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

Vlasios Doumpiotis

*Escription Incorporated*
*Needham, MA 02494 U.S.A.*

William Byrne *

*Cambridge University Engineering Department*
*Trumpington Street, Cambridge, CB2 1PZ, U.K.*

**Abstract**

Lattice segmentation techniques developed for Minimum Bayes Risk decoding in large vocabulary speech recognition tasks are used to compute the statistics needed for discriminative training algorithms that estimate HMM parameters so as to reduce the overall risk over the training data. New estimation procedures are developed and evaluated for both small and large vocabulary recognition tasks, and additive performance improvements are shown relative to maximum mutual information estimation. These relative gains are explained through a detailed analysis of individual word recognition errors.

*Key words:* Discriminative training, maximum mutual information (MMI) estimation, acoustic modeling, minimum Bayes risk decoding, risk minimization, large vocabulary speech recognition, lattice segmentation.

## 1 Introduction

Discriminative acoustic modeling procedures, such as maximum mutual information (MMI) estimation (Normandin (1996)), are powerful modeling tech-

⋆ Funded by National Science Foundation ITR No. IIS-9982329.

* Corresponding author. Tel.: +44 (0)1223 332651; fax: +44 (0)1223 332662.
*Email addresses:* vlasios@jhu.edu (Vlasios Doumpiotis), wjb31@cam.ac.uk (William Byrne).

niques that can be used to improve the performance of speech recognition systems that are created initially using Maximum Likelihood Estimation (MLE) algorithms. MMI is often motivated as an estimation procedure by observing that it increases the *a posteriori* probability of the correct transcription of the speech in the training set. This defines MMI as a parameter estimation procedure, but the overall value of incorporating MMI in system building is the reduction of words recognized incorrectly on an unseen test set.

Since the ultimate goal is to reduce the number of words in error, which we will define as the loss, estimation procedures that reduce loss rather than improve likelihood offer a modeling approach to improve systems under a criterion that is closely linked to overall ASR performance. One such risk-based parameter estimation procedure was developed by Kaiser et al. (2000, 2002) to reduce the expected loss, or risk, over the training set. Their approach is a generalization of MMI in that both are derived via the Extended Baum Welch algorithm (Gopalakrishnan et al. (1991)), and MMI is special case of risk minimization under the sentence error loss function.

The risk-based estimation algorithm of Kaiser et al. (2000, 2002) is not suited for direct application to large vocabulary speech recognition tasks. The difficulty arises from the need to compute the risk over many alternative hypotheses to obtain reliable statistical estimates. In small vocabulary tasks, N-Best lists are adequate to represent the space of hypotheses. However word lattices are needed in large vocabulary tasks. While lattice algorithms have been developed to compute the statistics needed for likelihood-based estimation procedures such as MMI (Woodland and Povey (2000)), risk-based estimation algorithms are not as easily formulated over lattices. The problem is that loss and likelihood are not computed in the same way. The lattice structures that make likelihood calculation easy and efficient do not help with the computation of risk.

The focus of this paper is the efficient computation of loss and likelihood in risk-based parameter estimation for large vocabulary speech recognition. We use lattice-cutting techniques developed for Minimum Bayes Risk decoding (Goel et al. (2001); Kumar and Byrne (2002); Goel et al. (2004)) to compute efficiently the statistics needed by the algorithm of Kaiser et al. (2000, 2002). We will show that the two techniques can be merged through a proper definition of the estimation problem to yield efficient and effective estimation procedures that can be used to obtain additive performance improvements over MMI for large vocabulary speech recognition.

The paper proceeds as follows. In Section 2 we discuss the relationships between prior work in minimum Bayes risk (MBR) decoding and in discriminative parameter estimation. As will be described, the efficacy of MBR decoding for large vocabulary speech recognition depends on the efficient computation of risk over large ASR lattices, and we suggest how risk-based lattice segmentation, developed for efficient MBR decoding, can be used to compute the statistics needed in MBR parameter estimation. In Section 3 we present Pinched Lattice Minimum Bayes Risk Discriminative Training procedures based on risk-based lattice segmentation. The general PLMBRT algorithm is presented first in Section 3, after which two PLMBRDT variants are presented in Sections 3.1 and 3.2: the first is intended for use with whole-word acoustic models and the second incorporates very aggressive lattice pruning. In Sections 4 and 5 we compare PLMBRDT variants to MMI: in Section 4 we study small vocabulary ASR recognition behavior and show that PLMBRDT can resolve errors in ways that the standard MMI does not, and in Section 5 we further show that PLMBRDT can be used to improve MMI acoustic models for large vocabulary ASR tasks.

## 2 Minimum Bayes Risk Discriminative Training

Risk-based parameter estimation procedures attempt to minimize the expected risk over the training set. Given a transcribed database $\{\bar{W},O\}$, the estimation objective is to find the optimum model parameters that minimize the expected risk

$$\theta^* = \operatorname*{argmin}_{\theta} R(\bar{W}, \mathcal{W}; \theta) \tag{1}$$

where

$$R(\bar{W}, \mathcal{W}; \theta) = \sum_{W' \in \mathcal{W}} l(\bar{W}, W') P(W'|O; \theta) \, . \tag{2}$$

$\mathcal{W}$ is taken to be a set of hypotheses being considered as alternatives to the truth $\bar{W}$, and we assume that their distance to the correct transcription $\bar{W}$ is measured by the string edit or Levenstein distance $l(\bar{W}, W)$ associated with the Word Error Rate (WER).

The estimation problem hinges on determining the contribution to the overall risk of each hypothesis $W'$ in $\mathcal{W}$. If a relatively likely hypothesis $W'$ differs significantly from $\bar{W}$ as measured by $l(\bar{W}, W')$, it will add substantially to the overall risk. Thus a successful estimation strategy is one that moves probability mass towards those hypotheses that are close to the reference while reducing the likelihood of those hypotheses that are far away. While the loss function

$l(\bar{W}, W')$ and the likelihood $P(W'|O;\theta)$ dominate the overall risk, $\mathcal{W}$ also plays an important role. Since the risk is measured over $\mathcal{W}$, it must provide a representative sample of hypotheses that are both likely and error-full. If $\mathcal{W}$ is not chosen well, the risk measurements will be biased. In particular there is a danger in having $\mathcal{W}$ too small and underestimating the risk.

### *2.1 Iterative Risk Minimization via the Extended Baum Welch Algorithm*

Kaiser et al. (2000, 2002) have shown how the Extended Baum Welch (Gopalakrishnan et al. (1991)) algorithm can be applied to obtain a risk-minimizing variant of the MMI re-estimation procedure for the parameters of state-dependent Gaussian observation distributions. The well-known MMI estimation equations for the means and variances of HMM Gaussian observation distributions are (Normandin (1996))

$$\bar{\mu}_s = \frac{\sum_\tau \gamma'_s(\tau;\ \theta) o(\tau) + D_s \mu_s}{\sum_\tau \gamma'_s(\tau;\ \theta) + D_s} \tag{3}$$

$$\bar{\Sigma}_s = \frac{\sum_\tau \gamma'_s(\tau;\ \theta) o(\tau)^2 + D_s \left(\Sigma_s + {\mu_s}^2\right)}{\sum_\tau \gamma'_s(\tau;\ \theta) + D_s} - \bar{\mu}_s^2 \tag{4}$$

where $\gamma'_s(\tau;\theta) = \gamma_s(\tau;\theta) - \gamma^g_s(\tau;\theta)$ ; $\gamma_s(\tau;\theta) = q_{s_\tau}(s|\bar{w}_0^{K-1}, o_1^{\hat{l}};\theta)$ is the conditional occupancy probability of state $s$ at time $\tau$ given the training acoustic observation vector sequence $o_0^{\hat{l}}$ and the reference $\bar{W} = \bar{w}_0^{K-1}$ ; and $\gamma^g_s(\tau;\theta) = q_{s_\tau}(s|o_1^{\hat{l}};\theta)$ is the conditional occupancy probability of state $s$ at time $\tau$ given only the training acoustic data. Taken together, the parameters of the state observation distribution are $\theta = \{\mu_s, \Sigma_s\}$.

The effect of MMI is that the new parameters improve the posterior distribution of the reference transcription: $P(\bar{W}|O;\bar{\theta}) \geq P(\bar{W}|O;\theta)$. By observing that the overall risk $R(\bar{W}, \mathcal{W};\theta)$ is a rational function similar to the posterior probability $P(W|O)$, Kaiser et al. (2000, 2002) derived the following MMI-variant to reduce the overall risk, i.e. so that $R(\bar{W}, \mathcal{W};\bar{\theta}) \leq R(\bar{W}, \mathcal{W};\theta)$ :

$$\bar{\mu}_s = \frac{\sum\limits_{W' \in \mathcal{W}} K(W', \mathcal{W};\theta) \sum_\tau \gamma_s(\tau; W') o(\tau) + D_s \mu_s}{\sum\limits_{W' \in \mathcal{W}} K(W', \mathcal{W};\theta) \sum_\tau \gamma_s(\tau; W') + D_s} \tag{5}$$

$$\bar{\Sigma}_s = \frac{\sum\limits_{W' \in \mathcal{W}} K(W', \mathcal{W};\theta) \sum_\tau \gamma_s(\tau; W') o(\tau)^2 + D_s \left(\Sigma_s + \mu_s {\mu_s}^T\right)}{\sum\limits_{W' \in \mathcal{W}} K(W', \mathcal{W};\theta) \sum_\tau \gamma_s(\tau; W') + D_s} - \bar{\mu}_s \bar{\mu}_s^T \tag{6}$$

where

$$K(W', \mathcal{W};\theta) = \left[ \sum_{W'' \in \mathcal{W}} P(W''|O;\theta) l(\bar{W}, W'') - l(\bar{W}, W') \right] P(W'|O;\theta)\ . \tag{7}$$

The quantity $K(W', \mathcal{W}; \theta)$ determines the contribution of each hypothesis $W'$ to the gradient of the loss. It plays the same role in the Extended Baum Welch update rule as the gradient of the likelihood does in the derivation of MMI:

$$-\nabla_\theta R(\bar{W}, \mathcal{W}; \theta) = \sum_{W' \in \mathcal{W}} K(W', \mathcal{W}; \theta) \, \nabla_\theta \log P(O|W'; \theta) \, . \tag{8}$$

Note that all the quantities in the above update relationships depend on the set of competing hypotheses $\mathcal{W}$. $K(W', \mathcal{W}; \theta)$ clearly depends on $\mathcal{W}$, as does the posterior distribution over the competing hypotheses:

$$P(W'|O) = \frac{P(O|W')P(W')}{\sum_{W'' \in \mathcal{W}} P(O|W'')P(W'')} \, . \tag{9}$$

We refer to $\mathcal{W}$ as the *evidence space* since it specifies the hypotheses over which the risk will be estimated.

### *2.2 Computing Statistics Over the Evidence Space*

In large vocabulary speech recognition tasks, $\mathcal{W}$ is often a lattice generated by the ASR decoder (Woodland and Povey (2000)). Lattices are used because the most likely hypotheses are so numerous that listing them explicitly is impractical. Through the conditional independence assumptions underlying the ASR system, quantities such as Equation (9) can be found by summing over the lattice arcs so that estimation procedures such as lattice-based MMI are feasible. Dense lattices are needed to obtain robust and unbiased estimates of the quantities needed in Equations 3 and 4.

However the risk minimizing estimation procedure of Equations 5, 6, and 7 is not readily realized over lattices. The source of the problem is the presence of the term $l(\bar{W}, W')$ in the quantity $K(W', \mathcal{W}; \theta)$; this term must be found for all $W' \in \mathcal{W}$. If $l(\bar{W}, W')$ was a likelihood based quantity, it could be easily computed over the ASR lattice. However, a separate trellis must be constructed for the dynamic programming (DP) alignment of each string $W'$ to $\bar{W}$ (Sankoff and Kruskal (1983)). This trellis is not consistent with the structure of ASR lattices; the latter reflect the conditional independence assumptions underlying the ASR models and not the requirements of the DP alignment. As a result, the quantity $K(W', \mathcal{W}; \theta)$ must be computed and maintained for each path $W' \in \mathcal{W}$. Beyond the computational difficulties in finding the distances for all these $W'$, there are complications in the computation of the mean and variance updates of Equations 5 and 6. The summation over $\mathcal{W}$ must be performed path-wise by explicitly enumerating all the hypotheses $W'$ so that the terms $K(W', \mathcal{W}; \theta)$ can be incorporated correctly into the statistics.

Given the formulation thus far, the only possibility for lattice-based estimation is simply to expand the lattices into N-Best lists so that the string-to-string comparisons and the gathering of statistics (done via the Forward-Backward procedure over each $W' \in \mathcal{W}$) can be carried out exactly. This is the approach proposed, investigated, and validated by Kaiser et al. (2000, 2002). It is effective for tasks for which N-Best lists can be created that contain a significant portion of the likely hypotheses. While correct, this approach is not feasible for large vocabulary continuous speech recognition tasks in which these N-Best lists would have to be extremely deep to contain the most likely hypotheses. As the depth of the N-Best lists increases, performing the Levenshtein alignments in order to compute the loss $l(\bar{W}, W')$ for all paths $W' \in \mathcal{W}$ becomes overwhelmingly expensive, as does the gathering of statistics needed to perform the parameter updates of Equations 5 and 6.

This problem of merging the computation of loss and likelihood also arises in the application of Minimum Bayes Risk decoding to large vocabulary ASR tasks (Goel et al. (2001); Kumar and Byrne (2002)). We next discuss how efficient techniques to compute risk over lattices can be used to obtain the statistics needed to implement the risk-based MMI variants for parameter estimation in large vocabulary speech recognition tasks.

### *2.2.1 Efficient Computation of Risk in MBR Decoding*

Minimum Bayes Risk decoders (Goel et al. (2001); Kumar and Byrne (2002)) find a sentence hypothesis with the least expected error under a loss function as

$$\hat{W} = \underset{W \in \mathcal{W}}{\operatorname{argmin}} \sum_{W' \in \mathcal{W}} l(W, W') P(W'|O; \theta) \tag{10}$$

This is essentially a large search problem in which $\mathcal{W}$ are N-Best lists or lattices that incorporate $P(W'|O)$ as a posterior distribution on word strings, typically obtained using an HMM acoustic model and an N-gram language model (Stolcke et al. (1997); Goel and Byrne (2000)).

In implementing an MBR decoder, there are conceptually two distinct steps (which can be combined for efficiency (Goel and Byrne (2000)):
**Step 1** For each $W \in \mathcal{W}$, find its risk :

$$R(W,\ \mathcal{W};\ \theta) = \sum_{W' \in \mathcal{W}} l(W, W')\ P(W'|O; \theta) \tag{11}$$

**Step 2** Select the minimum risk hypothesis :

$$\hat{W} = \underset{W \in \mathcal{W}}{\operatorname{argmin}}\, R(W,\ \mathcal{W};\ \theta)\ . \tag{12}$$

Efficient algorithms have been developed to compute the risk $R(W,\ \mathcal{W}\ ; \theta)$

of a hypothesis $W$ under the Levenshtein loss function (Goel et al. (2004)). Since it is straightforward to compute $P(W'|O;\theta)$ over lattices, the key is an efficient *lattice-to-string alignment* algorithm to find $l(W, W')$ for all $W'$ in any lattice $\mathcal{W}$. Such an algorithm has been developed, and it yields the (nearly) optimum alignment of every $W'$ to $\bar{W}$.

The lattice-to-string alignment algorithm is described in detail in Goel et al. (2004). For the purposes of this paper the algorithm can be summarized pictorially. The top lattice in Figure 1 shows a lattice generated by an ASR system. The lattice arcs are labelled by word hypotheses and these arcs carry the negative log likelihood of each word. In this example, the lattice will be aligned to the reference string $\bar{W}$: HELLO HOW ARE YOU ALL TODAY; it appears in the lattice marked in bold. The output of the lattice-to-string alignment algorithm is a lattice itself, as shown in the second lattice of Figure 1. The alignment of an arbitrary string from the first lattice, e.g.
$W'$ : WELL O NOW ARE YOU ALL TODAY
can be read from the corresponding string in the alignment lattice:
WELL.INS:0/1 O:0/1 NOW:1/1 ARE:2/0 YOU:3/0 ALL:4/0 TODAY:5/0
The notation WELL.INS:0/1 indicates that WELL is aligned as an insertion, with a cost of 1, to the word at position 0 in the reference string (which is HELLO - the alignment index starts at 0). Similarly, O is also aligned to HELLO and NOW is aligned to HOW, each with a substitution cost of 1. The overall alignment between the lattice path and the reference is

| Index $i$: | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Reference $\bar{W}_i$ : | HELLO | HOW | ARE | YOU | ALL | TODAY |
| $W'_i$ : | WELL O | NOW | ARE | YOU | ALL | TODAY |
| Per Segment Cost : | 2 | 1 | 0 | 0 | 0 | 0 |

with a total loss of $l(\bar{W}, W') = \sum_i l(\bar{W}_i, W'_i) = 3$. By tracing a path through the lattice and accumulating the Levenshtein alignment costs and weighting them by the arc likelihoods (which are preserved from the original ASR output lattice), the risk $R(\bar{W}, \mathcal{W};\, \theta)$ of $\bar{W}$ can be computed.

The connection between MBR decoding and the Minimum Risk estimation algorithm of Equations 5 and 6 becomes apparent when we note that a key quantity in the risk-based estimation procedure can be written in terms of the same lattice-based risk $R(\bar{W}, \mathcal{W};\, \theta)$ needed for MBR decoding :

$$K(W', \mathcal{W}; \theta) = [R(\bar{W}, \mathcal{W}; \theta) - l(\bar{W}, W')]P(W'|O;\theta)\,. \tag{13}$$

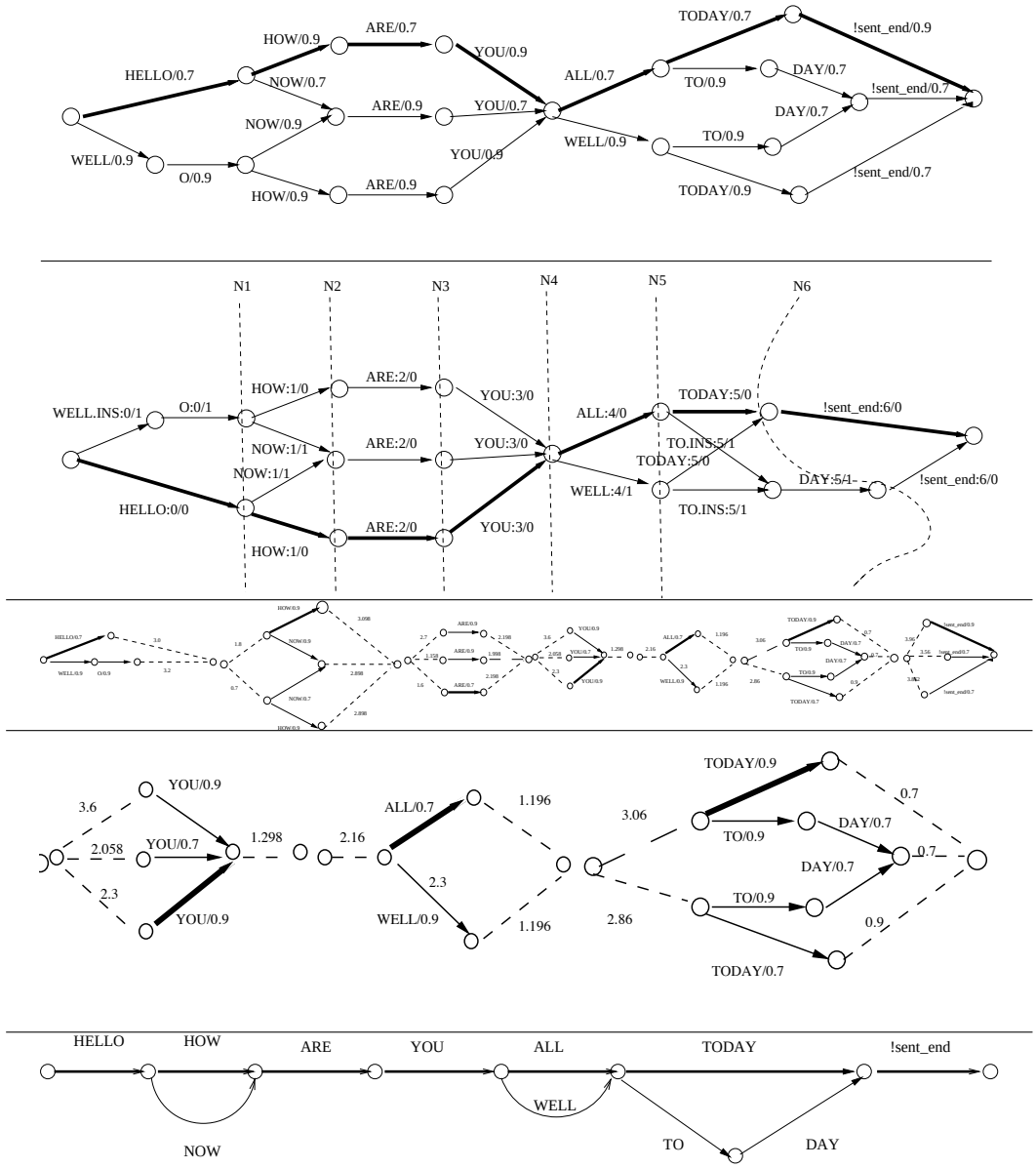


Fig. 1. Alignment of lattice to the reference string and the pinching and pruning operations that produce a lattice with confusion pairs. From top: original lattice with reference path in bold; aligned lattice with node cut sets; pinched lattice; a blown up segment of the pinched lattice spanning node cut sets N3 to N6; the final pinched and pruned lattice.

### *2.2.2 Risk-Based Pruning of the Evidence Space*

Even with the aid of the lattice-to-string alignment algorithm, computing lattice-based risk can be computationally challenging for long or deep lattices. We have developed risk-based lattice segmentation techniques to simplify MBR decoding and we now discuss how these methods can be applied to risk-based parameter estimation.

Risk-based lattice segmentation proceeds by segmenting the lattice with respect to the reference string by following the lattice-to-string alignments. For the $K$ words of the reference string $\bar{W}$, we identify K-1 *node cut sets*. To form the cut set N$i$ , $i = 1, \ldots, K-1$
- identify all lattice subpaths that are aligned to the reference word $\bar{W}_{i-1}$
- the cut set N$i$ consists of the final lattice nodes of all these subpaths
The second panel of Figure 1 shows the cut sets for that lattice. The paths between adjacent cut sets are tied at their ends so that they form sublattices, and these are then concatenated to form a *pinched lattice* as shown in the third panel of Figure 1 (and the second panel of Figure 2). Each of these sublattices contains one word from the reference string and the other word sequences which aligned to it. An expanded subsection of the pinched lattice from the cut sets N3 to N6 is given in the fourth panel of Figure 1. The dashed arcs show the likelihood of the word hypotheses. For instance, $-\log P(W_5 = ALL, O) = 2.16 + 0.7 + 1.196$, is the log likelihood of all the paths whose fifth word is ALL.

The pinched lattice is a sequence of sublattices each of which is aligned to a single word in the reference string. These sublattices are called *confusion sets* because they contain likely and errorful hypothesis segments that the ASR system might confuse with the reference words. It is important to stress that all the sentence hypotheses from the original ASR lattice are preserved in creating the pinched lattice and that no paths are removed by pinching. In fact, pinching may actually introduce new paths by piecing together subpaths from the original lattice; however these new paths are insignificant from a modeling point of view, in that they should be of lower probability than any of the original lattice paths.

The evidence space is pruned in two steps. In the first step, the likelihood of each lattice arc is used to discard all paths through every confusion set so that only the most likely alternative to the reference word remains. This is illustrated in the transition from the second to the third panel of Figure 2. When the confusion sets are pruned to contain binary alternatives, we call them *confusion pairs*. In the second pruning step, we simply count all the confusion pairs in the training set lattices, and if any pair has occurred fewer times than a set threshold, that pair is everywhere pruned back to the reference transcription. As an example, the bottom panel of Figure 1 shows that the confusion pair (WELL O, HELLO) is pruned back to HELLO; similarly, the bottom panel of Figure 2 shows that 4 is removed as an alternative to OH.

The result is a greatly reduced evidence space $\tilde{\mathcal{W}}$ derived from the original lattice $\mathcal{W}$. The reduction is controlled by the occurrence threshold, and we usually determine through experimentation what value gives a reasonable sized N-Best expansion of $\tilde{\mathcal{W}}$. For example, the 3 binary confusion pairs appearing in the example of Figure 1 give an N-Best list of depth $2^3$, and the depth can

be varied by the number of hypotheses pruned away.

### *2.2.3 Induced Loss Functions*

Our original motivation to refine the evidence space was to speed up MBR search. However lattice pinching also allows us to redefine the string-to-string loss within $\mathcal{W}$. Suppose the reference string $\bar{W}$ has $K$ words $\bar{W}_0 \dots \bar{W}_{K-1}$. After pinching, a string $W' \in \tilde{\mathcal{W}}$ is not allowed to be aligned completely freely to $\bar{W}$; its alignment must follow the constraints of $\tilde{\mathcal{W}}$. We refer to the corresponding loss as the *induced loss function*:

$$l_I(\bar{W}, W') = \sum_{i=0}^{K-1} l(\bar{W}_i, W'_i) \tag{14}$$

where $W'_i$ is the portion of $W'$ that is aligned to $\bar{W}_i$. If the initial lattice-to-string alignment was good, $l_I(\bar{W}, W')$ will be a good approximation to $l(\bar{W}, W')$.

## 3 Pinched Lattice Minimum Bayes Risk Discriminative Training

The induced loss function and the pinched and pruned evidence space produced by lattice segmentation can be used to reduce the computational cost of minimum Bayes risk discriminative training in large vocabulary speech recognition. By approximating the original lattice $\mathcal{W}$ by the pinched lattice $\tilde{\mathcal{W}}$ and by using the induced loss function Equation 14 in place of the Levenstein distance, the initial training objective of Equation 2 becomes

$$\theta^* \simeq \operatorname*{argmin}_{\theta} \sum_{W' \in \tilde{\mathcal{W}}} l_I(\bar{W}, W') P(W'|O;\theta) \,. \tag{15}$$

Estimation is via Equations (5) and (6) with $\mathcal{W}$ replaced by $\tilde{\mathcal{W}}$, and taking $K(W', \tilde{\mathcal{W}})$ as

$$K(W', \tilde{\mathcal{W}}; \theta) = [\sum_{W'' \in \tilde{\mathcal{W}}} P(W''|O;\theta) l_I(\bar{W}, W'') - l_I(\bar{W}, W')] \, P(W'|O;\theta) \,. \tag{16}$$

Given the induced loss function, this last quantity can be written as

$$K(W', \tilde{\mathcal{W}}; \theta) = [R_I(\bar{W}, \tilde{\mathcal{W}}; \theta) - l_I(\bar{W}, W')] P(W'|O;\theta) \;\; \forall \, W' \in \tilde{\mathcal{W}} \tag{17}$$

where $R_I(\bar{W},\ \tilde{\mathcal{W}};\theta) = \sum_{W''\in\tilde{\mathcal{W}}} P(W''|O;\theta) l_I(\bar{W}, W'')$ is the expected induced loss. This leads to the following training algorithm.

**Pinched Lattice Minimum Bayes Risk Discriminative Training: The PLMBRDT Algorithm**
**Step 1** Generate lattices over the training set
**Step 2** Align the training set lattices to the reference transcriptions
**Step 3** Segment the lattices
**Step 4** Prune the confusion sets to confusion pairs
**Step 5** Discard infrequently occurring confusion pairs
**Step 6** Expand each pinched lattice into an N-Best list, keeping $l_I(\bar{W}, W')$
**Step 7** Compute $R_I(\bar{W};\tilde{\mathcal{W}})$ as defined above
**Step 8** For each $W' \in \tilde{\mathcal{W}}$, compute $K(W', \tilde{\mathcal{W}};\theta)$ by Equation 17
**Step 9** Perform a Forward-Backward pass for each $W' \in \tilde{\mathcal{W}}$
**Step 10** Perform reestimation via Equations 5 and 6, replacing $\mathcal{W}$ by $\tilde{\mathcal{W}}$

This implementation does expand lattices into N-Best lists of sentence hypotheses. However, it is the pinched lattices that are expanded, not the original lattices generated by the large vocabulary ASR decoder. The pinched lattices are much reduced relative to the original lattice and, since we have control over the degree of pinching and pruning, we can control the size of the N-Best lists. In this way we can reduce the evidence space drastically so that the original formulation by Kaiser et al. (2000, 2002) can be applied directly to large vocabulary ASR, albeit under the induced loss function.

We next consider two algorithmic variants that arise from simplifications of $\tilde{\mathcal{W}}$. The first variant is *Pinched Lattice MMIE* which is appropriate for small vocabulary ASR tasks based on whole-word models. The second variant is *One-Worst Pinched Lattice MBRDT* which is a form of corrective training against a competing hypothesis extracted from the pinched lattice.

### *3.1 Pinched Lattice MMIE for Whole Word Acoustic Models*

Lattice cutting segments the original hypothesis space into a concatenation of $K$ sublattices: $\tilde{\mathcal{W}} = \tilde{\mathcal{W}}_0 \cdot \tilde{\mathcal{W}}_2 \cdots \tilde{\mathcal{W}}_{K-1}$. In regions of low confidence, the evidence space contains portions of the MAP hypothesis along with confusable alternatives. In regions of high confidence, the search space is restricted to follow the reference itself. We can express the empirical risk under the induced loss (Equation 15) as:

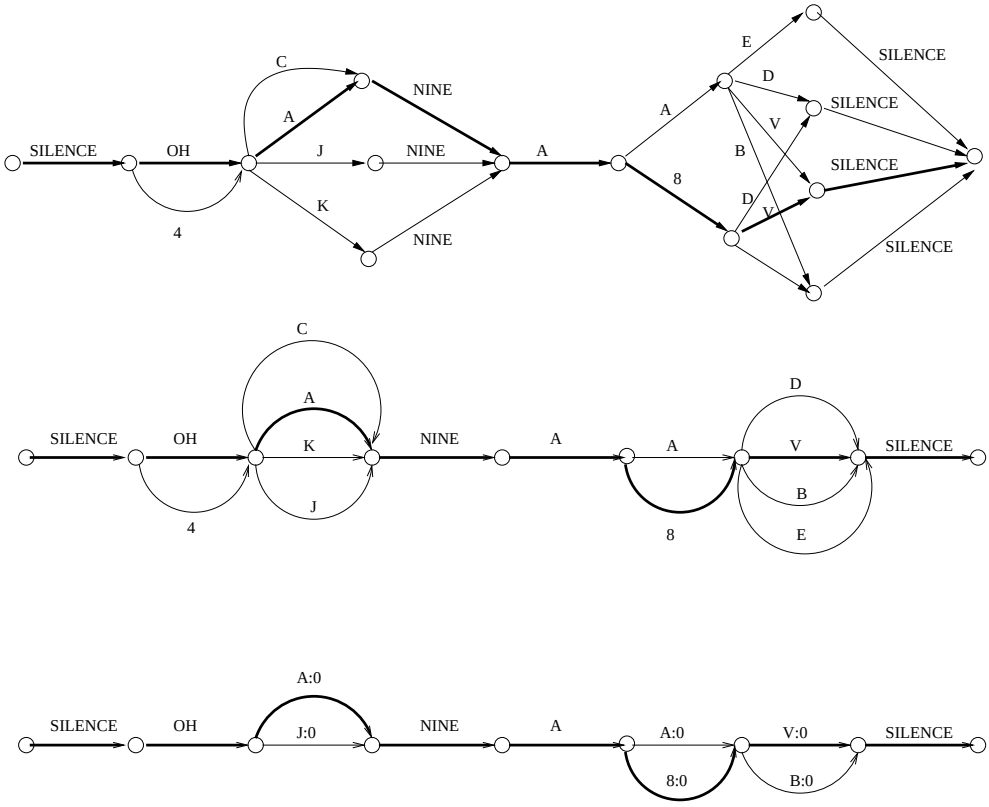


Fig. 2. Derivation of a pinched lattice for the Alphadigits task showing tagged word labels in the confusion pairs for Pinched Lattice MMI. The pinching procedure follows that of Figure 1. The top lattice is produced by a small-vocabulary ASR system; the lattice path corresponding to the ML hypothesis is marked in bold. The middle lattice shows a lattice aligned to the reference (ML) hypothesis; note that all paths are preserved from the original (top) lattice. The bottom lattice shows a pinched and pruned lattice containing confusion pairs with single word alternatives to the reference hypothesis; the word labels in these sets are tagged simply by appending the label ':0'.

$$\theta^* = \operatorname*{argmin}_{\theta} \sum_{W' \in \tilde{\mathcal{W}}} l_I(\bar{W}, W') P(W'|O;\theta) \tag{18}$$

$$= \operatorname*{argmin}_{\theta} \sum_{W'_0 \in \tilde{\mathcal{W}}_0} \cdots \sum_{W'_{K-1} \in \tilde{\mathcal{W}}_{K-1}} \sum_{i=0}^{K-1} l(\bar{W}_i, W'_i) P(W'|O;\theta) \tag{19}$$

$$= \operatorname*{argmin}_{\theta} \sum_{i=0}^{K-1} \sum_{W' \in \tilde{\mathcal{W}}_i} l(\bar{W}_i, W') P_i(W'|O;\theta) \tag{20}$$

where $P_i(W'|O;\theta)$ is the posterior probability that $W'$ is found in the $i^{th}$ lattice segment.

Next we introduce the *global confusion class* $C \subset \{0, \ldots, K-1\}$ to indicate the sublattices that permit alternatives to the truth, i.e. $i \in C$ implies that $\mathcal{W}_i$ contains at least one segment not in the reference hypothesis; for example,

in Figure 2, $C$ is $\{3, 6, 7\}$. We can then write the objective as

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \sum_{i \in C} \sum_{W' \in \tilde{\mathcal{W}}_i} l(\bar{W}_i, W') P_i(W'|O, \tilde{\mathcal{W}}; \theta) \tag{21}$$

since confusion sets that have no alternatives to the truth do not contribute to the overall risk. Finally, we assume that we have a 0/1 loss function over the confusion pairs and arrive at the "pinched lattice" MMI objective function

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \sum_{i \in C} P_i(\bar{W}_i|O, \tilde{\mathcal{W}}; \theta) \, . \tag{22}$$

Therefore, under all these assumptions, the empirical risk is minimized by maximizing the likelihood of the correct hypothesis in the confusable segments. This can be done by a simple modification to the 'normal' MMI procedure that forces it to focus on the low confidence regions identified by lattice pinching.

**Pinched Lattice MMI for Whole Word HMMs:**
**The PLMMI Algorithm**
**Step 1** Generate lattices over the training set (Figure 2, top)
**Step 2** Align the training set lattices to the reference transcriptions
**Step 3** Segment the lattices (Figure 2, middle)
**Step 4** Prune the confusion sets to confusion pairs
**Step 5** Discard infrequently occurring confusion pairs
**Step 6** Tag word hypotheses in confusion pairs (Figure 2, bottom). The labels A:0 and J:0 make it possible to distinguish an A confused with a J from a 'high confidence' A.
**Step 7** Regenerate lattices over the training set using the tagged and pinched lattices to constrain recognition (in contrast to Step 1). If the task requires a grammar, compose the tagged and pinched lattices with the task grammar before lattice regeneration/rescoring. The grammar should be (trivially) extended to cover the tagged words.
**Step 8** Perform lattice-based MMI (Woodland and Povey (2000)) using the word boundary times obtained from the lattice. The procedure differs from regular MMI in that statistics needed in Equations 3 and 4 are gathered only over tagged word hypotheses. Statistics from the un-tagged word hypotheses, which correspond to the high-confidence regions in the pinched lattice, are discarded.

In PLMMI the Levenshtein distance is not used explicitly in the reestimation procedure. It is used to create a pruned search space that contains only the confusable pairs identified by lattice segmentation. Statistics are compiled over these lattices as usual for lattice-based MMI, with the exception that statistics are gathered only for those word instances that appear in confusion sets. This is how we enforce the requirement that statistics be gathered only over segments in the global confusion class $C$.

The PLMMI technique is most appropriate for whole-word models, since it requires reconciling word start and end times with the HMMs that model the word. This is easily done with whole word models, but is much more difficult with subword models in which the multiple HMMs come together to form a single word model during decoding.

### *3.2 'One Worst' Pinched Lattice MBRDT*

The final approximation leads to a poor-man's version of corrective training. After lattice pinching and pruning, we select the *worst* sentence hypothesis from $\tilde{\mathcal{W}}$

$$W^* = \underset{W' \in \tilde{\mathcal{W}}}{\operatorname{argmax}}\ l_I(\bar{W}, W') \,. \tag{23}$$

and we redefine the evidence space to be $\tilde{\mathcal{W}} \simeq \{\bar{W}, W^*\}$. This forces the training procedure to reduce the likelihood of $W^*$ relative to the truth $\bar{W}$. Since only two hypotheses are considered, this approximation greatly simplifies the estimation procedure. The training objective function (15) then becomes

$$\underset{\theta}{\operatorname{argmin}}\, l_I(\bar{W}, W^*) P(W^*|O;\theta) \tag{24}$$

By simple arithmetic in (17) it follows that

$$\begin{aligned} K(\bar{W}, \tilde{\mathcal{W}};\theta) &= [R_I(\bar{W}, \tilde{\mathcal{W}};\theta) - l_I(\bar{W}, \bar{W})] P(\bar{W}|O) \\ &= l_I(\bar{W}, W^*)\ P(\bar{W}|O) P(W^*|O) \\ K(W^*, \tilde{\mathcal{W}};\theta) &= [R_I(\bar{W}, \tilde{\mathcal{W}};\theta) - l_I(\bar{W}, W^*)] P(W^*|O) \\ &= -l_I(\bar{W}, W^*)\ P(\bar{W}|O) P(W^*|O) \end{aligned}$$

leading to $K(\bar{W}, \tilde{\mathcal{W}};\theta) = -K(W^*, \tilde{\mathcal{W}};\theta)$. Note that in the above we restrict the acoustic likelihood $P(O)$ to the two word sequences $\bar{W}$ and $W^*$, so that $P(O) = P(O|\bar{W})P(\bar{W}) + P(O|W^*)P(W^*)$. A further approximation is to discard the terms $l_I(\bar{W}, W^*)\ P(\bar{W}|O)P(W^*|O)$ so that the update equations become

$$\bar{\mu}_s = \frac{\sum_\tau \gamma_s(\tau;\ \bar{W}) o(\tau) - \sum_\tau \gamma_s(\tau;\ W^*) o_\tau + D_s \mu_s}{\sum_\tau \gamma_s(\tau;\ \bar{W}) - \sum_\tau \gamma_s(\tau;\ W^*) + D_s} \tag{25}$$

$$\bar{\Sigma}_s = \frac{\sum_\tau \gamma_s(\tau;\ \bar{W}) o(\tau)^2 - \sum_\tau \gamma_s(\tau;\ W^*) {o_\tau}^2 + D_s\,(\Sigma_s + {\mu_s}^2)}{\sum_\tau \gamma_s(\tau;\ \bar{W}) - \sum_\tau \gamma_s(\tau;\ W^*) + D_s} - \bar{\mu}_s^2. \tag{26}$$

**The 'One Worst' Pinched Lattice MBRDT Algorithm**
**Step 1** Generate lattices over the training set
**Step 2** Align the training set lattices to the reference transcriptions
**Step 3** Segment the lattices

**Step 4** Prune the confusion sets to confusion pairs
**Step 5** Discard infrequently occurring confusion pairs
**Step 6** Extract the most errorful hypothesis $W^*$
**Step 7** Perform Forward-Backward passes with respect to $\bar{W}$ and $W^*$
**Step 8** Perform MMI as in Equations 25 and 26

This approach is very similar to Minimum Classification Error training (Juang and Katagiri (1992)) in that we attempt to improve the likelihood of the $\bar{W}$ relative to the 'worst' hypothesis $W^*$. What distinguishes this approach from other forms of corrective training is not the update procedure itself, but rather the way in which the competing hypothesis is obtained as the most errorful sequence found by lattice pinching and pruning rather than as the best (most likely) incorrect hypothesis.

### *3.3 Summary of MBRDT Algorithms*

We have used the induced loss functions and pinched and pruned lattices that can be derived from lattice segmentation to simplify the implementation of Minimum Bayes Risk Discriminative Training for large vocabulary ASR systems. The first algorithm, PLMBRDT, is a direct application of the minimum risk estimation procedure of Kaiser et al. (2000, 2002) under the induced loss function. The second procedure, PLMMI, is a modified version of MMI for whole word acoustic models that is performed over pinched lattices with binary confusion pairs. The third procedure, 'One Worst' PLMBRDT is a simple form of corrective training in which the MMI-variant improves the likelihood of the reference hypothesis relative to the worst competing candidate found in the pinched lattice.

Both the PLMBRDT and the 'One Worst' Pinched Lattice MBRDT are carefully constructed so that they can be applied to large vocabulary ASR tasks with sub-word acoustic models. Once the pinched and pruned evidence space $\tilde{W}$ is expanded into an N-Best list of sentence hypotheses $W'$, the Forward-Backward algorithm is performed with respect to each hypothesis to generate the statistics needed for minimum risk reestimation. In this way we do not need to keep track of the word or subword model boundary times found in the initial lattice generation. Applying the estimation procedure to a large vocabulary task is as straightforward as performing Forward-Backward passes with respect to the transcriptions in the N-Best list extracted from the pruned evidence space and weighting the resulting statistics by the $K(W', \tilde{\mathcal{W}}, ; \theta)$ factor.

## 4 Small Vocabulary ASR Performance and Analysis

Our basic estimation procedures were developed on the OGI Alphadigits (Noel (1997)) small-vocabulary speech recognition task. The lattice cutting techniques we employ attempt to identify regions of confusion and likely recognition errors. By studying a small vocabulary problem we restrict the diversity of recognition errors so that we can analyze MBRDT performance in detail.

### *4.1 Baseline MMIE System Description*

The baseline is a whole word HMM system built using the HTK Toolkit (Young et al. (2000)). The Alphadigits training set consists of 46730 utterances parameterized as 13 element MFCC vectors with first and second order differences. The baseline maximum likelihood models contain 12 mixtures per state, estimated according to the usual HTK training procedure.

The Alphadigits test set consists of 3112 utterances. Because the Alphadigits task does not have a specific language model, recognition both for MMI lattice generation and test set decoding is performed using an unweighted word loop over the vocabulary. The AT&T Large Vocabulary Decoder (Mohri et al. (2001)) was used to generate lattices for the training set which are then transformed into word posteriors based on the lattice total acoustic score. Using the lattices obtained by the AT&T decoder, word level posteriors were then estimated based on the lattice total acoustic score. MMIE was then performed at the word level using the word time boundaries taken from the lattices. The Gaussian model means and variances are updated by equations (3) and (4). An effective lower bound on $D_s$ is the value which ensures that all variances remain positive; a Gaussian specific value was used in these experiments as suggested by Woodland and Povey (2000).

Figure 3 shows that significant improvement over the baseline can be obtained by MMI: the initial ML performance of 10.42% WER is reduced to 8.41% before overtraining is observed in the test set WER.

### *4.2 Patterns of Binary Word Errors and Confusions*

As described in Sections (3, 3.1), our training procedure attempts to create models that can resolve the recognition errors represented by the confusion sets that result from lattice pinching. The effectiveness of our overall modeling approach depends on the reliability with which these confusion pairs can be associated with ASR errors. If this can be done, there is the possibility

Table 1
Dominant Alphadigit test set error pairs in unconstrained recognition after five MMI iterations

| Rank | Error Pair ($A$+$B$) | $A$ Wrongly Hypothesized | $B$ Wrongly Hypothesized | Occurrences of Each Pair |
|---|---|---|---|---|
| 1. | F+S | 35 | 89 | 124 |
| 2. | V+Z | 51 | 42 | 93 |
| 3. | M+N | 24 | 56 | 80 |
| 4. | P+T | 28 | 39 | 67 |
| 5. | B+V | 30 | 37 | 67 |
| 6. | 8+H | 15 | 32 | 47 |
| 7. | L+OH | 10 | 30 | 40 |
| 8. | A+8 | 20 | 18 | 38 |
| 9. | C+V | 15 | 16 | 31 |
| 10. | B+D | 11 | 17 | 28 |

There are a total of 646 errors identified as belonging to one of these pairs out of a total of 1571 errors.

of training discriminative models on the segmented training set lattices and applying these models to the test data to attempt to resolve errors made by the baseline system. We now investigate the degree to which the confusion pairs identified by lattice pinching agree with actual word errors.

Table 1 lists the most frequent word errors (*error pairs*) observed after five iterations of MMI estimation. The models are chosen after the fifth iteration because performance is nearly optimal at that point. The analysis indicates, for example, that there are 35 instances in which 'S' is hypothesized by the system, when the true word was actually 'F'. Similarly, there are 89 instances in which the system produced an 'F' rather than an 'S'. This simple error analysis is found through unconstrained decoding over the test set. The error pairs are extracted from the hypothesis-to-reference alignments under the Levenshtein distance used to compute the recognition WER. It provides a reference against which we can assess the viability of lattice cutting as a strategy to identify potential errors.

We analyze the distribution of confusion pairs over the Alphadigits test set by extracting them using an unsupervised version of the lattice cutting procedure described in Sections (2.2.1, 3.1). The process starts by identifying the MAP path in a first-pass ASR lattice (e.g. the bold path in the top panel of Figure 2). We obtain confusion sets by aligning lattice paths to this hy-

Table 2
The ten most frequently observed confusion pairs found by lattice cutting in the Alphadigits test and training sets

| Test Set | | | Training Set | | |
|---|---|---|---|---|---|
| Error Pair Rank | Confusion Pair | Occurrences | Error Pair Rank | Confusion Pair | Occurrences |
| 1 | F+S | 699 | 1 | F+S | 15 165 |
| 4 | P+T | 660 | 4 | P+T | 10 728 |
| 6 | 8+H | 650 | 6 | 8+H | 10 290 |
| 3 | M+N | 584 | 3 | M+N | 10 146 |
| 2 | V+Z | 493 | 2 | V+Z | 8 038 |
| 10 | B+D | 344 | 10 | B+D | 5 961 |
| 7 | L+OH | 300 | 7 | L+OH | 5 077 |
| 5 | B+V | 319 | 5 | B+V | 4 939 |
| - | A+K | 238 | - | 5+I | 4 327 |
| - | 5+I | 236 | - | J+K | 3 618 |

pothesis, and prune them to binary confusion pairs. We performed the sanity check of rescoring the pinched test set lattices with the MMI models used to generate the baseline lattices. Since the pinched test set lattices contain the MAP hypothesis we found that after rescoring performance was identical to unconstrained decoding. This confirms that the search space reduction introduces no new errors, and that the selection of the MAP hypothesis is not influenced by pinching. However, pruning does reduce the lattice search space substantially. As a result the Lattice Word Error Rate (LWER) (the minimal number of insertions deletions and substitutions with respect to the reference transcription) of the original lattices which is 1.27%, increases to 3.11% after pinching and pruning. While this significant increase in LWER may seem at odds with the claim that no new errors are introduced, pruning is carefully performed so that this is the case: the baseline hypothesis is retained and only alternative hypotheses are discarded. The increase in LWER does indeed limit the scope for improvement relative to the baseline hypothesis, but when rescoring the pruned lattices with the baseline acoustic models, the baseline hypothesis should always result.

Confusion pairs are extracted from the training data in the same way, except that the lattices are aligned to the reference transcriptions. The confusion sets are pruned to binary confusion pairs, and the 50 most frequently occurring pairs are kept; all other confusion pairs are pruned back to the reference word. These are the lattices used for PLMMI.

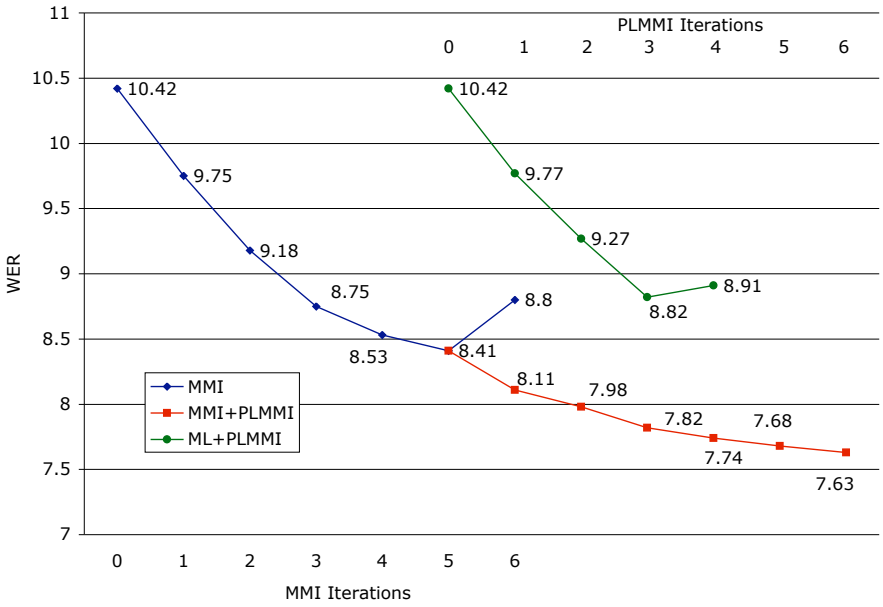


Fig. 3. Performance of MMI and PLMMI models in Alphadigit decoding. The MMI system is initialized from an ML baseline of 10.42% WER. Two different initialization points are chosen for the PLMMI procedure. ML+PLMMI training begins from the ML baseline HMM (10.42% WER), and MMI+PLMMI training follows 5 MMI iterations (at 8.41% WER).

The ten most frequent training and test set confusion pairs are given in Table 2. There is a strong agreement between the confusion pairs found in the training and test sets. The six most frequent pairs are in agreement across both sets, and eight of the pairs are in the top ten of both sets. Interestingly, there appears to be a systematic bias between the ordering of confusion pairs and the ordering of error pairs, in that the frequency of confusion pairs does not strictly follow the error pair frequency. However, apart from the difference in ordering, the frequent confusion pairs are also frequent error pairs, as desired. Confusion pairs reflect the confidence the ASR system has in its hypotheses, while error pairs reflect the accuracy of the one-best hypothesis. While these measures are not in exact agreement, we find that confusion pairs are in fact good indicators of the frequently occurring word errors.

### *4.3 Pinched Lattice MMI Performance and Within-Class Error Analysis*

Models trained after five MMI iterations were used to initialize the Pinched Lattice MMI (MMI+PL) estimation procedure. We observe in Figure 3 that

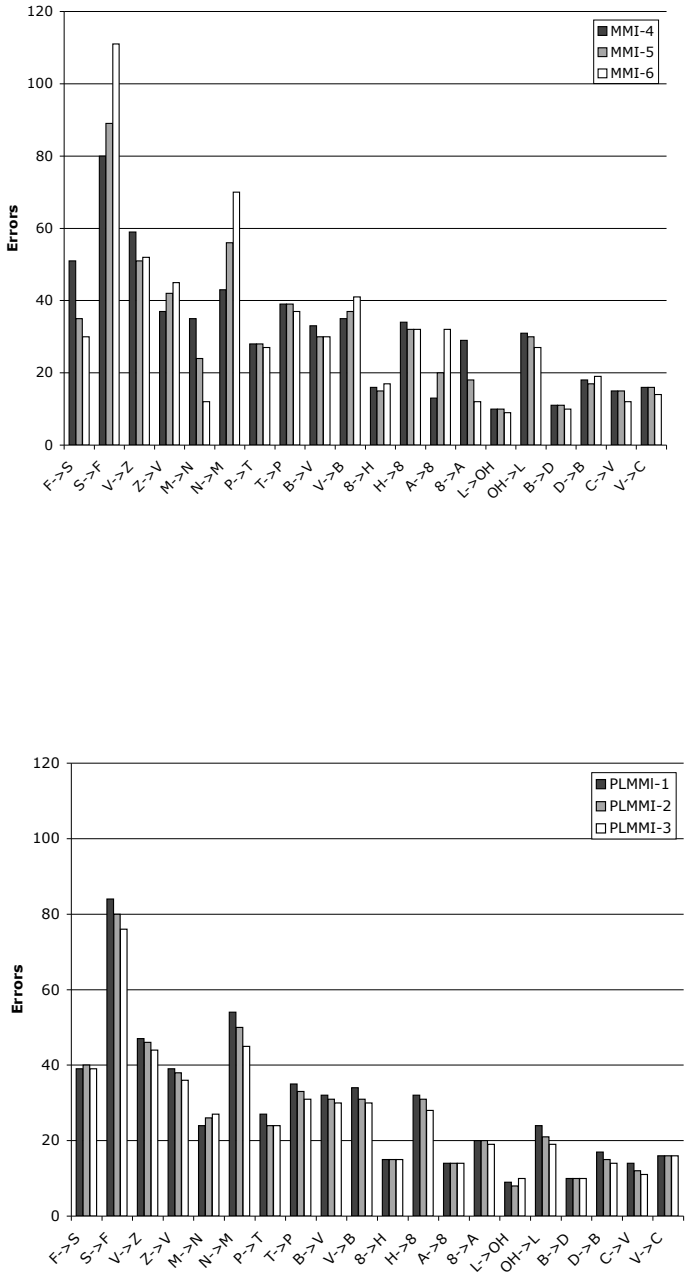


Fig. 4. Error counts within the most frequent error pairs for three iterations of MMI (top) and Pinched Lattice MMI (bottom). For example, the first three bars give the number of times 'F' is wrongly hypothesized as 'S' when decoding with models at each of the three training iterations.

the iterations of pinched lattice MMI estimation yield continued improvement in WER from the best MMI performance to 8.41% to 7.63%. This is contrast to "regular lattice" MMI, which shows evidence of overtraining beyond the fifth iteration. This is done as a fair comparison between pinched lattice and regular MMI, in that the systems being compared are of equal complexity and have the same number of parameters. The improved performance can be attributed to the use of lattice pinching in MMI estimation to refine the evidence space of competing word hypotheses. By contrast, PLMMI initialized from the ML baseline (ML+PLMMI) does not perform as well, and in fact performs worse than MMI itself. One explanation for this is that PLMMI only considers the 50 most frequent confusion pairs whereas MMI updates the models for all words. More important, however, is the correct initialization of these training procedures. Just as MMI works best when initialized from the best available ML models, our experience suggests that PLMMI (and other MBRDT training procedures) work best when initialized from well-trained MMI models.

We can analyze the behavior of the substitution errors made in rescoring with models trained with the MMI and pinched lattice MMI procedures. Each error pair has two types of errors: for example, within the error pair 'F+S', 'F' can be misrecognized as 'S', or 'S' can be misrecognized as 'F'. Ideally, both types of errors should decrease over each of the training iterations shown. However, as can be seen in Figure 4(top), despite the overall reduction in WER achieved by MMI training, error types are not reduced uniformly as training proceeds. For example, the decrease in $F \rightarrow S$ indicates that the number of times $F$ is incorrectly recognized as $S$ decreases sharply over the three MMI iterations. While this is good in itself, the complementary value of $S \rightarrow F$ indicates that it is gained at the cost of introducing errors in which $S$ is wrongly recognized as $F$. We find that this undesirable behavior less evident with the Pinched Lattice MMI models (Figure 4 bottom) in which the types of errors over each class are more balanced.

## 5 MBRDT for Large Vocabulary Automatic Speech Recognition

We will describe comparisons of the proposed risk-based discriminative training procedures 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 (Byrne et al. (2004)). Both systems are speaker independent continuous mixture density, tied state, cross-word, gender-independent, triphone HMM systems trained with HTK.

The Switchboard training set speech was parameterized into 39-dimensional

PLP cepstral coefficients with delta and acceleration coefficients (Hermansky (1990)). Cepstral mean and variance normalization was performed over each conversation side. There were 4000 unique triphone states with 6 Gaussian components per state. Lattice rescoring experiments were performed using the AT&T Large Vocabulary Decoder (Mohri and Riley (1999)), with a 33K-word trigram language model provided by SRI (Stolcke et al. (2000)). The baseline acoustic models used as seed models for our experiments, were built using HTK from 16.4 hours of Switchboard-1 and 0.5 hour of Callhome English data. This collection defined the minitrain development training set for the 2001 JHU LVCSR system (Byrne (2001)). The training set consists of 22580 utterances. The recognition tests were carried out on a subset of the 2000 Hub-5 Switchboard-1 evaluation set (SWBD1) (Martin et al. (2000)) and the 1998 Hub-5 Switchboard-2 evaluation set (SWBD2) (Martin et al. (1998)). The total SWITCHBOARD test set was 2 hours of speech.

The MALACH Czech baseline acoustic models were built from 62 hours of data with 24065 utterances, and the MALACH-CZ language model was a back-off bigram with a 83K word vocabulary (Byrne et al. (2004)). 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 and has approximately 2 hours of speech.

Lattice-based MMI was performed in each domain. MMI estimation was performed at the triphone model level, with triphone time boundaries extracted from ASR lattices generated over the training sets. The SWITCHBOARD lattices were generated once and the lattice link posteriors were fixed for all MMI iterations. In MALACH-CZ, the link posteriors were reestimated after each MMI iteration.

### *5.1 MMIE Performance on SWITCHBOARD and MALACH-CZ*

We first describe performance under MMIE training in the SWITCHBOARD domain. MMI training is seeded from a well-trained maximum likelihood (ML) system which achieves WERs of 41.1% on the SWBD1 test set and 51.1% on the SWBD2 test set. We update the Gaussian model parameters using Equations 3 and 4. As in the Alphadigits experiments, a Gaussian specific value of $D_s$ was used for our experiments. To validate our approach we calculated the WER and the value of the MMI objective function ($\log P(\bar{W}|O)$) over the training set at each iteration.

From Table 3 we see that MMIE performs as expected. As measured over the SWITCHBOARD training set, the overall MMI objective function increases as training proceeds and the WER over the training set decreases, as does

Table 3
MMIE Performance on the SWITCHBOARD and MALACH-CZ Tasks.

| | SWBD Training Set | | | | |
|---|---|---|---|---|---|
| MMIE | | MMI Objective | Test Set WER(%) | | |
| Iteration | WER (%) | Function | SWBD1 | SWBD2 | MALACH-CZ |
| ML | 29.42 | -2.37E05 | 41.1 | 51.1 | 44.3 |
| 1 | 27.55 | -2.05E05 | 40.6 | 50.5 | 43.4 |
| 2 | 26.24 | -1.81E05 | 40.5 | 50.0 | 42.4 |
| 3 | 25.62 | -1.647E05 | 39.9 | 49.7 | 42.1 |
| 4 | 25.66 | -1.53E05 | 40.2 | 50.5 | 41.9 |
| 5 | | | | | 41.6 |
| 6 | | | | | 41.5 |

The analysis of MMI performance over the SWITCHBOARD training set is given to verify that the MMI implementation performs as expected.

the WER over the SWBD1 and SWBD2 test sets, until there is evidence of overtraining at the third MMIE iteration. We see similar performance in the MALACH Czech ASR task, also reported in Table 3.

### *5.2 Risk-Based Pruning of the Evidence Sets*

The lattice segmentation procedure described in general in Section 2.2.2 and applied to the Alphadigits task in Section 3.1, can also be applied to the MALACH-CZ and SWITCHBOARD training sets. Following an initial lattice generation decoding pass over the training set, we use lattice cutting with respect to the reference transcription to produce pinched and pruned lattices with binary confusion pairs.

We performed two sets of experiments with the SWITCHBOARD system and one set with the MALACH-CZ system. In MALACH, a confusion pair occurrence threshold of 100 was used to create a single evidence set. In SWITCHBOARD, two threshold values, 5 and 75, were used to create two separate evidence sets. Following the pruning procedure of Section 2.2.2, all confusion pairs observed fewer times than the occurrence pruning threshold were pruned back to the reference transcription.

As the occurrence threshold increases more confusion pairs are pruned away. For example, Table 4 shows that the number of distinct confusion pairs in the SWITCHBOARD training data drops from 2139 to 159 when the threshold

Table 4
SWITCHBOARD and MALACH-CZ training set reduction by lattice pinching and pruning.

| | SWITCHBOARD | | MALACH-CZ |
|---|---|---|---|
| Acoustic training data (hours / utterances) | 16.9 / 22580 | | 62.4 / 24065 |
| Initial confusion pairs (types / tokens) | 25948 / 99199 | | 31467 / 120695 |
| Occurrence threshold used to select confusion pairs | 5 | 75 | 100 |
| Confusion pairs after filtering (types / tokens) | 2139 / 66349 | 159 / 33821 | 117 / 48302 |
| Avg. confusion pairs (per word / per utterance) | 0.35 / 3.37 | 0.2 / 2.14 | 0.13 / 3.12 |
| Reduced acoustic training data (hours / utterances) | 15.0 / 19687 | 13.0 / 15741 | 52.4 /15436 |
| Avg. depth of N-Best lists from pinched lattices | 48.8 | 13.1 | 36.5 |

increases from 5 to 75. The number of confusion pairs per utterance drops from 3.37 to 2.14. Due to this pruning of confusion pairs, many training set lattices are reduced to a single word sequence, i.e. if no confusion pairs remain, the pinched lattices will contain only the reference transcription. Since the loss over these lattices is zero, these utterances do not contribute to the overall training criterion and they are removed from the training data. As a result, by retaining only confusion pairs that occur 100 times or more, the MALACH-CZ training data is reduced from 62.4 hours to 52.4 hours, and similar reductions are found in SWITCHBOARD.

### *5.3 PLMBRDT Performance on SWITCHBOARD and MALACH-CZ*

The performance of the MBRDT training schemes is given in Table 5 for the SWITCHBOARD systems and in Table 6 for the MALACH-CZ systems. Both MBRDT and the 'One-Worst' approximation give improvements over the well-trained MMI system. For all systems we report the p-values in parentheses under the significance test between each system and the MMIE baseline system; the values in parentheses give the probability that there is no differ-

Table 5
Minimum Bayes Risk Training performance on SWITCHBOARD in WER(%)

| Occurrence Threshold | 5 | | 75 | |
|---|---|---|---|---|
| | PLMBRDT | One Worst | PLMBRDT | One Worst |
| Iteration | SWITCHBOARD1 - MMIE baseline 39.9 | | | |
| 1 | 39.6 (0.082) | 39.3 (0.01) | 39.6 (0.050) | 39.6 (0.080) |
| 2 | 39.3 (0.018) | 39.4 (0.11) | 39.5 (0.103) | 39.2 (0.011) |
| 3 | 39.5 (0.230) | – | 39.4 (0.112) | 39.8 (0.667) |
| Iteration | SWITCHBOARD2 - MMIE baseline 49.7 | | | |
| 1 | 49.7 (0.826) | 49.5 (0.230) | 49.7 (0.826) | 49.7 (0.726) |
| 2 | 49.5 (0.360) | 49.6 (0.520) | 49.4 (0.160) | 49.4 (0.184) |
| 3 | 49.4 (0.230) | – | 49.7 (0.834) | 49.8 (0.928) |

P-values relative to the MMI baselines are given in parentheses.

ence between that experiment and the baseline MMI system (Pallett et al. (1990)).

On the SWITCHBOARD experiments we see that the PLMBRDT algorithm and its One Worst variant perform comparably. There are not consistent differences in performance over the two evidence spaces, suggesting that the procedure is somewhat insensitive to the confusion pair occurrence pruning threshold, at least in these experiments.

### *5.3.1 Contribution of the Loss Function to Estimation*

A variant of PLMBRDT is applied to the MALACH-CZ system with the specific goal of investigating how the incorporation of the Levenshtein distance in the estimation criterion influences WER reduction. Table 6 shows PLMBRDT performance under the induced loss function $l_I(\bar{W}, W')$, which approximates the Levenshtein distance as described in Equation 15. This algorithm attempts to achieve

$$\underset{\theta}{\operatorname{argmin}} \sum_{W' \in \tilde{\mathcal{W}}} l_I(\bar{W}, W') P(W'|O;\theta) \tag{27}$$

which incorporates both the induced loss function $l_I$ as well as the restriction on the evidence space to the pinched lattice $\tilde{\mathcal{W}}$. We can replace the induced

Table 6
Minimum Bayes Risk Training performance on MALACH-CZ in WER(%)

| | PLMBRDT | | One-Worst |
|---|---|---|---|
| | Induced Loss | 0/1 Loss | |
| Iteration | MALACH-CZ - MMIE baseline 41.5 | | |
| 1 | 41.4 (0.114) | 41.4 (0.134) | 41.3 (0.107) |
| 2 | 41.3 (0.038) | 41.3 (0.129) | 41.2 (0.042) |
| 3 | 41.3 (0.112) | 41.3 (0.080) | 41.0 (0.003) |
| 4 | 41.3 (0.001) | 41.3 (0.197) | 41.1 (0.052) |
| 5 | 41.1 (0.031) | 41.4 (0.522) | — |
| 6 | 41.0 (0.013) | 41.5 (0.478) | — |

P-values relative to the MMI baseline are given in parentheses.

loss function by the sentence level loss function

$$l_{0/1}(\bar{W}, W') = \begin{cases} 0 \ \bar{W} = W' \\ 1 \ \bar{W} \neq W' \end{cases} \tag{28}$$

which transforms the training objective into the MMI variant

$$\underset{\theta}{\text{argmax}} \ \frac{P(\bar{W}, O; \theta)}{\sum_{W' \in \tilde{\mathcal{W}}} P(W', O; \theta)}.$$

This simply reduces to performing MMI over the N-Best lists extracted from the pinched lattice. Just to establish the relationship between the various procedures, this variant is not PLMMI, since the contribution of the loss is at the sentence level rather than the word level.

We perform this comparison to determine the relative contribution of the pinched and pruned evidence space $\tilde{\mathcal{W}}$ and the loss function $l_I$ used in estimation. Starting from the MMIE baseline of 41.5% WER, the complete PLMBRDT algorithm based on the induced loss function reduces WER to 41.0%, whereas the '0/1 Loss' variant reduces WER only to 41.3%. Loosely speaking, we conclude that the loss function contributes as much to the PLMBRDT gains as does the refinement of the evidence space. This is also consistent with the performance of the One Worst approach, which is constructed to pick the most errorful hypothesis from the refined search space. We conclude that it is beneficial to incorporate both the refined search space and the relative costs of the competing hypotheses in PLMBRDT.

## 6 Conclusion

We have demonstrated how techniques developed for Minimum Bayes Risk Decoding make it possible to apply risk-based parameter estimation algorithms to large vocabulary speech recognition tasks. Our approach starts with the original derivations of Kaiser et al. (2000, 2002) which show how the Extended Baum Welch algorithm can be used to derive a parameter estimation procedure to reduce expected loss over training data. That work focuses on incorporating the Levenstein distance into parameter estimation. However their formulation is very general and also supports other types of string-to-string loss functions. The link to Minimum Bayes Risk decoding is made through the induced loss functions that arise from the lattice segmentation algorithms developed for MBR search over large lattices. We use the formulation of Kaiser et al. (2000, 2002), but replace the Levenstein distance with the induced loss functions. Through this approximation we are able to compute the statistics needed to apply the risk-based parameter estimation algorithm to large vocabulary speech recognition tasks.

In these initial experiments we have focused on the most simple lattice pinching and pruning procedures. Each lattice path is aligned word-by-word against the reference transcription, and binary word confusion pairs are identified. These confusion pairs define the errors that the system will be trained to 'fix'. Many types of acoustic errors are excluded from this small number of confusion pairs and as a consequence these errors are not addressed by training. However, the value of this conservative approach is that it allows us to control and study the behavior of the estimation algorithms over a manageable number of word pairs. A PLMBRDT variant, Pinched Lattice MMI, was derived and applied to a whole word recognition task, and analysis of the performance shows that it does indeed reduce the individual types of word errors in a way that MMI does not. These same lattice pinching and pruning procedures can be applied to large vocabulary speech recognition. As in the small vocabulary case, we find that these PLMBRDT algorithms can be used to extend the gains obtained by MMI. These results are given on two large vocabulary recognition tasks, the conversational English SWITCHBOARD corpus, and the spontaneous Czech MALACH corpus. By varying the definition of the estimation algorithms, we find evidence that the improvement beyond MMI comes from both the inclusion of loss into estimation and from reducing the likelihood of the errorful hypotheses that are identified by pinching and pruning.

As mentioned earlier, MMI is a particular instance of risk-based estimation. Under the sentence level loss function, minimum risk estimation becomes

$$\operatorname*{argmin}_{\theta} \sum_{W'} l_{0/1}(\bar{W}, W') P(W'|O;\theta) = \operatorname*{argmax}_{\theta} P(\bar{W}|O;\theta) \ , \qquad (29)$$

which is the MMI objective function. From the view of minimizing risk, MMI is better matched to Sentence Error Rate than to Word Error Rate. This is clearly not a fatal shortcoming, in that MMI can be very effective in reducing Word Error Rate. However we find that MMI can be improved by using discriminative training procedures that are matched to the task metric, and we conclude that matching the estimation criterion to the task performance metric is beneficial for speech recognition performance.

**ACKNOWLEDGEMENTS** We would like to thank AT&T Research for use of the AT&T Large Vocabulary decoder and FSM libraries and S.Kumar of CLSP for helpful discussions and assistance with lattice cutting algorithms.

## References

Byrne, W., 2001. The JHU March 2001 Hub-5 Conversational Speech Transcription System. In: Proceedings of the NIST LVCSR Workshop. NIST.

Byrne, W., Doermann, D., Franz, M., Gustman, S., Hajič, J., Oard, D., Picheny, M., Psutka, J., Ramabhadran, B., Soergel, D., Ward, T., Zhu, W.-J., July 2004. Automatic recognition of spontaneous speech for access to multilingual oral history archives. IEEE Transactions on Speech and Audio Processing, Special Issue on Spontaneous Speech Processing.

Goel, V., Byrne, W., 2000. Minimum Bayes-Risk automatic speech recognition. Computer Speech and Language 14(2), 115–135.

Goel, V., Kumar, S., Byrne, W., 2001. Confidence based lattice segmentation and minimum bayes-risk decoding of lattice segments. In: European Conference on Speech Communication and Technology. Vol. 4. Aalborg, Denmark, pp. 2569–2572.

Goel, V., Kumar, S., Byrne, W., May 2004. Segmental minimum Bayes-risk decoding for automatic speech recognition. IEEE Transactions on Speech and Audio Processing.

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

Hermansky, H., Apr. 1990. Perceptual linear predictive (PLP) analysis of speech. Journal of the Acoustical Society of America 87 (4), 1738–1752.

Juang, B.-H., Katagiri, S., Dec. 1992. Discriminative learning for minimum error classification. IEEE Transactions on Signal Processing 40 (12), 3043–3054.

Kaiser, J., Horvat, B., Kacic, Z., 2000. A novel loss function for the overall risk criterion based discriminative training of HMM models. In: ICSLP. Beijing,China, pp. 888–890.

Kaiser, J., Horvat, B., Kacic, Z., 2002. Overall risk criterion estimation of

hidden Markov model parameters. Speech Communication 38 (3–4), 383–398.

Kumar, S., Byrne, W., 2002. Risk based lattice cutting for segmental minimum Bayes-risk decoding. In: ICSLP 2002. Denver, CO, USA, pp. 373–376.

Martin, A., Fiscus, J., Przybocki, M., Fisher, B., 1998. The evaluation: Word error rates and confidence analysis. In: Hub-5 Workshop. NIST, Linthicum Heights, Maryland, [Online]. Available: http://www.nist.gov/speech/tests/ctr/hub5e_98/hub5e_98.htm.

Martin, A., Przybocki, M., Fiscus, J., Pallett, D., 2000. The 2000 NIST evaluation of conversational speech recognition over the telephone. In: Proceeding of the Speech Transcription Workshop. NIST, [Online]. Available: http://www.nist.gov/speech/publications/tw00.

Mohri, M., Pereira, F., Riley, M., 2001. ATT General-purpose finite-state machine software tools. http://www.research.att.com/sw/tools/fsm/.

Mohri, M., Riley, M., 1999. Integrated context-dependent networks in very large vocabulary speech recognition. In: European Conference on Speech Communication and Technology. pp. 811–814.

Noel, M., 1997. Alphadigits. Tech. rep., Center for Spoken Language Understanding, Oregon Graduate Institue of Science and Technology, Portland, OR, http://www.cse.ogi.edu/CSLU/corpora/alphadigit.

Normandin, Y., 1996. Maximum mutual information estimation of hidden Markov models. In: Lee, C.-H., Soong, F. K., Paliwal, K. K. (Eds.), Automatic Speech and Speaker Recognition: Advanced Topics. Kluwer, Ch. 3, pp. 57–81.

Pallett, D., Fisher, W., Fiscus., J., 1990. Tools for the analysis of benchmark speech recognition tests. In: International Conference on Acoustics, Speech, and Signal Processing. Vol. 1. pp. 97–100.

Sankoff, D., Kruskal, J. (Eds.), 1983. Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison. Addison-Wesley.

Stolcke, A., Bratt, H., Butzberger, J., Franco, H., Rao Gadde, V. R., Plauché, M., Richey, C., Shriberg, E., S onmez, K., Weng, F., Zheng, J., 2000. The SRI march 2000 Hub-5 conversational speech transcription system. In: Proceeding of the Speech Transcription Workshop. NIST, [Online]. Available: http://www.nist.gov/speech/publications/tw00.

Stolcke, A., Konig, Y., Weintraub, M., 1997. Explicit word error minimization in n-best list rescoring. In: Eurospeech. Rhodes, Greece, pp. 163–166.

Woodland, P. C., Povey, D., 2000. Large scale discriminative training for speech recognition. In: Proceedings of the Tutorial and Research Workshop on Automatic Speech Recognition. ISCA, pp. 25–48.

Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P., Jul. 2000. The HTK Book, Version 3.0.