LVCSR RESCORING WITH MODIFIED LOSS FUNCTIONS: A DECISION THEORETIC PERSPECTIVE

Vaibhava Goel, William Byrne, Sanjeev Khudanpur

Center for Language and Speech Processing
The Johns Hopkins University
Baltimore, MD 21218
{vgoel, byrne, sanjeev}@mail.clsp.jhu.edu

ABSTRACT

The problem of speech decoding is considered here in a Decision Theoretic framework and a modified speech decoding procedure to minimize the expected risk under a general loss function is formulated. A specific word error rate loss function is considered and an implementation in an N-best list rescoring procedure is presented. Methods for estimation of the parameters of the resulting decision rules are provided for both supervised and unsupervised training. Preliminary experiments on an LVCSR task show small but statistically significant error rate improvements.

1. INTRODUCTION

The MAP decision rule used widely in statistical speech recognition [1] finds hypotheses according to

$$\hat{W} = \underset{W \in \mathcal{W}}{\operatorname{argmax}} P(X|W)P(W). \tag{1}$$

In this formulation the transcription W of an utterance is assumed to have been selected from a set of all possibilities $\mathcal{W}$ under the language model $P(W)$ which in the absence of acoustic evidence serves as a prior. For a given utterance the acoustics X are assumed to have been generated according to the likelihood $P(X|W)$.

In the Bayesian Decision Theoretic framework, the MAP rule is obtained by first considering real-valued, usually non-negative, loss functions $l(W, \delta(X))$ that penalize incorrect estimates of W produced by a decision rule $\delta(X)$

$$\delta(X) : X \rightarrow \mathcal{W}. \tag{2}$$

The decision rule minimizing the Bayes Risk

$$B(\delta(X)) = E_{P(W)}[E_{P(X|W)}[l(W, \delta(X))]] \tag{3}$$

is given by [2]

$$\delta(X) = \underset{W \in \mathcal{W}}{\operatorname{argmin}} \sum_{W' \in \mathcal{W}} l(W', W)P(W'|X). \tag{4}$$

This decision rule is known as Bayes decision rule. It is straightforward to show that the MAP decision rule of Equation 1 minimizes the Bayes Risk under the 0-1 valued loss function

$$l_{SER}(W, \delta(X)) = \begin{cases} 0 & \text{if } \delta(X) = W, \\ 1 & \text{otherwise.} \end{cases}$$

Since this function penalizes all incorrect sentences equally, it is termed the sentence error loss.

The shortcomings of this minimum sentence error rate MAP decision rule formulation have been long known [5] and have been addressed recently in [6]. One problem is that the l_{SER} loss function is only loosely linked to the recognition word error rate (WER) which is taken here to be the performance measure of interest. It assigns a loss of 1 for any incorrect classification without regard to the number of words found correctly and is therefore a fairly crude measure except when recognizer performance is very good. It is thus desirable to consider more general loss functions that take an appropriate form for the task at hand. If the intent is to minimize the WER, for example, $l(W, \delta(X)) = WER(W, \delta(X))$ may be an appropriate loss function; Stolcke *et al.* [6] have reported an N-best list rescoring algorithm which can be thought of as a Bayes decision rule under this criterion. In systems designed for speech understanding applications, where the *contents* of the speech are valued more than the exact word string recognized, a loss function might penalize according to semantic distance, *e.g.*,

$$l(\text{`hello'},\text{`hello there'}) < l(\text{`hello'},\text{`hello John'}).$$

For all such loss functions the decision rule is found via Equation 4.

Another general problem associated with minimum risk decision rules is that it is necessary to obtain a good estimate of the posterior distribution $P(W|X)$. The model architectures, parameterization, and training procedures needed to obtain accurate estimates of this distribution are not yet available for large speech recognition problems.

To address these two problems we propose here a decision rule for decoding under the minimum Bayes Risk associated with a WER-based loss function. The formulation of this modified decoding criterion is described first and its implementation as an N-best rescoring procedure is given. The loss-function is itself parameterized so that the resulting decision rule can be tuned to optimize word-error rate over a held-out training set. A method for the unsupervised optimization of these decision rule parameters on a test set is also described. Results on the Switchboard LVCSR task [3] are then given. While some modest but consistent improvements are reported, the main purpose of this paper is to present a framework in which model and decision rule parameters can be estimated under a minimum risk criterion that permits loss functions other than the sentence error rate and to motivate this approach by deriving and evaluating a decision rule under a modified WER loss function.

2. MODIFIED LOSS FUNCTIONS AND MINIMUM RISK DECODING

Direct derivation of decision rules from Equation 4 is impractical for at least two reasons: $P(W|X)$ is unknown; and both the search for the minimizer and the sum over all sentences for each candidate are impractical to implement in most cases. A consequence of not knowing the true posterior distribution $P(W|X)$ is that even if a loss function appropriate to the task such as WER is given, it may not be optimal to use it directly as $l(\cdot,\cdot)$ in Equation 4. The same argument, of-course, holds if $P(W|X)$ is known but the exact minimization is not carried out instead. We therefore propose that a class of loss functions be used which depend on the task-specific loss function but have an additional degree of freedom. This additional degree of freedom, suitably parameterized, may then be "tuned" for a given estimate of $P(W|X)$ and a given approximation of the minimization.

Modified loss functions

The focus of this paper is a set of decision rules that minimize risks associated with loss functions based on the word error rate. These functions are written as $l(W,\delta(X)) = f_{WER}(W,\delta(X))$. As discussed, fixed loss functions of any form may be not be appropriate, so the following simple alternative is proposed

$$l(W,\delta(X)) = [WER(W,\delta(X))]^x. \tag{5}$$

$WER(W,\delta(X))$ is the number of word errors between W and a hypothesis $\delta(X)$. The 'tilt' or exponential discounting parameter x is not specified beforehand but will be adjusted using observed data. The intent is to vary this parameter to minimize WER on available data directly.

N-best list rescoring

Decision criteria based on loss functions $f_{WER}(W,\delta(X))$ have the form

$$\delta(X) = \arg\min_{W\in\mathcal{W}} \sum_{W'\in\mathcal{W}} f_{WER}(W',W)P(W'|X).$$

If $f_{WER}(W,W) = 0$, then for each $W \in \mathcal{W}$ there is a term missing from the sum due to $W = W'$. Therefore it is plausible to assume that candidate hypotheses with higher $P(W|A)$ yield a smaller sum than those with lower $P(W|A)$. Should this admittedly optimistic assumption hold, the minimizer is likely to be among candidates with a high posterior probability. Hence it may suffice to perform the minimization only over relatively likely candidates, *e.g.*, the candidates present in a recognition N-best list. This list is denoted $\mathcal{W}_{hl}$ and the resulting decision rule is

$$\delta(X) = \arg\min_{W\in\mathcal{W}_{hl}} \sum_{W'\in\mathcal{W}} f_{WER}(W',W)P(W'|X).$$

A second, less optimistic, assumption is that for candidates W close to the true minimizer, the sum over $W' \in \mathcal{W}$ is well approximated by a small number of dominant terms – the N-best hypotheses generated by a moderately good recognizer. To see the plausibility of this assumption, observe that the acoustic likelihoods, and hence $P(W'|X)$, for word sequences W' vastly different from the true minimizer are expected to be smaller by orders of magnitude than those of the N-best hypotheses. Their contribution to the sum may therefore be ignored even if the corresponding $f_{WER}(W',W)$ is large. If $\mathcal{W}_{N-best}$ denotes a suitably large N-best list then, this amounts equivalently to assuming that

$$\sum_{W'\in\mathcal{W}-\mathcal{W}_{N-best}} f_{WER}(W',W)P(W'|X)$$

does not vary for $W \in \mathcal{W}_{hl}$.

Under these two simplifying assumptions the decision rule becomes

$$\delta(X) = \arg\min_{W\in\mathcal{W}_{hl}} \sum_{W'\in\mathcal{W}_{N-best}} f_{WER}(W',W)P(W'|X). \tag{6}$$

Supervised optimization of the loss function

As discussed thus far, the decision rule requires computation of quantities involving the distribution $P(W|X)$ which for acoustic HMMs can be approximated by $\frac{P(X|W)P(W)}{P(X)}$. As discussed in [6] the estimation of $P(X)$ can be difficult in both computational and modeling complexity as it requires finding $P(X) = \sum_{W'} P(X|W')P(W')$. However, it is possible to avoid this difficulty by minimizing the empirical Bayes Risk on available data. Note first that the recognition criterion found so far is equivalent to

$$\delta(X) = \arg\min_{W\in\mathcal{W}_{hl}} \sum_{W'\in\mathcal{W}_{N-best}} f_{WER}(W',W)P(X,W') \tag{7}$$

since $P(X)$ doesn't affect the minimization. The remaining likelihood scores $P(X,W)$ are available in the N-Best lists. The word-insertion penalties and grammar scale factors are unchanged from the values used during recognition; these parameters are assumed to be tuned already to the task.

The decision rule is further modified by adding a second tuning parameter so that the joint likelihood is of the form $P(X,Y)^{1/y}$. For the loss function given in Equation 5, the resulting decision rule is then parameterized by x and y so that

$$\delta_{x,y}(X) = \arg\min_{W\in\mathcal{W}_{hl}} \sum_{W'\in\mathcal{W}_{N-best}} WER(W',W)^x P(X,W')^{1/y}. \tag{8}$$

Rather than attempting to minimize the expected risk by estimating $P(W|X)$ and plugging it in to Equation 6, the form of the decision rule given in Equation 8 is taken and all necessary optimization of parameters is done to reduce the empirical risk by minimizing the risk

$$\frac{1}{|\mathcal{T}|} \sum_{(W_i,X_i)\in\mathcal{T}} WER(W_i,\delta_{x,y}(X_i)) \tag{9}$$

over a database $\mathcal{T} = \{(W_i,X_i)\}$ of labeled utterances. This optimization is performed using a grid search to find x and y.

This is a hybrid approach in that it does not attempt to avoid all estimation of the underlying distribution by finding decision rule parameters through the minimization of empirical risk [7]. Instead, this approach makes use of existing models when they are in convenient form and relatively reliable; all remaining decision rule parameters are then found to optimize measured performance directly on available data.

Unsupervised optimization

For some problems, notably speaker and channel adaptation, it may be desirable to optimize decision rule parameters without using a separate training set. In this case, the empirical risk is optimized taking the "truth" as the maximum likelihood recognition

hypothesis. For the loss function considered here, the parameters x and y are found to optimize

$$\frac{1}{|\mathcal{T}'|} \sum_{X_i \in \mathcal{T}'} WER(\hat{W}_i, \delta_{x,y}(X_i)) \quad (10)$$

over a test set $\mathcal{T}'$, where $\hat{W}_i = \text{argmax}_{W \in \mathcal{W}} P(W, X_i)$. This does not require generation of N-Best lists for held-out training data and, as will be described, has been found to work almost as well as supervised optimization for the problems studied.

3. EXPERIMENTS AND RESULTS

The proposed rescoring algorithm was evaluated in two experiments on the Switchboard LVCSR task [3]. The test consisted of 2427 utterances from 14 conversations (38 sides) that formed the dev-test at the 1997 Johns Hopkins University LVCSR Workshop; full details of the test set definition, language models used, and other details are given in [9]. The first experiment consisted of rescoring N-best lists obtained from bigram word lattices generated using an HTK-based 12-mixture speaker and gender independent cross-word triphone system [8] with 6973 triphone states. These lattices had a lattice word error rate of 9.6% with 0.9% OOVs. N-best lists denoted Set-I were then generated to a depth of 1000. It was not possible to find 1000 hypotheses for every utterance; the average list depth was 951.

To obtain the N-best lists rescored in the second experiment, bigram lattices were first generated using the above HMMs; the bigram lattices were pruned to reduce the number of lattices nodes to 1/3 of the original number, yielding lattices with a lattice word error rate of 10.1% and 0.8% OOVs. These lattices were then rescored using a trigram language model and HMMs adapted for each speaker using exponential warp maximum likelihood vocal tract normalization and global MLLR [4] based on the 1-best bigram hypotheses. N-best lists denoted Set-II were then generated; the average list depth was 722.

In both experiments the top 25 candidates in the N-best lists were considered for rescoring; i.e. the sets $\mathcal{W}_{hl}$ had 25 candidates. Table 1 lists the oracle word error rates for these two sets of

	WER	
N	Set-I	Set-II
1	44.80	38.50
10	34.80	28.70
20	31.20	26.50
25	31.20	25.90

Table 1: Oracle WER at increasing N-Best list depth.

N-Best lists at increasing depth. These are the minimum WERs attainable for a fixed size $\mathcal{W}_{hl}$ and so are lower bounds on the WER attainable in these experiments.

Supervised optimization and rescoring

For supervised optimization of the decision rule a set of N-best lists was generated for a held-out portion of the Switchboard acoustic training set. This consisted of 2109 gender-balanced utterances from 536 conversations (871 sides). As in the test conditions, the $\mathcal{W}_{hl}$ had 25 candidates and the $\mathcal{W}_{N-best}$ had depths of at most 1000. The decision rule parameters x and y were optimized via a grid search to minimize the WER on this set and the optimum values were found to be x = 1.7 and y = 35.

These parameter values were used in rescoring the two test sets of N-best lists, even though they were optimal only for Set-I. The results given in Table 2 show consistent reduction of WER, although the reduction is slightly less for the better system.

Unsupervised optimization and rescoring

The decision rule parameters x and y were found by optimization of Equation 10 directly over the two test sets of N-best lists. The optimal parameter values for Set-I were found to be x = 1.3 and y = 50. For Set-II they were x = 2.6 and y = 35. Table 2 lists the corresponding results. Note that the gain of about 0.5% carries over from the bigram experiment to the Trigram + ML-VTN + MLLR case and also that the error rate improvements in the unsupervised case are very similar to those in the supervised case for both experiments.

	Bigram no adapt.		Trigram ML-VTN+MLLR	
	SER	WER	SER	WER
Baseline	70.60	44.80	65.90	38.50
Supervised	71.10	43.90	66.70	38.00
% abs change	+0.50	-0.90	+0.80	-0.50
Unsupervised	71.30	44.00	67.10	38.00
% abs change	+0.70	-0.80	+1.20	-0.50

Table 2: WER and SER for minimum risk rescoring.

A final set of experiments was performed to compare the rescoring approach proposed here to that of Stolcke *et al*. [6]. To test the latter, the implementation provided in the SRI Language Modeling Toolkit was used. The size of the candidate set $\mathcal{W}_l$ was kept at 25. Rescoring was performed using both the Set-II N-best lists and a set of depth 2500 N-best lists generated in same manner, i.e. with trigram and global MLLR; the average depth of this latter set was 1661.

Baseline	N = 1000	N = 2500
38.5	38.5	38.4

Table 3: SRI LM Toolkit MWER Rescoring Results

As can be seen from Table 3, the procedure of [6] does not yield significant improvement in these experiments. This is consistent with observations reported in [6], namely that the technique is less effective as WER decreases. This is in contrast to the results given in Table 2, which demonstrate that the technique proposed in this paper remains effective. These results are not a definitive comparison of the two techniques; they are given only to suggest that the procedure proposed here may prove to remain effective as error rates decrease.

4. DISCUSSION AND CONCLUSION

We have proposed a modified decoding rule intended to address the limitations inherent in the MAP decoding criterion and have described its implementation as an N-best rescoring scheme. The explicit form of the decoding rule is derived within the minimum Bayes risk framework. This procedure allows existing model parameters to be used as appropriate and additional parameters that

modify both the loss function of interest and the model-based likelihoods can be added as needed and adjusted to minimize empirical risk on available data. It is argued that this scheme bypasses some computational and modeling difficulties inherent in finding a 'plug-in' minimum risk decoder.

Since the N-best rescoring scheme can be implemented for general loss functions, loss functions can be tailored to the requirements of the recognition task. Although the error rate improvements obtained are less than 1% absolute, the gains seem to be additive with other contributions due to improvements in language models and speaker adaptation. Undoubtedly the quality of the N-best lists to be rescored affects the error rate improvements. A problem of interest is to consider more candidates by enlarging $\mathcal{W}_{hl}$ without large increases in computational cost.

5. ACKNOWLEDGMENTS

Lattice generation and rescoring was performed using software provided by Entropic Cambridge Research Laboratory, Cambridge, UK in support of the 1997 JHU LVCSR Workshop. Assistance from A. Stolcke in obtaining and using of the SRI LM Toolkit is gratefully acknowledged.

6. REFERENCES

[1] L. R. Bahl, F. Jelinek, and R. L. Mercer. A maximum likelihood approach to continuous speech recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 5(2):179–190, 1983.

[2] P. J. Bickel and K. A. Doksum. *Mathematical Statistics: Basic Ideas and Selected topics*. Holden-Day Inc., Oakland, CA, 1977.

[3] J. J. Godfrey, E. C. Holliman, and J. McDaniel. Switch-board: Telephone speech corpus for research and development. *Proc. ICASSP*, pp. 517–520, San Francisco, CA, 1992.

[4] C. J. Leggetter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, pp. 171–185, 1995.

[5] A. Nadas. Optimal solution of a training problem in speech recognition. *IEEE Trans. on Acoustics, Speech, and Signal Processing.*, ASSP-33(1):326–329, 1985.

[6] A. Stolcke, Y. Konig, and M. Weintraub. Explicit word error minimization in N-best list rescoring. *Eurospeech-97*, pp. 163–165, Rhodes, Greece, 1997.

[7] V. Vapnik. *Estimation of dependences based on empirical data*. Springer-Verlag, New York, 1982.

[8] S. Young et al. *HTK 2.1*. Entropic Cambridge Research Laboratory Ltd., Cambridge, UK, 1997.

[9] *Proceedings of the 1997 Large Vocabulary Continuous Speech Recognition Workshop*, Johns Hopkins University, Baltimore, MD, 1997.