Posterior-Scaled MPE: Novel Discriminative Training Criteria

Markus Nussbaum-Thom [1], *Zoltán Tüske* [1], *Georg Heigold* [2], *Ralf Schlüter* [1], *Hermann Ney* [1]

[1]Computer Science Dept. 6, RWTH Aachen University, Aachen, Germany
[2]Google Research, Mountain View, CA, USA
{nussbaum, tuske, schlueter, ney}@cs.rwth-aachen.de, heigold@google.com

Abstract

We recently discovered novel discriminative training criteria following a principled approach. In this approach training criteria are developed from error bounds on the global error for pattern classification tasks that depend on non-trivial loss functions. Automatic speech recognition (ASR) is a prominent example for such a task depending on the non-trivial Levenshtein loss. In this context, the posterior-scaled Minimum Phoneme Error (MPE) training criterion, which is the state-of-the-art discriminative training criterion in ASR, was shown to be an approximation to one of the novel criteria.

Here, we describe the implementation of the posterior-scaled MPE criterion in a transducer-based framework, and compare this criterion to other discriminative training criteria on an ASR task. This comparison indicates that the posterior-scaled MPE criterion performs better than other discriminative criteria including MPE.

Index Terms: error bounds, discriminative training criteria, margin, MPE

1. Introduction

In ASR, discriminative training criteria [1, 2, 3] like Maximum Mutual Information (MMI), Minimum Classification Error (MCE) and MPE are commonly used to enhance the performance of the maximum-likelihood (ML) trained Gaussian mixture models (GMM). In statistical pattern recognition, a model-based posterior $p_\Lambda(c|x)$ is learned with parameters Λ from the training samples $(x_n, c_n)_{n=1}^N$ (i.e. class $c_n \in C$ is measured for observation $x_n \in \mathcal{X} \subseteq R^D$).

In [4] we derived novel discriminative training criteria from error bounds on the mismatch between the global error of the model-based decision rule and the Bayes error for non-trivial loss functions — the Weighted MMI (WMMI)

$$F^{(\text{WMMI})}(\Lambda) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c\in C}\mathcal{A}(c_n,c)\log p_\Lambda(c|x_n) \quad (1)$$

and the Weighted Squared Error criterion (WSE), involving accuracy $\mathcal{A} : C \times C \to \mathbb{R}$. At the same time, the posterior-scaled MPE criterion (κ-MPE)

$$F^{(\kappa\text{-MPE})}(\Lambda) = \frac{1}{N}\sum_{n=1}^{N}\sum_{c\in C}\mathcal{A}(c_n,c)p_\Lambda(c|x_n)^\kappa$$

was derived as an approximation to the WMMI criterion. This links the MPE criterion to error bounds for the first time, since for $\kappa = 1.0$ this is identical to the MPE criterion. A more detailed description of the error bound derivation and the relation between the WMMI and κ-MPE criteria will be given in Section 3.

For mathematical reasons, the considered non-trivial loss functions $\mathcal{L} : C \times C \to \mathbb{R}$ are assumed to be defined by accuracies $\mathcal{A} : C \times C \to \mathbb{R}$ that fulfill the requirements (with length value $|d|$):

$$\mathcal{L}(d,c) = |d| - \mathcal{A}(d,c)$$
$$\forall c,d : \mathcal{A}(d,c) \geq 0 \quad (2)$$
$$\forall c : \mathcal{A}(c,c) \geq 1 \quad (3)$$

In this work, the implementation of the κ-MPE criterion in a transducer-based framework via a modification of the Eisner semiring [5] is discussed. Furthermore, experiments are performed on an ASR task, comparing the MMI and the novel WMMI, WSE and κ-MPE discriminative training criteria.

2. Bayes Decision Theory

In this section, the foundations of Bayes decision theory will be summarized briefly. Bayes decision theory involving loss function $\mathcal{L}$ provides an optimal choice among all decision rules $r : \mathcal{X} \to C$: By minimizing the local error

$$E_\mathcal{L}(c|x) = \sum_{k\in C} pr(k|x)\mathcal{L}(k,c)$$

in the Bayes decision rule

$$c_\mathcal{L}(x) = \operatorname*{argmin}_{c\in C}\{E_\mathcal{L}(c|x)\} \quad (4)$$

the global error

$$E_\mathcal{L}(r) = \int pr(x)E_\mathcal{L}(r(x)|x)\,\mathrm{d}x \quad (5)$$

is minimized, given the joint true distribution $pr(x, c)$ of observations and classes. The minimum error $E_{\mathcal{L}}(c_{\mathcal{L}})$ is called the Bayes error. The terminology "true" symbolizes the distribution of observations and classes in nature, which is unknown in real-world applications. In [8] the Bayes decision rule optimization problem for non-trivial loss functions is shown to be NP-complete, therefore computationally expensive. In practice the efficient computable 0-1-loss model-based Bayes decision rule is used instead, which from now on will be referred to as the model-based decision rule:

$$c_{0\text{-}1}^{\Lambda}(x) = \underset{c \in C}{\operatorname{argmax}} \{p_{\Lambda}(c|x)\}$$

The next section gives a short overview of the derivation of the novel training criteria from error bounds.

3. On Error Bounds and Discriminative Training Criteria

In this section, the recent findings in [4] are summarized to give a better understanding of the derivation of the novel training criteria.

3.1. Error Bounds

In [4] upper bounds $F(\Lambda)$ are established on the mismatch

$$\Delta_{\mathcal{L}}^2 := \left[E_{\mathcal{L}}(c_{0\text{-}1}^{\Lambda}) - E_{\mathcal{L}}(c_{\mathcal{L}})\right]^2 \leq F(\Lambda) \tag{6}$$

between the global error of the model-based Bayes decision rule and the Bayes error. In order to reduce this mismatch, the parameters are chosen to minimize $F(\Lambda)$:

$$\widehat{\Lambda} = \underset{\Lambda}{\operatorname{argmin}} \{F(\Lambda)\} \tag{7}$$

Based on the definition of the local reward

$$R_{\mathcal{A}}(c|x) = \sum_{k \in C} pr(k|x)\mathcal{A}(k, c)$$

the Bayes decision rule can be reformulated by:

$$c_{\mathcal{A}}(x) := \underset{c \in C}{\operatorname{argmax}} \{R_{\mathcal{A}}(c|x)\} \tag{8}$$

Analogous, the true reward posterior,

$$qr(c|x) = \frac{R_{\mathcal{A}}(c|x)}{\sum_{k \in C} R_{\mathcal{A}}(k|x)}$$

which is well-defined since the accuracy is assumed to be positive (2), is used to formulate the 0-1-loss Bayes decision rule based on the true reward posterior $qr(c|x)$:

$$c_{0\text{-}1}^{qr}(x) := \underset{c \in C}{\operatorname{argmax}} \{qr(c|x)\} \tag{9}$$

On basis of these definitions, we have shown in [4] that decision rule (9), based on the true reward posterior, is equivalent to the Bayes decision rule (4):

$$\begin{aligned} c_{\mathcal{L}}(x) &= c_{\mathcal{A}}(x) \\ &= \underset{c \in C}{\operatorname{argmax}} \{qr(c|x)\} \\ &= c_{0\text{-}1}^{qr}(x) \end{aligned}$$

This identity suggests that the local and global error based on the true reward posterior $qr(c|x)$

$$\begin{aligned} E_{0\text{-}1}^{qr}(c|x) &= (1 - qr(c|x)) \\ E_{0\text{-}1}^{qr}(r) &= \int pr(x) E_{0\text{-}1}^{qr}(r(x)|x)\, \mathrm{d}x \end{aligned}$$

can be considered instead of the global error in (5) based on $pr(c|x)$. This also yields the 0-1-loss based mismatch (6) based on $qr(c|x)$:

$$\left[\Delta_{0\text{-}1}^{qr}\right]^2 := \left[E_{0\text{-}1}^{qr}(c_{0\text{-}1}^{\Lambda}) - E_{0\text{-}1}^{qr}(c_{0\text{-}1}^{qr})\right]^2 \tag{10}$$

Furthermore, for upper bounds $G(\Lambda) \geq \left[\Delta_{0\text{-}1}^{qr}\right]^2$ the following relationship is deduced in [4]

$$\Delta_{\mathcal{L}}^4 \leq \alpha^2 \left[\Delta_{0\text{-}1}^{qr}\right]^2 \leq \alpha^2 G(\Lambda)$$

with $Z(x) := \sum_{k \in C} R_{\mathcal{A}}(k|x)$ and $\alpha := \int pr(x) Z^2(x)\, \mathrm{d}x$. Similar to the findings in [6] for $pr(c|x)$, the Kullback-Leibler bound is such an upper bound,

$$\left[\Delta_{0\text{-}1}^{qr}\right]^2 \leq -2 \int pr(x) \sum_{c \in C} qr(c|x) \log \frac{p_{\Lambda}(c|x)}{qr(c|x)}\, \mathrm{d}x \tag{11}$$

which is derived for the true reward posterior $qr(c|x)$ instead of $pr(c|x)$ in this case.

In the next section, the derivation of the WMMI criterion from the Kullback-Leibler bound will be described briefly.

3.2. Empirical Training Criteria

Training criteria are obtained from bound (11) in a few further steps. These derivations can only be described briefly here, but are covered in all detail in [4]. First, bound (11) is simplified by dropping all terms independent to Λ in optimization (7). Second, the derived simplified bound is widened by substituting the denominator of $qr(c|x)$ with a lower bound. Third, the true distribution is replaced by the empirical distribution on the training samples using the discrete Kronecker delta and the continuous valued Dirac delta. Finally, the WMMI criterion (1) is derived using the sifting property [9] of the Dirac delta and by dropping all terms independent to Λ in the optimization (7).

Similarly to the derivation of WMMI criterion, the WSE criterion

$$F^{(\mathrm{WSE})}(\Lambda) = \frac{1}{N} \sum_{n=1}^{N} \sum_{c \in C} \mathcal{A}(c_n, c) \sum_{k \in C} \left[p_{\Lambda}(k|x_n) - \delta(k, c)\right]^2$$

was derived in [4] from bounds on the mismatch (10).

The WMMI and WSE criteria have some sound properties: Both criteria (as well as the corresponding error bounds) result in the MMI and squared error criteria (as well as the Kullback-Leibler and squared error bounds [6, pp.641]) for $A(c,d) = \delta(c,d)$ and therefore, they are generalizations to non-trivial loss functions. In case of infinite training data, the WMMI and WSE criteria minimize the global error based on the related loss function. Like in [3, 7], modified criteria can be derived from the WMMI, WSE and κ-MPE criteria by augmenting the joint distribution with a margin term. Furthermore, this connection induces an interpretation as smooth approximation to support vector machine loss functions. The κ-MPE criterion can be derived from the WMMI criterion using the power approximation of the logarithm $\log u = \lim_{\kappa \to 0} \left(u^\kappa - 1\right)/\kappa$.

The next section describes the implementation of the WMMI and κ-MPE criteria in a transducer-based framework.

4. Application to ASR

For the purpose of ASR, the classes c correspond to word sequences $w_1^N = (w_n)_{n=1}^N$. In addition, Hidden Markov Model (HMM) state sequences $s_1^T = (s_t)_{t=1}^T$ and features $x_1^T = (x_t)_{t=1}^T$ are used. The margin accuracy should scale with the length of the utterance but at the same time the requirements of positiveness (2) and boundedness (3) have to be met. We choose the frame-wise phone accuracy [10] to fulfill all requirements. The model-based posterior involving HMMs and the margin term is formulated by

$$p_{\Lambda,\gamma\rho}(w_1^N|x_1^T) = \sum_{s_1^T : w_1^N} \frac{\left[p_\Lambda(x_1^T, s_1^T, w_1^N) \exp\left(\rho \mathcal{A}\left(w_1^N, w_1^N\right)\right)\right]^\gamma}{Z_{\Lambda,\gamma\rho}\left(x_1^T\right)}$$

with the posterior renormalization $Z_{\Lambda,\gamma\rho}\left(x_1^T\right)$. Given the vector representation of feature functions $f(x_1^T, s_1^T)$ and parameter set $\Lambda = \{\lambda\}$, the choice $p_\Lambda(x_1^T, s_1^T, w_1^N) = p(w_1^N) \exp\left(\lambda^T f(x_1^T, s_1^T)\right)$ results in a log-linear posterior model. In order to avoid overfitting, an ℓ_2 regularization is used, centered around the generative maximum likelihood Gaussian mixture model. Only first order features $f_{t,s,d}(x_1^T, s_1^T) = \delta(s, s_t) x_{t,d}$ are used in combination with zeroth order features defined similarly. The training criteria are optimized using the gradient-based procedure using RPROP in a transducer-based framework.

4.1. Transducer-Based Framework Implementation

In this section, the efficient calculation of the gradient of the κ-MPE criterion is discussed by extending the Eisner expectation semiring [5]. All criteria have been implemented in the transducer-based framework of [3, 7].

We start with an extension of the expectation semiring which then is used to calculate the gradient of the κ-MPE and WMMI criterion.

κ-scaled expectation semiring. The κ-scaled expectation semiring is a multiplex semiring with weights $(p, v) \in \mathbb{R}^+ \times \mathbb{R}$, and

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

In addition, the inverse is defined to be $\mathrm{inv}\,(p, v) = \left(p^{-1}, -v p^{-2\kappa}\right)$. In terms of notations and definitions, we stick to [7]. In particular, the expectation transducer is denoted by $E_\mathcal{P}[Z]$ w.r.t. random variable Z and the probabilistic transducer $\mathcal{P}$. The expectation can be calculated by forward potentials α_q and backward potentials β_q at the state q (here q_0 denotes the initial state).

Gradient of the objective function. Let $\mathcal{P}$ be the word lattice with the joint probabilities $p_\Lambda\left(x_1^T, s_1^T, w_1^N\right)$. This is an acyclic transducer with probability semiring. The transducer $\mathcal{A}$ is the accuracy transducer corresponding to $\mathcal{P}$, having the same topology but different weights. Define transducer Z with the κ-scaled expectation semiring and assign the weights $w_Z[a] = \left(w_\mathcal{P}[a], w_\mathcal{P}^\kappa[a]\, w_\mathcal{A}[a]\right)$ to the arcs. Then, the gradient of the κ-MPE criterion can be calculated by:

$$\nabla F^{(\kappa\text{-MPE})}(\Lambda) = \sum_{e \in \mathcal{P}} w_{\nabla \log \mathcal{P}}[e] \cdot \left(\frac{w_{E_\mathcal{P}[Z]}[e]\,[v]}{\beta_{q_0}^\kappa[p]} - \frac{w_{E_\mathcal{P}[Z]}[e]\,[p] \beta_{q_0}[v]}{\beta_{q_0}^{1+\kappa}[p]} \right)$$

Furthermore, the choice $\kappa = 0$ results in the gradient of the WMMI criterion.

In the next section, the experimental results are discussed.

5. Experimental Results

The modified criteria are tested on the TIMIT phone recognition task. The corpus statistics are presented in Table 1.

Table 1: *Corpus statistics.*

task	corpus	data [h]	#run. words	frames
TIMIT	Train	3.14	30 k	1 M
	Test	0.16	1,5 k	100 k

The acoustic front-end is comprised of 16 Mel-Frequency Cepstral Coefficient (MFCC) features. Feature vectors from nine consecutive frames are concatenated and a Linear Discriminant Analysis (LDA) is used to reduce the dimension to 33.

The word-conditioned lattices for discriminative training were generated with the baseline Gaussian Mixture Model (GMM) system using a unigram language

model, acting as a weak margin. The language model scale γ during training was tuned on a hold-out set and kept fixed during training. The log-linear models were initialized with the ML model and trained with a gradient descent method using the RPROP algorithm. The regularization and margin were chosen to the point where the word error rate (WER) started to increase rapidly.

The GHMM baseline recognition system for TIMIT uses 114 monophone states plus one silence state, phoneme folding [11] is used. The emission probabilities are modeled by GMM with a total of about 20 k densities, all sharing a single diagonal covariance matrix. A trigram phoneme language model was used for recognition.

Table 2: *WER[%] results for the ML and modified criteria on the* TIMIT *test set.*

criterion					
ML	MMI	MPE	WSE	WMMI	0.6-MPE
32.0	30.6	30.6	31.2	31.7	30.1

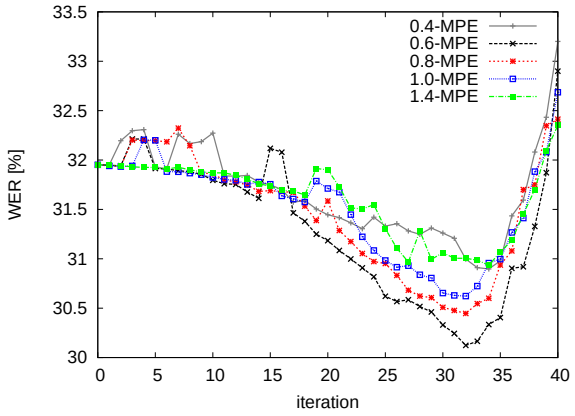

Figure 1: *WER progress of the κ-MPE criterion for different κ values on the TIMIT test set.*

Briefly, the experimental results in Table 2 indicate that the novel modified training criteria WMMI and WSE give some improvement but perform worse than the modified MMI and MPE criteria on TIMIT, which show a relative improvement of 5%. At the same time, the modified 0.6-MPE criterion performs even better than the modified MPE criterion with a relative improvement of 6%. These results could be expected since MPE is identical to the 1.0-MPE criterion, which is a numerical stable approximation of the WMMI criterion. This is also confirmed by the WER progress of the κ-MPE criteria in Figure 1: The 0.6-MPE and 0.8-MPE criteria perform noticeable better than MPE. However, the 0.4-MPE criterion performs worse, probably because $\kappa = 0.4$ is closer to the numerical more unstable WMMI criterion with $\kappa = 0$.

The next section concludes the paper.

6. Conclusion

We introduced novel error bounds on the global error mismatch. These bounds yield novel discriminative training criteria based on non-trivial loss functions, like the Levenshtein distance in the case of ASR. In addition, the posterior-scaled MPE criterion, which is an approximation to one of the proposed criteria, was presented. This connection to global error bounds gives a better theoretical justification for the good performance of the MPE criterion for the first time.

To investigate the performance of the posterior-scaled MPE criterion, the corresponding implementation was discussed in a transducer-based framework via a small modification of the Eisner expectation semiring. Experiments were performed comparing the posterior-scaled MPE criterion to other discriminative training criteria on an ASR task. Theoretical results were confirmed and they show that the posterior-scaled MPE criterion performs better than the state-of-the-art MPE criterion.

Acknowledgements – This work was partly realized under the QUAERO Programme, funded by OSEO, French State agency for innovation.

7. References

[1] R. Schlüter, “Investigations on Discriminative Training Criteria”, PhD thesis, RWTH Aachen University, Aachen, Germany, September, 2000.

[2] E. McDermott, and S. Watanabe and A. Nakamura, “Margin-space integration of MPE loss via differencing of MMI functionals for generalized error-weighted discriminative training”, Interspeech, pp. 224-227, 2009.

[3] G. Heigold, “A Log-Linear Discriminative Modeling Framework for Speech Recognition”, PhD thesis, RWTH Aachen University, Aachen, Germany, June, 2010.

[4] M. Nussbaum-Thom, G. Heigold, R. Schlüter, H. Ney, “On the Relation of Loss-Based Error Bounds to Discriminative Training Criteria”, submitted to the Neural Information Processing Systems (NIPS) Conference, 2012. (Available by personal communication in the case of rejection.)

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

[6] H. Ney, “On the Relationship between Classification Error Bounds and Training Criteria in Statistical Pattern Recognition”, Iberian Conference on Pattern Recognition and Image Analysis IbPRIA, Puerto de Andratx, Spain, 636-645, June, 2003.

[7] G. Heigold, T. Deselaers, R. Schlüter and H. Ney, “Modified MMI/MPE: A Direct Evaluation of the Margin in Speech Recognition”, ICML, pp. 384-391, Helsinki, Finland, July, 2008.

[8] M. Nussbaum-Thom, S. Wiesler, G. Heigold, R. Schlüter, H. Ney, “Novel Error Bounds and Discriminative Training Criteria”, submitted to Interspeech 2012.

[9] R. Bracewell, “The Sifting Property”, In The Fourier Transform and Its Applications, 3rd ed. New York, McGraw-Hill, pp. 74-77, 1999.

[10] M. Gibson, “Minimum Bayes risk acoustic model estimation and adaptation”, PhD thesis, University of Sheffield, Sheffield, UK, 2008.

[11] K.-F. Lee and H.-W. Hon,“Speaker-independent phone recognition using hidden Markov models”,IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 37, no. 11, pp. 1641-1648, 1989.