DISCRIMINATIVE TRAINING FOR SEGMENTAL MINIMUM BAYES RISK DECODING

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

A modeling approach is presented that incorporates discriminative training procedures within 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. Acoustic models specialized to discriminate between the competing words in these classes 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

The limitations of the Maximum Likelihood Estimation (MLE) procedures widely used in Hidden Markov Model (HMM) speech recognition systems are well known. One of the most commonly cited problems is the violation of the model correctness assumption. Parameterized models obtained via MLE can be employed optimally for detection and classification if the data encountered is generated by some distribution from the model family. The problem arises due to the various conditional independence assumptions that underlie HMM models. Given these assumptions, it is unlikely that the processes that actually generate speech can be closely modeled by HMMs. Therefore ML estimation of HMMs cannot be relied upon to yield models that are optimum for ASR.

As an alternative to relying on the asymptotic behavior of ML estimation under the model correctness assumption, there are modified estimation and decoding procedures that directly attempt to optimize ASR performance criteria. This paper describes a modeling framework that unifies and extends two such modeling approaches, Maximum Mutual Information (MMI) estimation and Minimum Bayes Risk (MBR) decoding.

2. DISCRIMINATIVE ESTIMATION AND DECODING

Maximum Mutual Information estimation [1, 2] attempts to improve the likelihood of the correct sentence hypothesis given the acoustic evidence. Given a labeled training set of word sequences and acoustic observations $\{W, A\}$, MMI iteratively optimizes the model parameters θ to increase $P(W|A;\theta)$ over $\mathcal{W}$, which is usually taken to be the set of all word strings allowed in the language. This training objective is directly related to reducing the Sentence Error Rate on the acoustic training set. This immediately suggests that beyond the usual difficulties of ensuring that performance obtained in training generalizes to the test set, there may also be issues in generalization under different performance criteria. While Sentence Error Rate is in some sense the ultimate performance criterion, there may be value in estimation procedures that minimize other criteria, such as Word Error Rate [3, 4].

Similar issues arise in the maximum a-posteriori (MAP) decoding criterion implemented by the Viterbi procedure. MAP decoding, which given an utterance A produces a sentence hypothesis according to $\tilde{W} = \text{argmax}_{W \in \mathcal{W}} P(W|A)$, is the optimum decoding criterion when performance is measured under the Sentence Error Rate criterion. However for other criteria, again such as Word Error Rate, other decoding schemes may be better.

2.1. Segmental Minimum Bayes-Risk Decoders

With this motivation, Minimum Bayes-Risk decoders [5, 6] 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 as

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

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 [5, 6].

Segmental Minimum Bayes Risk decoding was 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 decision rules

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

where $\hat{W}$ is the concatenation of $\hat{W}_i$, $i = \{1, 2, ..., N\}$, from which the term Segmental Minimum Bayes Risk follows.

There are a variety of possible segmentation schemes. Here we segment the lattice word strings by aligning each path in the lattice to the MAP sentence hypothesis [7, 8]: given the MAP hypothesis $\tilde{W}$, we segment the paths in the lattice to attain $l(\tilde{W}, W') =$

This work was supported by the National Science Foundation under Grant No. #IIS-9982329 and Grant No. #IIS-0122466.

$\sum_{i=1}^{N} l(\tilde{W}_i, W_i')$. This segmentation procedure is performed carefully so as to retain the structure of the original lattice in regions of low confidence [8].

2.1.1. Search Space Refinements

This procedure 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. In regions of high confidence, the search space is restricted to follow the MAP hypothesis itself. Because the structure of the original lattice is retained whenever we want to consider alternatives to the MAP hypothesis, we can perform acoustic rescoring over this pinched lattice.

2.1.2. Refined Discriminative Training for SMBR Decoding

We have the opportunity to train and apply extremely refined acoustic models trained specifically to resolve the errors encountered in the test set. In previous approaches to MBR, $P_i(W|A)$ was found via a lattice forward-backward procedure [7] using fixed likelihood scores obtained from the original ASR system. Even if this system was trained using MMI, it is still intended to discriminate between all sentences in the language that might be uttered.

Rather than derive these posteriors from general acoustic models, our goal is to estimate each $P_i(W|A)$ so that it is optimized for the distinct recognition problem to which it will be applied: $P_i(W|A)$ will be trained only to discriminate word sequences in $\mathcal{W}_i$. There are two problems here that arise. The first is the appropriate training criterion. The second is to find relevant training data. SMBR allows us to address them simultaneously. We generate lattices on the acoustic training set, and perform lattice segmentation with respect to the true transcription. This identifies patterns of recognition errors within the training set. Given a particular error pattern found in the test set, we can use training data associated with similar errors to train a discriminative model.

In summary, our goal is to develop a joint estimation and decoding procedure that improves over MMI. After an initial MAP decoding pass with MMI models, for each utterance we use lattice cutting to produce pinched lattices that identify the segment sets that are likely to contain recognition errors. We then turn to the training set to find all relevant data that can be used to train models $P_i(W|A)$ to pick the correct hypothesis from these segment sets. We finally apply these models in a full acoustic rescoring of the pinched lattice by applying each $P_i(W|A)$ in decoding over the appropriate segment set.

3. MMI BASELINE PERFORMANCE

To develop the basic estimation and decoding mechanisms, we present results on the OGI Alpha-Digits task [9]. This is a fairly challenging small vocabulary task on which we still encounter a relatively high baseline WER (approx. 10%). This ensures that we have a significant number of errors to identify and correct. We begin by presenting the MMI baseline system and analyzing its performance and the errors it makes.

Error Pairs		$\bar{c}_0^{(3)}$	$\bar{c}_1^{(3)}$	Error Pairs		$\bar{c}_0^{(3)}$	$\bar{c}_1^{(3)}$
1.	F+S	58	60	6.	8+H	17	34
2.	V+Z	54	42	7.	A+8	10	40
3.	M+N	45	35	8.	L+OH	12	33
4.	P+T	32	44	9.	B+D	16	23
5.	B+V	40	29	10.	C+V	16	17

Table 1. Dominant Confusion Pairs in Unconstrained Recognition after Three MMI Iterations.

The baseline system is built using the HTK Toolkit [10]. The data is parameterized as 13 element MFCC vectors with first and second order differences . The training set consists of 46,730 utterances. The baseline maximum likelihood models contain 12 mixtures per state, estimated according to the usual HTK training procedure. Starting from these models, several iterations of MMI estimation were performed. The AT&T Large Vocabulary Decoder [11] was used to generate lattices for the training set where are 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 test set consists of 3,112 utterances. 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. Table 3, Row 1 shows that significant improvement over the baseline can be obtained by MMI: the initial ML performance of 10.7% WER is reduced to 9.07%.

We now look closely at the changes in errors as MMI training proceeds. Table 1 presents the most frequently confused words ('*confusion pairs*') observed after three iterations of MMI estimation. Iteration 3 is chosen because MMI performance is nearly optimal at that point. We tabulate errors over each word in each class. The notation $\bar{c}_0^{(3)}(1) = 58$ indicates that there are 58 instances in which F is incorrectly recognized as S, and $\bar{c}_1^{(3)}(1) = 60$ indicates that there are 17 instances in which S is incorrectly recognized as F. The superscript indicates the MMI iteration.

As indicated in Table 3, overall WER does decrease as MMI training progresses. However, when the confusion pairs are monitored individually, it becomes apparent that the improvement is not uniform. Figure 1 tracks the change in confusion pair counts relative to performance at MMI iteration 2. The top plot indicates that $c_0^{(3)}(1)$ (the number of times F is misrecognized as S) decreases by 18 in going from the second to third MMI iteration, and by 30 in the fourth iteration. However, $c_1^{(4)}(1) - c_1^{(2)}(1)$ is positive and larger than $c_1^{(3)}(1) - c_1^{(2)}(1)$ which is also positive, which indicates that the improved recognition of F comes at the expense of errors in the recognition of S. Ideally, all these changes should be negative. However, that behavior is not guaranteed by the MMI training procedure, which is free to introduce performance degradation over individual confusion pairs so long as the overall sentence posterior score improves.

4. CONFUSION PAIRS VIA LATTICE CUTTING

The identification of ASR errors through confidence measurements is well-established [12, 13], and our training approach builds on this work. We need to establish first that lattice cutting finds segment sets that are similar to the dominant confusion pairs observed in MMI decoding. We also need to establish that the segment sets

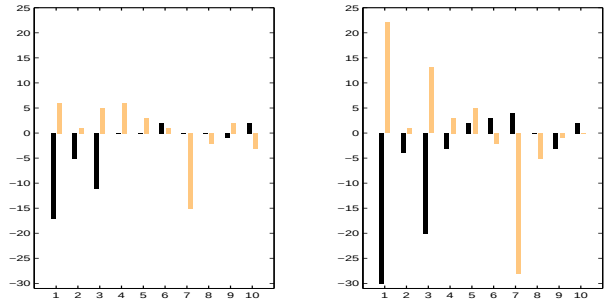

Fig. 1. Confusion Pair Errors in MMI Decoding. *Left*: $\bar{c}^{(3)}(k) - \bar{c}^{(2)}(k)$; *Right*: $\bar{c}^{(4)}(k) - \bar{c}^{(2)}(k)$. The abscissa k is the confusion pair index given in Table 1. For each confusion pair index, $\bar{c}_0^{(i)}(k) - \bar{c}_0^{(2)}(k)$ is given in the left (black) bar and $\bar{c}_1^{(i)}(k) - \bar{c}_1^{(2)}(k)$ is given in the right (white) bar.

identified in the test set are also found consistently in the training set. If these two conditions hold, there is the possibility of training discriminative models on the segment sets in the training data and applying them to the test data to resolve the dominant errors remaining after MMI training.

We establish the first point by comparing the dominant MMI confusion pairs in Table 1 with the test set segment sets found in Table 2 by lattice cutting. There is good agreement among the top eight sets identified in each case, after which there is some divergence. A similar relationship holds between the segment sets identified in test and training reported in Table 2.

4.1. Unsupervised Selection of Segment Sets

As described earlier we obtain segment sets by aligning lattice paths to the MAP hypothesis [8]. 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 in the middle panel of Fig. 2. This is suboptimal in that better WER is can be by optimizing the cutting period [8], 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.

The process starts by identifying the MAP path in a first-pass ASR lattice (Fig. 2, *Top*). Period-1 risk-based lattice cutting is used to reduce the lattice to a sequence of segment sets. In some regions only the MAP path remains (Fig. 2, *Middle*); each arc also contains

Test Set		Count	Training Set		Count
1	F+S	1089	1	F+S	15197
4	P+T	843	4	P+T	10744
6	8+H	784	6	8+H	10370
3	M+N	772	3	M+N	10242
2	V+Z	557	2	V+Z	8068
9	B+D	389	9	B+D	5996
8	L+OH	343	8	L+OH	5108
5	B+V	314	5	B+V	4963
-	A+K	292	-	5+I	4413
-	5+I	289	-	J+K	3653

Table 2. Frequent confusion pairs found by lattice cutting. Indices provided for pairs in the dominant MMI confusable pairs.

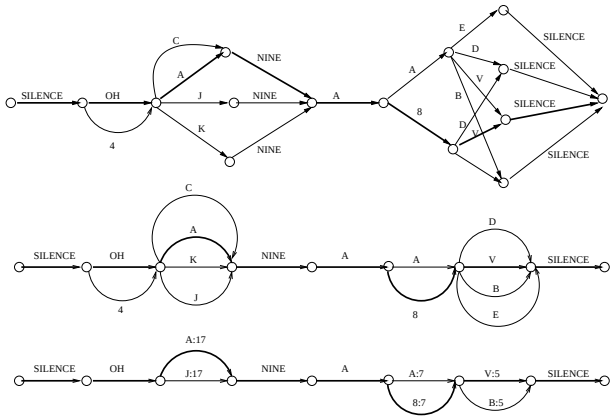

Fig. 2. Lattice Segmentation for Estimation and Search. *Top*: First-pass lattice of likely sentence hypotheses with MAP path in bold; *Middle*: Alignment of lattice paths to MAP path; *Bottom*: Refined search space $\tilde{\mathcal{W}}$ consisting of segment sets selected for discriminative training.

a word posterior derived from the original lattice. Segment sets that occur less than ten times are discarded.

We then perform the same process on the training set to obtain a collection of segment sets representative of recognition errors found in the training data. 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 frequently in the training set.

The final step in the search space refinement is to restrict the segment sets initially identified in the test set to the final 50 that also occur frequently in the training set (Fig. 2, *Bottom*). Some segment sets not in the final collection (e.g. OH+4) are discarded.

The word hypotheses in the refined search space are identified by the segment set to which they belong. This makes it simple to perform discriminative training and to apply the discriminatively trained models appropriately in rescoring. There will be several models for A, for instance. The model A:17 will be used whenever the word hypothesis A is found in segment set 17. Model A:17 is trained to distinguish A's from J's, and is therefore different from A:7, which is trained to distinguish A's from 8's.

5. SMBR TRAINING AND DECODING

Our goal is to perform SMBR as described in Equation 2 using models $P_i(W|A)$ trained to minimize the expected loss over hypotheses drawn from $\mathcal{W}_i$. The estimation is difficult in general, although procedures are available [3, 4]. However Period-1 lattice cutting reduces this problem to MMI estimation over the competing word hypotheses in $\mathcal{W}_i$. This can be seen simply by noting that the loss function over the strings in $\mathcal{W}_i$ is the 1-0 loss function (trivially) consistent with Levenshtein distance between strings of length 1. The minimum risk decoder is therefore the MAP decoder, and empirical risk is minimized by maximizing the likelihood of the correct hypothesis.

We use MMI to estimate word models $P_i(A|W)$ for $W \in \mathcal{W}_i$. Models are initialized using word models trained by three 'normal' MMI iterations (MMI-3 models). The $P_i(A|W)$ are refined using the training set segments identified for each $\mathcal{W}_i$, as described in

Iteration	0	1	2	3	4	5
MMI	10.7	9.98	9.36	9.07 $*$	9.03	9.27
DT+SMBR	$*$	8.47	8.17	8.01	7.92	7.86

Table 3. MMI vs. SMBR Training and Decoding in WER(%).

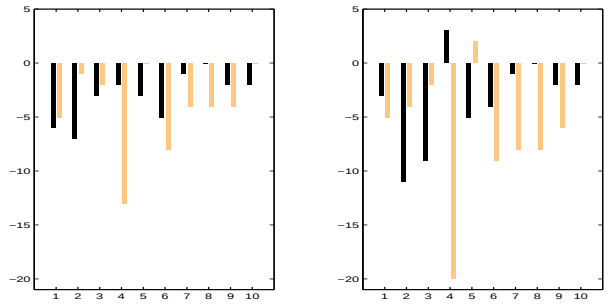

Fig. 3. Confusion Pair Errors in DT+SMBR Decoding. *Left*: $\bar{c}^{(4)}(k) - \bar{c}^{(3)}(k)$; *Right*: $\bar{c}^{(5)}(k) - \bar{c}^{(3)}(k)$. (see Fig. 1 caption).

the previous section. The training objective for each set of distributions is to maximize $P_i(A|W)/\sum_{W' \in \mathcal{W}_i} P_i(A|W')$, which is done using MMI over the appropriate training set segments.

The Period-1 cutting used to identify the segment sets also simplifies the SMBR decoding procedure of Equation 2. In a similar way as was observed in the estimation problem, rescoring is simply Viterbi search over the refined search spaces $\tilde{\mathcal{W}}$. When the search space is constrained to follow the MAP hypothesis, the MMI-3 models are used. In regions of the search space corresponding to a segment set $\mathcal{W}_i$, models $P_i(A|W)$ are used.

The results of SMBR training and decoding are given in Table 3. We first discuss the search space refinements. We performed the 'sanity check' of rescoring the pinched lattices with the MMI-3 models: performance was identical to unconstrained rescoring. This verifies that the search space refinement introduces no new errors. Pinching does reduce the lattice search space substantially, however. The Lattice Word Error Rate of the original lattices is 1.27%, which increases to 3.11% after pinching. Despite this restriction in the search space, we still see more than a 1% WER reduction beyond the best MMI performance. We also note that the discriminatively trained models are inextricably bound up with the SMBR segmentation process. Performance degrades drastically if these models are used in unconstrained search decoding pass.

Finally, we note that the improvement over the confusion pairs is more uniform than under MMI estimation. Figure 3 shows that nearly all the error counts are decreasing over all words within the confusion classes. Overall performance gains found with SMBR are not being achieved at the expense of words in individual classes.

6. CONCLUSION

We have presented an ASR modeling framework that incorporates discriminative training in SMBR rescoring. It is a divide-and-conquer approach to identifying and eliminating ASR errors. SMBR decoding is used first to identify distinct regions in the search space that are likely to contain errors, and then used in rescoring with models trained specifically to resolve these errors. We have shown on a small vocabulary recognition task that this refinement of the search space allows us to improve the effectiveness of the widely used MMI estimation procedure.

Casting the ASR problem as a minimum Bayes-risk decision problem provides a rigorous framework for the integration of discriminative search and estimation procedures. Although we have selected a simple recognition task to develop and present our approach, our ultimate goal is apply these techniques to large vocabulary ASR. Due to the great diversity of ASR errors in large vocabulary tasks, we expect the primary challenge to be robust estimation of discriminative models from sparse training data. We expect that constrained, discriminative estimation procedures will prove useful in these problems [14].

Acknowledgments We gratefully acknowledge discussions with S. Kumar in formulating these experiments. We also thank M. Riley and M. Saraclar of AT&T Research for assistance with the FSM tools and ASR decoder.

7. REFERENCES

[1] Y. Normandin, "Maximum Mutual Information Estimation of Hidden Markov Models," in *Automatic Speech and Speaker Recognition: Advanced Topics*, Chin-Hui Lee, Frank K. Soong, and Kuldip K. Paliwal, Eds. Kluwer, 1996.

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

[3] 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.

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

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

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

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

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

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

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

[11] M. Mohri, F. Pereira, and M. Riley, *AT&T General-purpose Finite-State Machine Software Tools*, 2001, [Online]. Available: http://www.research.att.com/sw/tools/fsm/.

[12] 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.

[13] 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.

[14] S. Tsakalidis, V. Doumpiotis, and W. Byrne, "Discriminative Linear Transforms for Feature Normalization and Speaker Adaptation in HMM Estimation," in *ICSLP*, 2002.