conventional large margin formulation of SVMs. In the proposed approach, we can take advantage of the generalization guarantees of large margin classifiers while keeping the existing framework for the discriminative training, including the efficient algorithms for conventional MPE or MMI. On the conceptual side, this allows for a *direct* evaluation of the margin term. Finally, experimental results are presented for different large vocabulary continuous speech recognition tasks (one of which is trained on a very large amount of training data) using these modified criteria.

**Index Terms**— training criteria, large margin, weighted finite state transducer, speech recognition

## 1. INTRODUCTION

The parameter estimation problem has two important ingredients: the loss term (*e.g.* phoneme error) and a term to control the model complexity (*e.g.* regularization and margin). The margin concept has proven to be useful for many applications. However, depending on the training/test conditions, we expect different relative importance of the two terms, see Tab. 1.

This work is directly based on our previous work in [1] but presents additional experimental results for Large Vocabulary Continuous Speech Recognition (LVCSR). More explicitly, we would like to experimentally find out the potential of a margin term for LVCSR. The experiments were designed to meet the following objectives. First, a direct evaluation of the margin term. Ideally, we can turn on/off the margin term in the optimization problem. In particular, we want to avoid effects caused by different loss functions, optimization algorithms, model parameterization, convergence speed *etc.* as far as possible. Unfortunately but like for (most) other approaches, the effect of spurious local optima cannot be excluded. Second, the evaluation of the margin on state-of-the-art LVCSR systems. Ideally, we directly improve on the best discriminative criterion. In our case, this is the MPE rather than the MMI criterion. Third, show a clear relationship of our approach to existing large margin classifiers (SVMs). As discussed in [1], existing work on large margin training in ASR is not in complete agreement with these objectives. Finally and as a matter of form, no tuning (*e.g.* the determination of the best training iteration) on the test data, in particular because the margin works on the generalization issue.

To achieve the above-listed objectives, we transform a standard large margin optimization problem including constraints and slack

**Table 1.** Relative importance of loss and margin terms under different training/test conditions.

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

variables into an equivalent optimization problem that resembles conventional training criteria, *e.g.* MMI or MPE. Like for MCE, the loss function is replaced with a smooth approximation [2]. The practical advantage of this approach is that the incorporation of the margin term leads only to small modifications of the conventional training criteria. From the technical point of view, we consider this approach attractive because the same efficient algorithms as for conventional MMI and MPE can be used. Moreover, we even do not need to modify our transducer-based implementation. Keep in mind that the modifications concern only the training criteria and do not affect the underlying model, *i.e.*, the search remains unchanged.

The remainder of the paper is organized as follows. First, the large margin optimization problem is formulated in Sec. 2. Based on the ideas in this section, the conventional training criteria MPE and MMI are modified to incorporate a margin term in Sec. 3. Then, our transducer-based implementation is described in Sec. 4, and used to evaluate the effect of the margin term for LVCSR in Sec. 5. Finally, conclusions are drawn in Sec. 6.

## 2. COMMON LARGE MARGIN CLASSIFIERS

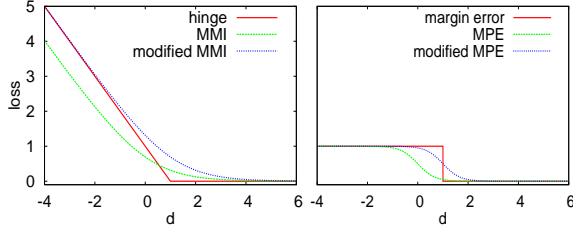
Different formulations of the large margin optimization problem can be found in the literature, *e.g.* [3, 4]. Here, the optimization problem is based on the Hidden Markov SVM introduced in [5] because it best fits our needs. The associated optimization problem is directly stated in the form that best serves our purpose. The reader is referred to the literature for the original formulation. Then, we replace the non-smooth loss function by a smooth approximation such that we can use standard gradient-based algorithms, *e.g.* Extended Baum Welch (EBW) or RProp.

### 2.1. Support Vector Machines (SVMs)

According to [5], for  $C$  classes,  $N$  labeled observations  $(x_n, c_n)$ , and feature functions  $f_i(x, c)$ , the optimization problem of SVMs can be formulated as follows

$$\hat{\Lambda} = \arg \min_{\Lambda} \left\{ \frac{1}{2} \|\Lambda\|^2 + \frac{J}{N} \sum_{n=1}^N l(c_n, d_n; 1) \right\}.$$

For SVMs, the distance vector  $d_n$  has the components  $d_{nc} = \lambda^{\top} (f(x_n, c_n) - f(x_n, c))$ . Note that this parameterization,  $\lambda^{\top} f(x, c)$  also includes (log-)linear models of the type  $\lambda_c^{\top} x$  with  $\Lambda = \{\lambda_c\}$ . The empirical constant  $J > 0$  is used to balance the



**Fig. 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  $d = d_{nc_n}$ .*

margin and the loss terms. The hinge loss function is the typical loss function used in the context of SVMs and is defined as [5]

$$l^{(hinge)}(c_n, d_n; \rho) := \max_{c \neq c_n} \{ \max\{-d_{nc} + \rho, 0\} \}. \quad (1)$$

In this formulation, the margin  $\rho = 1$  is kept fixed and the model parameters  $\Lambda$  are scaled to indirectly maximize the margin. In the ASR experiments below,  $\rho$  shall be set to another (empirical) value to have a better initialization. In alternative formulations, the margin may be directly maximized [4].

## 2.2. Smooth Loss Functions/Limits

As mentioned above, we would like to use gradient-based optimization algorithms, which require differentiable training criteria. For this reason, we define smooth loss functions  $l_\gamma$  with the parameter  $\gamma$  to control the smoothness of the loss function. In addition to the differentiability, we require that  $l_\gamma(\cdot) \xrightarrow{\gamma \rightarrow \infty} l(\cdot)$ , i.e., the smooth loss function converges to the original loss function.

The hinge loss function in Eq. (1) is replaced with the soft-max approximation

$$l_\gamma^{(hinge)}(c_n, d_n; \rho) = \frac{1}{\gamma} \log \left( 1 + \sum_{c \neq c_n} e^{\gamma(-d_{nc} + \rho)} \right).$$

We feel that the weak point about the hinge loss in pattern recognition is that it is unclear how this loss is correlated with the recognition error, which is the typical measure used to evaluate a recognition system eventually. In other words, there is some guarantee regarding the generalization for the hinge loss, but *not* the recognition error which we are actually interested in. In addition, the hinge loss is not robust against outliers because even a single observation can dominate the criterion, see Fig. 1. For these reasons, we prefer the recognition error over the hinge loss. The ideal margin error is defined as

$$l^{(error)}(c_n, d_n; \rho) := E[\hat{c}_n(d_n) | c_n].$$

Here,  $\hat{c}_n$  denotes the recognized class, including the margin  $\rho$ . It is the class that maximizes  $p_\rho(c|x) \propto \exp(\gamma(-d_{nc} - \rho\delta(c, c_n)))$ . In the simplest case,  $E[c|c_n]$  counts the recognition errors,  $1 - \delta(c, c_n)$ . For ASR, a string-based error is used instead, e.g. the phoneme error.

Below, we shall use this loss function in the MPE sense

$$l_\gamma^{(error)}(c_n, d_n; \rho) = \sum_c E[c|c_n] p_\rho(c|x_n).$$

## 3. MODIFIED TRAINING CRITERIA IN ASR

In this section, we apply the results from the last section to ASR. Keep in mind that the SVM formulation is more intuitive for HCRFs

than Gaussian HMMs (GHMMs). The equivalence of GHMMs and Gaussian-like HCRFs [6], however, allows us to apply this approach to GHMMs as well.

In ASR, the data is sequential. A natural extension of the margin from simple to sequential observations is to replace the above 0-1 margin by the string accuracy  $\mathcal{A}(c, c_n)$  between  $c$  and  $c_n$  [7, 4]. This choice makes sure that the SVMs for i.i.d. sequences [5] is consistent with the original SVMs [3] for single observations. Unfortunately, there is no natural choice of accuracy. Two convenient definitions are the approximate accuracy also used for MPE [1, 8] and the frame-based accuracy (Hamming distance) [4, 9]. In the remainder of this section, two variants of the above smooth optimization criterion are introduced and discussed. For this purpose, some additional notation is required.  $W$  denotes a word sequence and  $X$  the sequence of feature vectors  $x_t$ . Furthermore,  $p(W)$  represents the language model and  $p(X|W)$  stands for the acoustic model (without language model scaling factor for simplicity).

### 3.1. Modified MMI

First, a variant of the large margin optimization problem based on the hinge loss is presented. Modified MMI is defined as

$$\mathcal{F}_\gamma^{(MMI)}(\Lambda) = -\frac{1}{2} \|\Lambda\|^2 + \frac{J}{N} \sum_{n=1}^N \frac{1}{\gamma} \log \left( \frac{\{p(W_n) p_\Lambda(X_n|W_n) \exp(-\rho \mathcal{A}(W_n, W_n))\}^\gamma}{\sum_W \{p(W) p_\Lambda(X_n|W) \exp(-\rho \mathcal{A}(W, W_n))\}^\gamma} \right)$$

For HCRFs, the acoustic model is log-linear. For this choice and  $\gamma \rightarrow \infty$ , it can be shown that the optimization problem converges to the SVM optimization problem in Sec. 2.1 using the hinge loss function [1]. In general, the L2-norm regularization is replaced with the stronger regularization  $\|\Lambda - \Lambda_0\|^2$  where  $\Lambda_0$  is some reasonable initial estimate, and is similar to i-smoothing used for GHMMs [1]. Due to the equivalence of GHMMs and Gaussian-like HCRFs [6], we can also use GHMMs for the acoustic model and i-smoothing for regularization [1]. Note that it can be shown that this variant of Modified MMI is theoretically equivalent to the heuristically motivated Boosted MMI [8]. In practice, Boosted MMI differs from Modified MMI in the way the iteration constants are set and the choice of the acoustic model for i-smoothing.

### 3.2. Modified MPE

In contrast to the hinge loss, the recognition error is bounded as illustrated in Fig. 1. This means that a single observation cannot dominate the criterion and thus, is expected to be more robust than the hinge loss.

We define an MPE-like criterion representing a smoothed margin phoneme error with regularization according to Sec. 2.2

$$\mathcal{F}_\gamma^{(MPE)}(\Lambda) = \frac{1}{2} \|\Lambda\|^2 + \frac{J}{N} \sum_{n,W} \frac{E[W|W_n] \{p(W) p_\Lambda(X_n|W) \exp(-\rho \mathcal{A}(W, W_n))\}^\gamma}{\sum_{W'} \{p(W') p_\Lambda(X_n, W') \exp(-\rho \mathcal{A}(W', W_n))\}^\gamma}.$$

Using log-linear acoustic models, it can be shown that this criterion converges to the SVM optimization problem in Sec. 2.1 using the associated margin phoneme error [1]. As for Modified MMI, we can also use GHMMs for the acoustic model and i-smoothing instead [1]. The accuracy  $\mathcal{A}(\cdot, \cdot)$  for the margin may be the same approximate phoneme accuracy as for MPE.

**Justification of MPE heuristics?** Observe that formally, Modified MPE is similar to conventional MPE. Indeed, Modified MPE

gives some new insight into several heuristics typically used for conventional discriminative training: the weak language model can be considered an approximation of the margin term, i-smoothing is a refined regularization term, and the smoothing parameter  $\gamma$  corresponds with the scaling factor for the posteriors.

#### 4. TRANSDUCER-BASED IMPLEMENTATION

In this section, our transducer-based implementation for discriminative training is described. The implementation is based on the finite state automaton library FSA [10]. The outstanding feature of this implementation is that a large class of lattice-based training criteria including standard and Modified MMI/MCE/MPE, share the same algorithms in combination with different semirings. More specifically, the MMI and MPE recursion formulae for GHMMs found in the literature are unified in a single forward/backward (FB) algorithm for acyclic transducers operating on different semirings. The material of this section is not essential for the comprehension of the ideas in the previous sections. However, this generalization will facilitate future research and implementations concerning alternative models (e.g. HCRFs [1], probably also in other domains) or more refined training criteria, cf. the *unified criterion* in [11]. Here, we would like to focus on the different steps required to estimate objective functions of the simple form  $E_{\mathcal{P}}[\mathcal{A}]$  (key quantity of the unified criterion) where  $E$  stands for the expectation of the random variable  $\mathcal{A}$  (some "accuracy") w.r.t. the probability distribution  $\mathcal{P}$ . The optimization of this objective function is conducted with gradient-based techniques. If  $\mathcal{A}$  is the approximate phoneme accuracy and  $\mathcal{P}$  stands for the scaled joint probabilities  $\{p(W)p(X|W)\}^\gamma$ , the MPE criterion is recovered. In the general context of the above-mentioned unified criterion, however,  $\mathcal{A}$  and  $\mathcal{P}$  might represent different quantities, probably without explicit interpretation.

Assume three weighted transducers to represent the word lattices  $\mathcal{P}$ , the accuracies  $\mathcal{A}$ , and the margins  $\mathcal{M}$ . The word boundaries are known and kept fixed. The path weights  $w_{\mathcal{P}}[\pi]$  of the probabilistic transducer  $\mathcal{P}$  with paths  $\pi$  are the joint probabilities  $\{p(W)p(X|W)\}^\gamma$ . In addition, it is assumed that these three transducers are acyclic and share their topology, i.e., differ only in the arc weights. The above choices for  $\mathcal{A}$  and  $\mathcal{M}$  (approximate phoneme accuracy or Hamming distance) satisfy these two constraints, see [1, 8, 9]. The transducer-based objective function reads

$$E_{\mathcal{P}}[\mathcal{A}] = \frac{\sum_{\pi \in \mathcal{P}} w_{\mathcal{P}}[\pi] w_{\mathcal{A}}[\pi]}{\sum_{\pi \in \mathcal{P}} w_{\mathcal{P}}[\pi]}. \quad (2)$$

It can be shown that the gradient of this objective function is the covariance  $Cov_{\mathcal{P}}(\mathcal{A}, \nabla \log \mathcal{P})$  where  $\nabla \log \mathcal{P}$  is the gradient transducer of  $\mathcal{P}$  with path weights  $w_{\nabla \log \mathcal{P}}[\pi] = \nabla \log p(X, W)$ . The covariance is defined in a similar fashion as the expectation in Eq. (2). As shown in [1], the covariance can be efficiently calculated by the standard FB algorithm using the multiplex expectation semiring [12] instead of the probability semiring known from the MMI training. This algorithm provides the posterior transducer  $\mathcal{Q}(\mathcal{Z})$  derived from the transducer  $\mathcal{Z} := (\mathcal{P}, \mathcal{A})$  with the arc weights  $(w_{\mathcal{Z}}[a]_p, w_{\mathcal{Z}}[a]_v) := (w_{\mathcal{P}}[a], w_{\mathcal{P}}[a]w_{\mathcal{A}}[a])$  [1]. With this formalism, we are in the position to state our transducer-based discriminative training in terms of FSA algorithms from [10] in Tab. 2 and Tab. 3. Obviously, MPE and Modified MPE differ only in the composition 'o' of  $\mathcal{P}$  with the scaled margins obtained by arc-wise multiplication of  $w_{\mathcal{M}}[a]$  with the scalar  $\tilde{\rho} := \gamma\rho$ ,  $\text{multiply}(\mathcal{M}, \tilde{\rho})$ . MMI/MPE and Modified MMI/MPE are based on the same FB algorithm for the posteriors but use different semirings:  $\text{posterior}_p(\cdot)$  (probability semiring) vs. ( $v$ -component of)  $\text{posterior}_E(\cdot)$  (multiplex expectation semiring).

**Table 2.** Comparison of MMI, MPE, and Modified MMI/MPE.

	MPE	Modified MPE	Modified MMI	MMI
$\mathcal{P}'$	$\mathcal{P}$	$\mathcal{P} \circ \text{multiply}(\mathcal{M}, \tilde{\rho})$		$\mathcal{P}$
$\mathcal{Z}$	$(\mathcal{P}', \mathcal{A})$		$\mathcal{P}'$	
$\mathcal{Q}$	$\text{posterior}_E(\mathcal{Z})_v$		$\text{posterior}_p(\mathcal{Z})$	

**Table 3.** Transducer-based discriminative training in a nutshell, step 3 is implemented with a Depth First Search. See text for details.

Step	Description
1.	Define $\mathcal{P}'$ and $\mathcal{Z}$ according to Tab. 2.
2.	$\mathcal{Q} = \text{posterior}(\mathcal{Z})$ using respective semiring.
3.	For each arc $a$ and for each frame $t$ : accumulate feature $x_t$ with weight $w_{\mathcal{Q}}[a]$ for state $s_t$ .

#### 5. EXPERIMENTAL RESULTS

The effect of the margin term is tested on three different tasks. Compared with [1], this work provides additional experimental results to confirm our findings for LVCSR there. The German digit string task is used for a few control experiments whereas the other two are LVCSR tasks. In the latter two tasks, the HMM states are modeled by Gaussian densities with globally pooled variances. This allows us to produce rather good ML baselines consisting of a fairly high number of densities, cf. Tab. 4. The language model scale, the best MPE iteration, and the optimal margin parameter  $\rho$  are all tuned on the training and development data [1]. All test data are reserved for the final evaluation of the acoustic models. The same setups as described in [1, 13] are used for the discriminative training. Only for BNBC Cn, we use RProp instead of the EBW algorithm to optimize the GHMMs. The approximate phoneme accuracy is chosen for the margin  $\mathcal{A}$  because experiments have shown that the word error rates are rather insensitive to the choice of the margin [1]. Unless otherwise stated, the scaling factor  $\rho$  is set to 0.5.

##### 5.1. German Digit Strings

In a control experiment, Modified MMI was applied to the Sietill task. This simple digit string recognition task severely suffers from overfitting. 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 digit (including the pronunciation variant 'zwo'). The setup and corpus statistics are described in [1] and summarized in Tab. 4. The ML baseline system uses Gaussian mixtures with globally pooled variances and serves as initialization of the log-linear models. As expected from Tab. 1, the margin term with  $\rho = 25$  helps significantly. For the best log-linear HMM, the ML baseline digit error rate is 1.81% whereas conventional MMI and Modified MMI yield 1.77% and 1.59%, respectively. A similar effect is observed for the log-linear setup using single densities and all  $n$ -th order features up to degree  $n = 3$ . Here, the baseline is the frame-based estimated system with 1.75%. MMI reduces the digit error rate to 1.68%. Again, the best result is achieved with Modified MMI, 1.53%.

##### 5.2. English Parliament Plenary Sessions (EPPS)

This task contains recordings from the European Parliament Plenary Sessions (EPPS). The setup and corpus statistics are described in detail in [1]. A summary of the information can be found in Tab. 4. For recognition, a lexicon with 50k entries in combination with a 4-gram language model is used. Keep in mind that the evaluation data from the evaluation campaign 2007 (Eval07) are used only for testing but not tuning. The results for Modified MPE are summarized

**Table 4.** Corpus statistics and acoustic model. Test data are Eval07 and Eval06 (GALE) for EPPS En and BNBC Cn, respectively.

Task	Train/Test Data [h]	#States/#Densities Features
Sietill	5.5/5.5	430/27k 25 LDA(MFCC)
EPPS En	92/2.9	4,500/830k 45 LDA(MFCC+voicing)+SAT
BNBC Cn 230h	230/2.2	4,500/1,100k 45 LDA(MFCC+tone)+SAT
BNBC Cn 1500h	1500/2.2	4,500/1,200k 45 LDA(PLP+tone)+42 NN+SAT

**Table 5.** Word Error Rate (WER) for EPPS English (Eval07) and BNBC Mandarin (Eval06).

Criterion	WER [%]		
	EPPS En	BNBC Cn	
		230h	1500h
ML	12.0	21.9	17.9
MPE	11.5	20.6	16.5
Modified MPE	11.3	20.3	16.3

in Tab. 5. Refer also to [1] where the interdependence of the weak unigram language model and the margin was investigated.

### 5.3. Mandarin Broadcasts

The second LVCSR task consists of Mandarin broadcasts news and conversations. The experiments are based on the same setup as described in [13]. The corpus statistics of the two systems under consideration are shown in Tab. 4. The BNBC Cn 230h/BNBC Cn 1500h use MFCC/PLP features augmented with a tone feature. In either system, 9 consecutive frames are concatenated and projected to 45 dimensions by means of a Linear Discriminant Analysis (LDA). In addition to these base features, the BNBC Cn 1500h system has 50 Neural Network (NN) based posterior features. The features are warped using Vocal Tract Normalization (VTLN). On top of this, Speaker Adaptive Training (SAT) is applied. The lexicon with 60k entries and the 4-gram language model from [13] are used for recognition. The results for the two different setups are shown in Tab. 5. These results are in agreement with our expectations from Tab. 1 and the results in [1].

## 6. CONCLUSIONS

We modified existing training criteria, *e.g.* MMI, MCE, and MPE to include a margin term. These modified criteria were shown to be closely related with standard large margin classifiers, *i.e.*, SVMs. This approach to large margin optimization has the advantage to fit into our existing transducer-based framework where the training criteria under consideration mainly differ in the choice of the semiring (probability *vs.* expectation semiring), and allows for a direct evaluation of the margin term. As expected, Modified MMI clearly outperforms conventional MMI (up to 75% of the discriminative improvement comes from the margin) on a simple digit string recognition task where overfitting is an issue. For LVCSR, the additional margin term leads to more limited but nevertheless consistent improvements (less than 25% of the discriminative improvements) compared with standard MPE, even for a very large amount of training data. Reasons for this outcome might be that the loss term dominates for LVCSR as illustrated in Tab. 1, and that the margin concept is already well approximated by several heuristics (*e.g.* weak language model) used for conventional discriminative training.

## 7. 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.

## 8. REFERENCES

- [1] G. Heigold, T. Deselaers, R. Schlüter, and H. Ney, "Modified MMI/MPE: A direct evaluation of the margin in speech recognition," in *Int. Conf. on Machine Learning ICML*, Helsinki, Finland, July 2008, pp. 384–391.
- [2] J. Zhang, R. Jin, Y. Yang, and A.G. Hauptmann, "Modified logistic regression: An approximation to SVM and its applications in large-scale text categorization," in *Int. Conf. on Machine Learning ICML*, 2003.
- [3] V. Vapnik, *The nature of statistical learning theory*, Springer-Verlag, 1995.
- [4] F. Sha and L.K. Saul, "Comparison of large margin training to other discriminative methods for phonetic recognition by hidden Markov models," in *Proc. of the Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Honolulu, Hawaii, Apr. 2007.
- [5] Y. Altun, I. Tsochantaridis, and T. Hofmann, "Hidden Markov support vector machines," in *Int. Conf. on Machine Learning ICML*, 2003.
- [6] G. Heigold, P. Lehnen, R. Schlüter, and H. Ney, "On the equivalence of Gaussian and log-linear HMMs," in *Proc. of the Int. Conf. on Spoken Language Processing (Interspeech)*, Brisbane, Australia, Sept. 2008.
- [7] B. Taskar, C. Guestrin, and D. Koller, "Max-margin Markov networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2003.
- [8] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, "Boosted MMI for model and feature-space discriminative training," in *Proc. of the Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, NV, USA, 2008.
- [9] G. Saon and D. Povey, "Penalty function maximization for large margin HMM training," in *Proc. of the Int. Conf. on Spoken Language Processing (Interspeech)*, Brisbane, Australia, Sept. 2008.
- [10] S. Kanthak and H. Ney, "FSA: An efficient and flexible C++ toolkit for finite state automata using on-demand computation," in *Ann. Meeting of the Ass. for Computational Linguistics (ACL)*, Barcelona, Spain, July 2004, pp. 510–517.
- [11] W. Macherey, L. Haferkamp, R. Schlüter, and H. Ney, "Investigations on error minimizing training criteria for discriminative training in automatic speech recognition," in *Proc. of the Int. Conf. on Spoken Language Processing (Interspeech)*, Sept. 2005, pp. 2133–2136.
- [12] J. Eisner, "Expectation semirings: Flexible EM for finite-state transducers," Helsinki, Finland, Aug. 2001.
- [13] B. Hoffmeister, C. Plahl, P. Fritz, G. Heigold, J. Löff, R. Schlüter, and H. Ney, "Development of the 2007 rwth mandarin lvcsr system," in *IEEE Automatic Speech Recognition and Understanding Workshop*, Kyoto, Japan, Dec. 2007.