Lattice Segmentation and Minimum Bayes Risk Discriminative Training

Vlasios Doumpiotis, Stavros Tsakalidis, William Byrne

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

Abstract

Modeling approaches are presented that incorporate discriminative training procedures in segmental Minimum Bayes-Risk decoding (SMBR). SMBR is used to segment lattices produced by a general automatic speech recognition (ASR) system into sequences of separate decision problems involving small sets of confusable words. We discuss two approaches to incorporating these segmented lattices in discriminative training. We investigate the use of acoustic models specialized to discriminate between the competing words in these classes which are then applied in subsequent SMBR rescoring passes. Refinement of the search space that allows the use of specialized discriminative models is shown to be an improvement over rescoring with conventionally trained discriminative models.

1. Introduction

Minimum Bayes Risk Decoding (MBR) is an alternative ASR search strategy that produces hypotheses in an attempt to minimize the empirical risk of speech recognition errors [1, 2]. The measurement of risk derives from a loss function that is appropriately chosen for the recognition task; for example, in ASR the Levenshtein distance is most commonly used. MBR decoding has been found to consistently provide improved performance relative to straightforward maximum likelihood (ML) decoding procedures. This is usually credited to the integration of the task performance criterion directly into the decoding procedure. The minimum risk formulation can also be used to explain the effectiveness of ASR hypothesis and system combination procedures such as ROVER [3, 4, 5]. The effectiveness of MBR decoding appears to be fairly independent of how the underlying models are trained. Both ML and discriminatively trained models can benefit from MBR decoding. However there is the possibility of developing estimation procedures for models intended specifically for use in MBR decoding.

We use the Segmental Minimum Bayes Risk (SMBR) decoding framework [6, 7] to develop discriminative estimation and decoding procedures. SMBR is a divide-and-conquer approach to ASR that transforms continuous speech recognition into a connected sequence of separate recognition problems. This approach can be thought of as identifying the recognition problems that remain after the initial recognition pass. Rescoring can then be done with specialized sets of models, each of which is discriminatively trained to "solve" these separate sub-tasks. We call this coupled recognition and estimation strategy Segmental Minimum Bayes Risk Discriminative Training (SMBR-DT) [8]. In this paper we analyze SMBR-DT and its relationship to Maximum Mutual Information estimation

(MMI) [9, 10]. We show that SMBR-DT can yield improvement over MMI both in the overall word error rate and in the distribution of individual word errors.

2. Risk-Based Estimation and Decoding

MBR decoders attempt to find the sentence hypothesis with the least expected error under a given task specific loss function. If $l(W, W')$ is the loss function between word strings W and W', the MBR recognizer seeks the optimal hypothesis given the acoustic data A as

$$\hat{W} = \underset{W \in \mathcal{W}}{\operatorname{argmin}} \sum_{W' \in \mathcal{W}} l(W, W') P(W'|A). \qquad (1)$$

Prior work in MBR decoding has treated it essentially as a large search problem in which $\mathcal{W}$ are N-Best lists or lattices that incorporate $P(W'|A)$ as a posterior distribution on word strings obtained using an HMM acoustic model and an N-gram language model [1, 2].

MAP decoding, which given an utterance A produces a sentence hypothesis according to $\tilde{W} = \operatorname{argmax}_{W \in \mathcal{W}} P(W|A)$, is the optimum decoding criterion when performance is measured under the Sentence Error Rate criterion, i.e. when $l(W, W')$ is the 0/1 valued identity function.

2.1. Segmental Minimum Bayes-Risk Decoders

Segmental Minimum Bayes Risk decoding was initially developed [7] to address the MBR search problem over very large lattices. We assume that each word string $W \in \mathcal{W}$ is segmented into N substrings of zero or more words $W_1, \ldots, W_N$. Since each lattice path is a word string $W \in \mathcal{W}$, this segments the original lattice into N segment sets $\mathcal{W}_i$, $i = 1, 2, ..., N$. Given a specific lattice segmentation, the MBR hypothesis $\hat{W}$ can then be obtained as a sequence of independent decisions [7]

$$\hat{W}_i = \underset{W \in \mathcal{W}_i}{\operatorname{argmin}} \sum_{W' \in \mathcal{W}_i} l(W, W') P_i(W'|A). \qquad (2)$$

$P_i(W'|A)$ is the probability of observing the string W' in the ith segment set: $P_i(W'|A) = \sum_{W \in \mathcal{W}: W_i = W'} P(W|A)$. The complete sentence hypothesis is obtained as $\hat{W} = \hat{W}_1 \cdots \hat{W}_N$.

Once a lattice has been segmented, the distance between two strings in the lattice is constrained by the segmentation, i.e. the distance between W and W' is found as $\sum_{i=1}^{N} l(W_i, W'_i)$. Ideally, we should satisfy the strong requirement that the loss function between any two word sequences $W, W' \in \mathcal{W}$ is not affected by the lattice cutting, i.e. that $l(W, W') = \sum_{i=1}^{N} l(W_i, W'_i)$. This is difficult to achieve and by way of approximation we segment the lattice word strings by aligning each path in the lattice to the MAP sentence hypothesis [7, 11]:
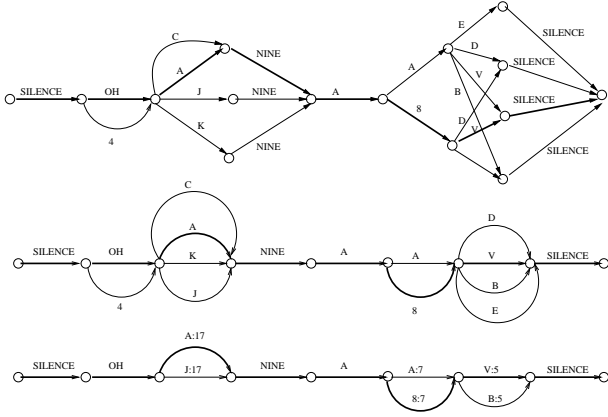
Figure 1: Lattice Segmentation for Estimation and Search. *Top*: First-pass lattice of likely sentence hypotheses with MAP path in bold; *Middle*: Alignment of lattice paths to MAP path; *Bottom*: Refined search space $\tilde{\mathcal{W}}$ consisting of segment sets selected for discriminative training and rescoring.

given the MAP hypothesis $\tilde{W}$, we segment the paths in the lattice to attain $l(\tilde{W}, W') \approx \sum_{i=1}^{N} l(\tilde{W}_i, W'_i)$.

2.1.1. Unsupervised Search Space Refinement and Selection of Confusion Sets

Lattice segmentation can be used both to identify potential errors in the MAP hypothesis and to derive a new search space for the subsequent decoding passes. For each utterance that is to be decoded, we define a new search space, called a "pinched lattice", by concatenating the segment sets found by lattice cutting: $\tilde{\mathcal{W}} = \mathcal{W}_1 \cdots \mathcal{W}_N$. In regions of low confidence, the search space contains portions of the MAP hypothesis along with confusable alternatives; we call these *Confusion Sets*. In regions of high confidence, the search space is restricted to follow the MAP hypothesis itself (Fig. 1). Because the structure of the original lattice is retained whenever we consider alternatives to the MAP hypothesis, we can perform acoustic rescoring over this pinched lattice.

2.2. Risk-Based Training

Suppose we have a database $\{\bar{W}, A\}$ of transcribed speech. One approach to discriminative estimation is to estimate model parameters as follows

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \sum_{W' \in \mathcal{W}} l(\bar{W}, W') P(W'|A; \theta) \qquad (3)$$

As was observed in the MAP decoder, under the 0/1 loss function, Equation 3 becomes the MMI objective function $\operatorname{argmax}_\theta P(\bar{W}|A; \theta)$. The close relationship between MMI and minimum risk estimation is widely known, and MMI-variants for the training criterion of Equation 3 have been developed [12, 13].

2.2.1. Pinched Lattice MMI

An alternative approach to direct risk minimization is to first incorporate lattice pinching, with the goal of focusing both training and decoding procedures on individual recognition errors. Following the approach developed for SMBR decoding, we segment and pinch the lattice paths, so that the original lattice $\mathcal{W}$ is approximated by the pinched lattice $\tilde{\mathcal{W}}$, although in training,

the segmentation is found relative to the correct transcription. When restricted to the pinched lattice, Equation 3 becomes

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{W' \in \tilde{\mathcal{W}}_i} l(\bar{W}_i, W') P_i(W'|A, \tilde{\mathcal{W}}; \theta) \quad (4)$$

where $P_i(W'|A, \tilde{\mathcal{W}}; \theta) = \sum_{W \in \tilde{\mathcal{W}}: W_i = W'} P(W|A; \theta)$. Note that this makes use of the loss function induced by lattice segmentation with respect to the transcription.

Pinching has the further effect that no contribution to the loss function comes from any of the high-confidence segment sets that were constructed so as to agree with the MAP hypothesis. To exclude these, we introduce the global confusion class $C \subseteq \{1, \ldots, N\}$ to indicate the segment sets that permit alternatives to the MAP path, i.e. $i \in C$ implies that $\mathcal{W}_i$ contains at least one segment not in the MAP hypothesis. We can then write the objective as

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \sum_{i \in C} \sum_{W' \in \tilde{\mathcal{W}}_i} l(\bar{W}_i, W') P_i(W'|A, \tilde{\mathcal{W}}; \theta). \quad (5)$$

Finally, we assume that we have a 0/1 loss function and arrive at the "pinched lattice" MMI objective function

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \sum_{i \in C} P_i(\bar{W}_i|A, \tilde{\mathcal{W}}; \theta). \qquad (6)$$

We implement this as lattice-based MMI [10], with two simple modifications. As described above, the posterior statistics are computed over the pinched lattice. The introduction of the global confusion class excludes from the MMI calculation any contribution from the high confidence segment sets; only statistics from those in the global confusion class are accumulated. These two modifications force the MMI procedure to focus on the low confidence regions identified by lattice pinching.

2.2.2. SMBR Discriminative Training

We begin by partitioning the global confusion set C into a collection of error equivalence classes $\{C_j\}_{j=1}^{J}$. We then clone the model parameters J times so that there is a separate, independent parameter set θ_j for each of the J error equivalence classes, i.e. $\theta \rightarrow \{\theta_1, \ldots, \theta_J\}$. Ideally, each class will represent ASR errors that can be "fixed" in the same way. In the simple task we address in this paper, the classes are binary word errors, i.e. the segment sets are generated so that they contain only pairs of words. We note that individual words will appear in multiple equivalence classes, as the equivalence is over errors, not words. For instance, there may be two equivalence classes $C_{17} = \{A, J\}$ and $C_7 = \{A, 8\}$. The models θ_{17} are trained to discriminate between 'A' and 'J', whereas the models θ_7 should discriminate between 'A' and '8'.

Assuming that the parameter sets θ_j can be optimized independently, the pinched lattice MMI criterion of Equation 6 becomes the objective for SMBR Discriminative Training

$$\theta_j^* = \underset{\theta_j}{\operatorname{argmax}} \sum_{i \in C_j} P_i(\bar{W}_i|A, \tilde{\mathcal{W}}; \theta_j) \quad \text{for } j = 1, \ldots, J. \quad (7)$$

We note that pinched lattice MMI is a special case of this form of SMBR-DT. We also emphasize that the HMM model architecture is retained in pinched lattice MMI and SMBR-DT, and that only the training criterion changes. The goal is to estimate HMMs for $P(A|W; \theta_j)$ to optimize Equations 7 and 4.

2.3. SMBR Discriminative Training and Decoding

We have developed the following decoding strategy [8] to integrate the estimation and decoding procedures described in the previous section.

1). Following an initial lattice generation MAP decoding pass, we use lattice cutting with respect to the MAP hypothesis to produce pinched lattices that identify low-confidence segment sets $\mathcal{W}_i$ that are likely to contain recognition errors. These sets are then grouped into equivalence classes $\{C_j\}_{j=1}^J$.

2). We then turn to the training set to find all relevant data that can be used to train models $P(A|W; \theta_j)$ that can be applied for all $W \in \mathcal{W}_i$ and $i \in C_j$.

3). We finally apply these models in a full acoustic rescoring of the pinched lattice $\tilde{\mathcal{W}}$. For each class j, the models θ_j are used for the words in the segment sets $W_i \in C_j$; the first-pass models are used for words in segment sets not in C.

3. Performance and Error Analysis

We present results on the OGI Alpha-Digits task [15]. This is a fairly challenging small vocabulary task on which we the ML trained system has a relatively high baseline Word Error Rate (WER) (approx. 10%). This ensures that we have a significant number of errors to identify and correct. The baseline system is built using the HTK Toolkit [16]. The training set consists of 46,730 utterances. The data is parameterized as 13 element MFCC vectors with first and second order differences. The baseline ML whole word models contain 12 mixtures per state. The AT&T Large Vocabulary Decoder [17] was used to generate word lattices for the training set which were then transformed into word posteriors based on the lattice total acoustic score. MMI is then performed at the word level using the word time boundaries taken from the lattices. The Gaussian means and variances are updated as described by Woodland and Povey [10] (Sec. 3, scheme *ii* with $E = 2$). The Alpha-Digits task does not have a specific language model, thus recognition both for MMI lattice generation and test set decoding is performed using an unweighted word loop over the vocabulary. The test set consists of 3,112 utterances. ASR performance for 5 iterations of MMI training is presented in Figure 2. Significant improvement over the baseline can be obtained by MMI: the initial ML performance of 10.7% WER is reduced to 9.07% before overtraining is observed in the WER.

3.1. Unsupervised Selection of Segment Sets

Both SMBR-DT and pinched lattice MMI require lattice segmentation; lattices are generated using MMI-3 models and fixed for all subsequent rescoring and training iterations. As described earlier, we obtain segment sets by aligning lattice paths to the MAP hypothesis [11]. We use a particular version of the algorithm, known as "Period-1" cutting. This yields segment sets that contain word sequences of length at most one word, as described in Section 3.1 and the middle panel of Figure 1. This is suboptimal in that better WER can be obtained by optimizing the cutting period [11], however the Period-1 case is the simplest to study. We further simplify the problem by restricting the segment sets to contain only two competing word sequences; segment sets that occur less than 10 times are discarded. This process is performed on both the test and training set lattices. We use these two collections to identify the 50 test segment sets that were also observed most frequently in training. In this way we identify a final collection of segment sets that are likely to contain recognition errors and that also occur
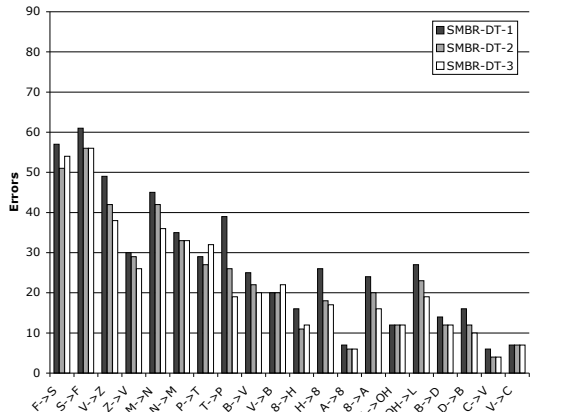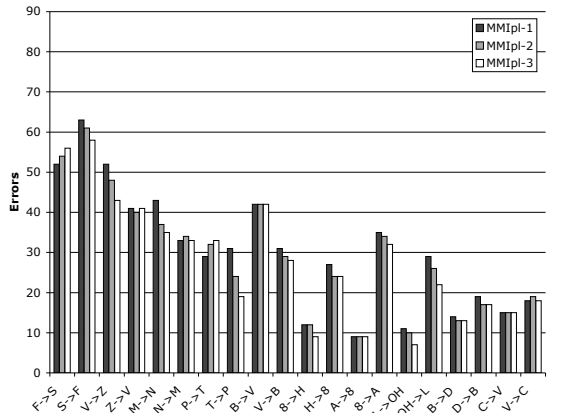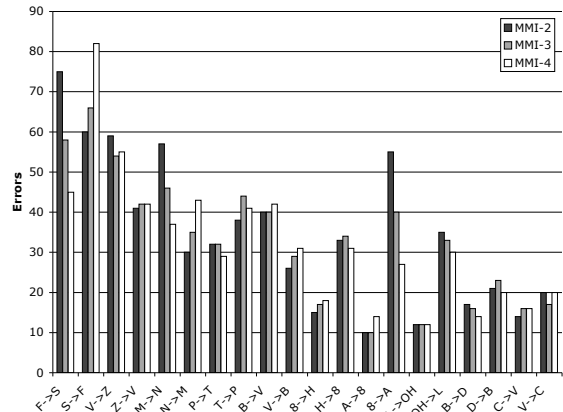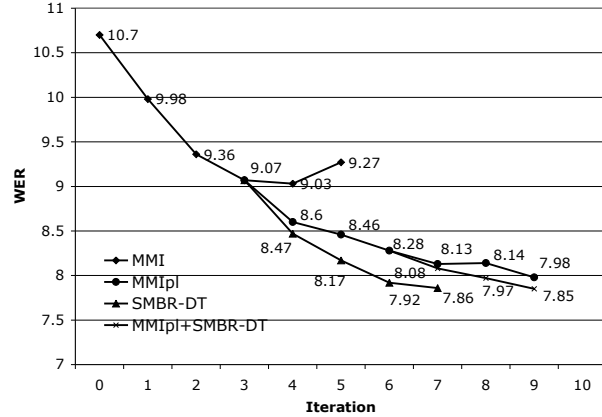

Figure 2: WER and Class-Specific Errors for MMI, Pinched Lattice MMI, and SMBR-DT.

frequently in the training set. The 10 most frequent sets can be found in Figure 2 (see also [8]).

We emphasize that the process of selecting confusion sets is unsupervised. The effectiveness of the approach depends on the unsupervised selection of segment sets and the reliability with which they can be associated with ASR errors. The identification of ASR errors through confidence measurements is well-established [3, 14]. Our work follows these approaches and we have verified it does indeed identify recognition errors consistently and reliably (see [8], Section 4.1).

3.2. Rescoring After Pinched Lattice MMI and SMBR-DT

Models trained after three MMI iterations (MMI-3) were used to initialize the pinched lattice MMI and SMBR-DT procedures. We observe in Figure 2 that the iterations of pinched lattice MMI estimation (MMIpl) yield continued improvement in WER. This is in sharp contrast to "regular lattice" MMI which shows evidence of overtraining beyond the fourth 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 therefore be attributed to the use of lattice pinching in MMI estimation to refine the space of competing hypotheses. SMBR-DT yields further improvements, whether initialized by MMI or by pinched lattice MMI (MMI-pl+SMBR-DT); these results suggest that the pinched lattice hypothesis space refinement and the use of specialized discriminative models in rescoring are complementary.

3.3. Within-Class Error Analysis

We now present an analysis of the substitution errors made in rescoring with models trained with MMI, pinched lattice MMI, and SMBR-DT procedures. Ideally, all error types should decrease over each of the 3 training iterations shown in Figure 2. However, despite the overall reduction in WER achieved by MMI training, error types are not reduced uniformly. For example, the decrease in $F{\rightarrow}S$ indicates that the number of times F is recognized as S decreases sharply over the 3 MMI iterations. However, the complementary plot of $S{\rightarrow}F$ indicates that this takes place at the cost of introducing errors in which S is recognized as F. We find that this undesirable behavior is greatly reduced by the pinched lattice MMI models and nearly entirely eliminated with the SMBR-DT models. In Section 2 we note that MMI is associated with Sentence Error Rate, while pinched lattice MMI and SMBR-DT are motivated by minimization of WER. This provides experimental support in favor of estimation procedures that are tuned to the task performance measure.

4. Conclusion

We have shown that lattice segmentation and estimation techniques based on empirical risk minimization can be integrated into a divide-and-conquer strategy for ASR. We introduce pinched lattice MMI which yields improved performance relative to MMI by focusing on errors identified in the training set and does so without any introduction of additional model complexity. Both pinched lattice MMI and SMBR-DT are shown to yield improved performance over MMI not only in overall WER but also in individual substitution errors.

The potential gains in our experiments are limited by our use of binary confusion sets; this choice was made to simplify analysis and presentation. We note anecdotally that further WER improvements are possible by using more and richer confusion sets.

5. References

[1] A. Stolcke, Y. Konig, and M. Weintraub, "Explicit Word Error Minimization in N-Best List Rescoring," in *Eurospeech*, Rhodes, Greece, 1997.

[2] V. Goel and W. Byrne, "Minimum Bayes-Risk Automatic Speech Recognition," *Comp. Spch. & Lang.*, vol. 14(2), 2000.

[3] J. Fiscus, "A Post-processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER)," in *IEEE Wkshp. Spch. Recog. & Und.*, 1997.

[4] E. Brill L. Mangu and A. Stolcke, "Finding consensus in speech recognition: Word Error Minimization and other applications of confusion networks," *Comp. Spch. & Lang.*, vol. 14(4), 2000.

[5] V. Goel and W. Byrne, "Minimum Bayes-risk automatic speech recognition," in *Pattern Recognition in Speech and Language Processing*, W. Chou and B.-H. Juang, Eds. CRC Press, 2003.

[6] V. Goel, S. Kumar, and W. Byrne, "Segmental Minimum Bayes-Risk ASR Voting Strategies," in *ICSLP*, Beijing, China, 2000.

[7] V. Goel, S. Kumar, and W. Byrne, "Confidence Based Lattice Segmentation and Minimum Bayes-Risk Decoding," in *Eurospeech*, Aalborg, Denmark, 2001.

[8] V. Doumpiotis, S. Tsakalidis, and W. Byrne, "Discriminative Training for Segmental Minimum Bayes Risk Decoding," in *ICASSP*, 2003.

[9] Y. Normandin, "Maximum Mutual Information Estimation of Hidden Markov Models," in *Automatic Speech and Speaker Recognition: Advanced Topics*, C.-H. Lee, F. K. Soong, and K. Paliwal, Eds. Kluwer, 1996.

[10] P. C. Woodland and D. Povey, "Large Scale Discriminative Training for Speech Recognition," in *Proc. ITRW ASR*. ISCA, 2000.

[11] S. Kumar and W. Byrne, "Risk Based Lattice Cutting for Segmental Minimum Bayes-Risk Decoding," in *ICSLP*, Denver, Colorado, USA, 2002.

[12] J. Kaiser, B. Horvat, and Z. Kačič, "A Novel Loss Function for the Overall Risk Criterion Based Discriminative Training of HMM Models," in *ICSLP*, Beijing, China, 2000.

[13] D. Povey and P. C. Woodland, "Minimum Phone Error and I-Smouthing for Improved Discriminative Training," in *ICASSP*, 2002.

[14] T. Hain, P.C. Woodland, T.R. Niesler, and E.W.D. Whittaker, "The 1998 HTK System for Transcription of Conversational Telephone Speech," in *ICASSP*, 1999.

[15] M. Noel, "Alphadigits," CSLU, OGI, 1997, http://www.cse.ogi.edu/CSLU/corpora/alphadigit.

[16] S. Young et. al., *The HTK Book, Version 3.0*, July 2000.

[17] Mehryar Mohri, Fernando Pereira, and Michael Riley, "Weighted Automata in Text and Speech Processing," in *ECAI*, 1996.