

# Pinched Lattice Minimum Bayes Risk Discriminative Training for Large Vocabulary Continuous Speech Recognition

Vlasios Doumptiotis, William Byrne

Center for Language and Speech Processing  
Johns Hopkins University, Baltimore, MD 21218, USA  
{vlasios,byrne}@jhu.edu

## Abstract

Iterative estimation procedures that minimize empirical risk based on general loss functions such as the Levenshtein distance have been derived as extensions of the Extended Baum Welch algorithm. While reducing expected loss on training data is a desirable training criterion, these algorithms can be difficult to apply. They are unlike MMI estimation in that they require an explicit listing of the hypotheses to be considered and in complex problems such lists tend to be prohibitively large. To overcome this difficulty, modeling techniques originally developed to improve search efficiency in Minimum Bayes Risk decoding can be used to transform these estimation algorithms so that exact update, risk minimization procedures can be used for complex recognition problems. Experimental results in two large vocabulary speech recognition tasks show improvements over conventionally trained MMIE models.

## 1. Introduction

Discriminative estimation procedures such as MMI [1] are powerful acoustic modeling techniques that, while computationally expensive, are efficient enough to be used for large vocabulary speech recognition. This efficiency derives from the existence of lattice-based estimation algorithms [2]. In practice, lattice-based algorithms are necessary if a modeling approach that considers multiple alternative hypotheses is to be applied to large vocabulary speech recognition. However, there are discriminative training procedures which, unlike MMI, do not have readily derived lattice-based estimation procedures. In this paper we explore how a such discriminative training algorithm intended to minimize an overall risk criterion can be applied to large vocabulary speech recognizer by using techniques originally developed to improve Minimum Bayes Risk decoding [3, 4]. We will show that merging the two techniques yields an efficient and effective estimation procedure that can be used to obtain additive performance improvements over MMI for large vocabulary speech recognition.

## 2. Risk-Based Discriminative Training

Risk based parameter estimation procedures attempt to minimize the expected risk over a training set. Given a transcribed acoustic training set  $(\bar{W}, O)$  the estimation objective is

$$\operatorname{argmin}_{\theta} \sum_{W' \in \mathcal{W}} l(\bar{W}, W') P(W'|O; \theta). \quad (1)$$

The goal is to minimize the empirical loss by reducing the likelihood of any competing hypotheses  $W'$  that are far from the truth  $\bar{W}$  under the loss function. Throughout this paper  $l(\bar{W}, W')$  is the Levenshtein distance related to Word Error Rate.

An iterative estimation procedure for minimizing this objective function has been developed by Kaiser *et al.* [5]. They applied the Extended Baum Welch algorithm to obtain a risk-minimizing variant of the MMI re-estimation procedure for the parameters of state-dependent Gaussian observation distributions [1]. The reestimation equations for the Gaussian means and variances  $\{\mu_s, \Sigma_s\}$  are

$$\bar{\mu}_s = \frac{\sum_{W' \in \mathcal{W}} K(W') p_s(W') + D_s \mu_s}{\sum_{W' \in \mathcal{W}} K(W') \gamma_s^{W'} + D_s} \quad (2)$$

$$\tilde{\Sigma}_s = \frac{\sum_{W' \in \mathcal{W}} K(W') q_s(W') + D_s \tilde{\Sigma}_s}{\sum_{W' \in \mathcal{W}} K(W') \gamma_s(W') + D_s} - \bar{\mu}_s^2, \quad (3)$$

where  $K(W')$  is defined as

$$\left[ \sum_{W'' \in \mathcal{W}} P(W''|O) l(\bar{W}, W'') - l(\bar{W}, W') \right] P(W'|O),$$

$\tilde{\Sigma}_s = \Sigma_s + \mu_s \mu_s^T$ , and the usual HMM sufficient statistics computed wrt each hypothesis  $W'$  are  $\gamma_s(W') = \sum_{\tau} \gamma_s(\tau; W')$ ,  $p_s(W') = \sum_{\tau} \gamma_s(\tau; W') o_{\tau}$ , and  $q_s(W') = \sum_{\tau} \gamma_s(\tau; W') o_{\tau}^2$ .

All quantities in the update relationships above are dependent on the set of competing hypotheses  $\mathcal{W}$ .  $K(W')$  clearly depends on  $\mathcal{W}$ , as does the posterior distribution over the competing hypotheses

$$P(W'|O) = \frac{P(O|W')P(W')}{\sum_{W'' \in \mathcal{W}} P(O|W'')P(W'')}. \quad (4)$$

## 2.1. Computational Issues

In large vocabulary speech recognition tasks,  $\mathcal{W}$  is often a lattice generated by the ASR decoder. Lattices are used because the most likely hypotheses are so numerous that listing them explicitly is impractical. However probabilities such as Equation 4 can be found by summing over lattice paths so that procedures such as lattice-based MMI are feasible [2].

The minimum risk re-estimation procedure of Equations 2 and 3 are not as readily realized over lattices. For instance, the quantity  $K(W', \mathcal{W})$  requires finding the Levenshtein distance between the reference  $\bar{W}$  and every other path  $W'$  in  $\mathcal{W}$ . These distances are not as easily computed as path likelihoods, since Levenshtein distance between two strings does not distribute over lattice arcs in the manner of path likelihoods. One possibility is simply to expand first-pass ASR lattices into N-Best lists so that the string-to-string comparisons can be carried out explicitly. These N-Best lists would have to be extremely deep to contain a significant portion of the most likely hypotheses, and the computation of loss over them would also be costly. As an alternative, we investigate algorithms developed within MBR decoding to overcome similar difficulties.

## 2.2. Computation of Risk for MBR Decoding

MBR decoders [4] find the sentence hypothesis with the least expected error under a loss function as

$$\hat{W} = \operatorname{argmin}_{W \in \mathcal{W}} \sum_{W' \in \mathcal{W}} l(W, W') P(W'|O).$$

Conceptually there are two distinct steps (although they can be combined [4]). First, the risks are computed:

$$E(W; \mathcal{W}) = \sum_{W' \in \mathcal{W}} l(W, W') P(W'|O) \quad \forall W \in \mathcal{W}$$

and then there is a search:  $\hat{W} = \operatorname{argmin}_{W \in \mathcal{W}} E(W; \mathcal{W})$ .

Efficient algorithms have been developed to compute the risk  $E(W; \mathcal{W})$  of a hypothesis  $W$  under the Levenshtein loss function [6]. We can thus find the discriminative training objective function of Equation 1 as  $E(\bar{W}; \mathcal{W})$ . The key is to find  $l(\bar{W}, W')$  for all  $W'$  in a lattice  $\mathcal{W}$ . This yields an (nearly) optimum alignment of every  $W'$  to  $\bar{W}$  called the *lattice-to-string alignment*.

This alignment makes it possible to segment  $\mathcal{W}$  into a series of sublattices as outlined in Figure 1. We first align each original lattice to the correct hypothesis to obtain ‘pinched’ lattices. The pinched lattice is a sequence of sublattices, each aligned to a single word in the reference transcription. These sublattices provide alternative hypotheses to the reference words. We chose not keep all these alternatives, and we discard many of the sublattices by pruning them back to the truth (as described in Sec. 3). The result is a greatly reduced hypothesis space  $\tilde{\mathcal{W}}$  derived from the original lattice  $\mathcal{W}$ .

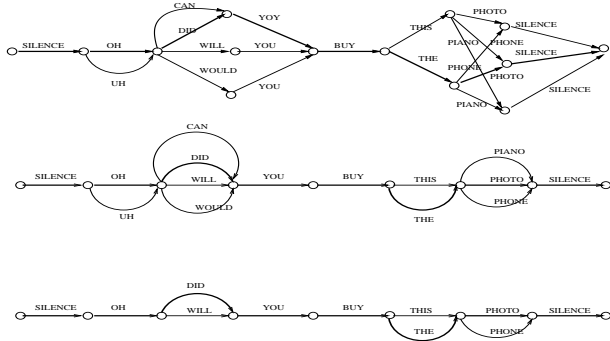


Figure 1: Lattice Segmentation for MBR Training. *Top*: First-pass lattice of likely sentence hypotheses with correct path in bold; *Middle*: Alignment of lattice paths to correct path; *Bottom*: Refined hypothesis space  $\tilde{\mathcal{W}}$  used for Minimum Bayes Risk Discriminative Training.

Our original motivation was to speed up MBR search, but this approach also allows us to redefine the string-to-string loss within  $\tilde{\mathcal{W}}$ . Suppose the reference string  $\bar{W}$  has  $N$  words  $\bar{W}_1 \dots \bar{W}_N$ . Another string  $W' \in \tilde{\mathcal{W}}$  is not allowed to be aligned arbitrarily to  $\bar{W}$ ; it must follow the constraints of  $\tilde{\mathcal{W}}$ . We call this the *induced loss function*

$$l_I(\bar{W}, W') = \sum_{i=1}^N l(\bar{W}_i, W'_i)$$

where  $W'_i$  is the substring of  $W'$  that aligns with the  $i^{th}$  word of  $\bar{W}$ .  $l_I(\bar{W}, W')$  is an easy to compute approximation to  $l(\bar{W}, W') \quad \forall W' \in \tilde{\mathcal{W}}$ .

In summary, lattice segmentation produces both a reduced hypothesis space as well as an induced loss function. We next discuss how to use these quantities to reduce the computational cost of discriminative training.

## 2.3. Pinched Lattice MBRDT

Our approach to large vocabulary minimum Bayes risk discriminative training is simply to incorporate our method of selecting hypothesis lists into the MBRDT estimation procedure of Kaiser *et al.* [5]. We call this procedure Pinched Lattice Minimum Bayes Risk Discriminative Training (PLMBRDT).

If we restrict the hypothesis space to the pinched lattice, the Minimum Bayes Risk Discriminative Training objective under the induced loss function becomes

$$\operatorname{argmin}_{\theta} \sum_{W' \in \tilde{\mathcal{W}}} \sum_{i=1}^N l(\bar{W}_i, W'_i) P(W'|O; \theta). \quad (5)$$

In this way we can reduce the hypothesis space significantly so that the original formulation by Kaiser *et al.* [5] can be applied to large vocabulary ASR, albeit under an approximate loss function.

Estimation is via Equations 2 and 3, with  $\mathcal{W}$  replaced by  $\tilde{\mathcal{W}}$ , computing  $P(W'|O)$  over  $\tilde{\mathcal{W}}$ , and taking

$K(W', \tilde{\mathcal{W}})$  as

$$[ \sum_{W'' \in \tilde{\mathcal{W}}} P(W''|O) l_I(\bar{W}, W'') - l_I(\bar{W}, W') ] P(W'|O),$$

This calculation requires expanding the pinched lattices into lists of hypotheses. However we have control over the degree of lattice pinching and thus also over the depths of the hypothesis lists.

#### 2.4. ‘One-Worst’ PLMBRDT

We can obtain further computational savings by restricting the hypothesis space to only two hypotheses. We first identify the ‘worst’ hypothesis within the pinched lattice

$$W^* = \operatorname{argmax}_{W' \in \tilde{\mathcal{W}}} l_I(\bar{W}, W').$$

We then restrict  $\mathcal{W}$  to the truth and this one competitor:  $\tilde{\mathcal{W}} = \{\bar{W}, W^*\}$ . If  $\tilde{\mathcal{W}}$  has  $n$  segments, then  $l_I(\bar{W}, W^*) = n$ , and Equation 5 simplifies to

$$\operatorname{argmin}_{\theta} \sum_{W' \in \{\bar{W}, W^*\}} \sum_{i=1}^N l(\bar{W}_i, W'_i) P(W'|O; \theta)$$

which is  $\operatorname{argmin}_{\theta} n P(W^*|O; \theta)$ . It follows that

$$K_{\bar{W}} P(O|\bar{W}) = P(\bar{W}|O) n P(W^*|O) \\ K_{W^*} P(O|W^*) = -P(\bar{W}|O) n P(W^*|O)$$

and  $P(O) = P(O|\bar{W})P(\bar{W}) + P(O|W^*)P(W^*)$ . The update rule for  $\bar{\mu}_s$  becomes

$$\frac{P(\bar{W}|O) n P(W^*|O) (p_s(\bar{W}) - p_s(W^*)) + D_s \bar{\mu}_s}{P(\bar{W}|O) n P(W^*|O) (\gamma_s(\bar{W}) - \gamma_s(W^*)) + D_s}.$$

A final, rather brutal, approximation is to discard the terms  $P(\bar{W}|O) n P(W^*|O)$ . This yields a simple corrective training update procedure that we term ‘One-Worst PLMBRDT’

$$\bar{\mu}_s = \frac{p_s(\bar{W}) - p_s(W^*) + D_s \bar{\mu}_s}{\gamma_s(\bar{W}) - \gamma_s(W^*) + D_s} \\ \bar{\Sigma}_s = \frac{q_s(\bar{W}) - q_s(W^*) + D_s \bar{\Sigma}_s}{\gamma_s(\bar{W}) - \gamma_s(W^*) + D_s} - \bar{\mu}_s^2.$$

What distinguishes this approach from other forms of correct training is not the update procedure itself, but rather the way in which the competing hypothesis is obtained.

### 3. Experimental Procedures and Results

We will present results on two large vocabulary speech recognition systems. The first system is trained and evaluated in the SWITCHBOARD conversational English domain and the second system is trained and evaluated in the MALACH spontaneous Czech domain [7]. Both

systems are speaker independent continuous mixture density, tied state, cross-word, gender-independent, using tri-phone HMMs trained by HTK. The AT&T Large Vocabulary Decoder was used to generate lattices over the training and test sets.

The SWITCHBOARD training set consisted of 16.4 hours of SWITCHBOARD-1 and 0.5 hour of CallHome English data, with 22580 utterances in total. The speech was parameterized into 39-dimensional, PLP cepstral coefficients, with delta and acceleration coefficients. The SWITCHBOARD recognition tests were carried out on a subset of the 2000 Hub-5 SWITCHBOARD-1 evaluation set (SWBD1) with 866 utterances and the 1998 Hub-5 SWITCHBOARD-2 evaluation set (SWBD2) with 913 utterances (approx. 2 hours of speech, total). The MALACH Czech baseline acoustic models were built from 40 hours of data with 24065 utterances. The speech was parameterized into 39-dimensional, MFCC coefficients, with delta and acceleration coefficients. The test set consisted of 954 utterances selected from held-out speakers (approx. 2 hours of speech).

The SWITCHBOARD language model was a back-off trigram with a 33K word vocabulary, while the MALACH language model was a back-off bigram with a 83K word vocabulary.

Lattice-based MMI [2] was performed in each domain. The SWITCHBOARD lattices were generated once and the link posteriors were fixed for three iterations of MMI. In MALACH, the link posteriors were reestimated after each of six MMI iterations.

#### 3.1. PLMBRDT Training Steps

The identification of binary word confusion pairs followed the procedures developed for small vocabulary tasks [8]. We first generate lattices over the acoustic training set using MMIE models. We then align the lattices to the reference transcription under the Levenshtein distance and then pinch them as depicted in Figure 1.

This produces a very large number of ‘confusable’ pairs and in these experiments we focused only on the most frequently observed pairs. We identify those confusion pairs that are observed more than 75 times in the SWITCHBOARD training data and more than 100 times in MALACH. The less frequently occurring pairs are discarded. As an example, suppose that the pair {PHONE, PHOTO} was observed less than 75 times in the pinched training set lattices. In each observed instance, the link corresponding to the incorrect word hypotheses (PHONE) would be discarded and only the single link corresponding to the correct word (PHOTO) would be retained. This reduces the number of different types of binary confusions in SWITCHBOARD from 31467 to 159 and from 25847 to 117 in MALACH. This corresponds to a rate of 0.3 and 0.17 confusion pair per correct word in SWITCHBOARD and MALACH, resp;

Iter	PLMBRDT		OneWorst PLMBRDT	
	SWBD1	SWBD2	SWBD1	SWBD2
1	39.6	49.7	39.6	49.7
2	39.5	49.4	39.2	49.4
3	39.4	49.7	39.8	49.8

Table 1: PLMBRDT Performance on SWITCHBOARD (WER%). On the SWBD1/SWBD2 sets, the ML WER is 41.1% / 51.1% and the MMI WER is 39.9% / 49.7%.

i.e. in SWITCHBOARD, there is a binary confusion for roughly every third word.

We observed that due to this aggressive filtering, many training set lattices are reduced to a single word sequence, i.e. the reference transcription. These utterances do not contribute to the overall training criterion and they are therefore removed from the PLMBRDT training data. The MALACH training set is reduced from 24,065 to 15,436 utterances, and the SWITCHBOARD training set is reduced from 22,580 to 15,741 utterances. We found that after filtering the average number of binary confusion pairs in each pinched training set lattice is 2.1 in SWITCHBOARD and 3.2 in MALACH. Hypothesis lists are then generated from these pinched lattices, resulting in an average transcription list depth of 13.1 in SWITCHBOARD and 36.5 in MALACH. The PLMBRDT calculations of Equations 2 and 3 are carried out over these lists of hypotheses, and the hypothesis needed for the ‘One Worst’ algorithm is also extracted from them. Both algorithms are then be carried out as in Sections 2.3 and 2.4.

### 3.2. PLMBRDT Performance

The models trained by PLMBRDT and its ‘One Worst’ variant are used with the baseline language models to rescore the test sets. Results are reported in Tables 1 and 2, the latter with MLLR adaptation. Since the MALACH Czech ASR task is not widely studied, we also report p-values (in parentheses) which give the probability that there is no performance difference between each system and the MMIE system. Even though PLMBRDT attempts to correct a much smaller set of hypotheses than MMI, both versions of PLMBRDT give improvements relative to the lattice-MMIE baseline. The performance of the OneWorst approach in particular suggests that, even though sparse, the sets of competing hypotheses identified by lattice pinching can be used for discriminative training.

## 4. Conclusion

Lattice pinching techniques developed to find suitable search spaces for Minimum Bayes Risk decoding can also be used to generate competing hypotheses for discriminative training. The induced loss function defines a training objective under which an exact parameter update procedure can be obtained. The result is an iterative estimation procedure that minimizes an approximate loss function and is efficient enough to be applied to discrim-

Iter	PLMBRDT	OneWorst PLMBRDT
1	41.4 (0.114)	41.3 (0.107)
2	41.3 (0.038)	41.2 (0.042)
3	41.3 (0.112)	41.0 (0.003)
4	41.3 (0.001)	41.1 (0.052)
5	41.1 (0.031)	—
6	41.0 (0.013)	—

Table 2: PLMBRDT Performance on MALACH (WER%). The ML WER is 44.3% and the MMI WER is 41.5%. p-values wrt the MMI baseline are in parens.

inative estimation of large vocabulary continuous speech systems. The experiments reported here are conservative and early explorations of possible modeling approaches. Due to the the use of heavily pinched lattices and the cautious selection of confusion sets, these experiments focused on a small collection of recognition errors. PLMBRDT nevertheless improves well trained MMI systems and there is the expectation of further gains through more aggressive modeling.

**Acknowledgements** Thanks to M. Mohri of AT&T for the AT&T decoder and FSM libraries, and to S. Kumar of CLSP.

## 5. References

- [1] Y. Normandin, *Hidden Markov Models, Maximum Mutual Information, and the Speech Recognition Problem*, Ph.D. thesis, McGill University, 1991.
- [2] P. C. Woodland and D. Povey, “Large scale discriminative training for speech recognition,” in *Proc. ITRW ASR*. ISCA, 2000.
- [3] A. Stolcke, Y. Konig, and M. Weintraub, “Explicit word error minimization in N-Best list rescoring,” in *Proc. EUROSPEECH*, 1997.
- [4] V. Goel and W. Byrne, “Minimum Bayes-Risk automatic speech recognition,” *Computer Speech and Language*, vol. 14(2), 2000.
- [5] J. Kaiser, B. Horvat, and Z. Kacic, “A novel loss function for the overall risk criterion based discriminative training of HMM models,” in *Proc. ICSLP*, 2000.
- [6] V. Goel, S. Kumar, and W. Byrne, “Segmental minimum Bayes-risk decoding for automatic speech recognition,” *IEEE Trans. Speech and Audio Proc.*, To Appear.
- [7] W. Byrne *et al.*, “Automatic recognition of spontaneous speech for access to multilingual oral history archives,” *IEEE Trans. Speech and Audio Proc.*, July 2004.
- [8] V. Doumptiotis, S. Tsakalidis, and W. Byrne, “Lattice Segmentation and Minimum Bayes Risk Discriminative Training,” in *Proc. EUROSPEECH*, 2003.