EXPLICIT WORD ERROR MINIMIZATION IN N-BEST LIST RESCORING

Andreas Stolcke *Yochai König* *Mitchel Weintraub*

Speech Technology and Research Laboratory
SRI International, Menlo Park, CA, U.S.A.
http://www.speech.sri.com/
{stolcke,konig,mw}@speech.sri.com

ABSTRACT

We show that the standard hypothesis scoring paradigm used in maximum-likelihood-based speech recognition systems is not optimal with regard to minimizing the word error rate, the commonly used performance metric in speech recognition. This can lead to sub-optimal performance, especially in high-error-rate environments where word error and sentence error are not necessarily monotonically related. To address this discrepancy, we developed a new algorithm that explicitly minimizes expected word error for recognition hypotheses. First, we approximate the posterior hypothesis probabilities using N-best lists. We then compute the expected word error for each hypothesis with respect to the posterior distribution, and choose the hypothesis with the lowest error. Experiments show improved recognition rates on two spontaneous speech corpora.

1. INTRODUCTION

The standard selection criterion for speech recognition hypotheses aims at maximizing the posterior probability of a hypothesis W given the acoustic evidence X [1]:

$$\begin{aligned} W^* &= \operatorname*{argmax}_W P(W|X) \\ &= \operatorname*{argmax}_W \frac{P(W)P(X|W)}{P(X)} \qquad (1) \\ &= \operatorname*{argmax}_W P(W)P(X|W) \qquad (2) \end{aligned}$$

Here $P(W)$ is the prior probability of a word sequence according to a *language model*, and $P(X|W)$ is given by the acoustic model. Equation (1) is Bayes' Rule, while (2) is due to the fact that $P(X)$ does not depend on W and can therefore be ignored during maximization. Bayes decision theory (see, e.g., [2]) tells us that this criterion (assuming accurate language and acoustic models) maximizes the probability of picking the correct W; i.e., it minimizes *sentence* error rate. However, speech recognizers are usually evaluated primarily for their *word* error rates.

Empirically, sentence and word error rates are highly correlated, so that minimizing one tends to minimize the other. Still, if only for theoretical interest, two questions arise:

(A) Are there cases where optimizing expected word error and expected sentence error produce different results?

(B) Is there an effective algorithm to optimize expected word error explicitly?

Note that question (A) is not about the difference between word and sentence error in a particular instance of X and its correct transcription, since obviously the two error criteria would likely pick different best hypotheses in any given instance. Instead, we are concerned with the *expected* errors, as they would be obtained by averaging over many instances of the same acoustic evidence with varying true word sequences, i.e., if we sampled from the true posterior distribution $P(W|X)$.

We will answer question (A) first by way of a constructed example, showing that indeed the two error metrics can diverge in their choice of the best hypothesis. Regarding question (B), we develop a new N-best rescoring algorithm that explicitly estimates and minimizes word error. We then verify that the algorithm produces lower word error on two benchmark test sets, thus demonstrating that question (A) can be answered in the affirmative even for practical purposes.

2. AN EXAMPLE

The following is a hypothetical list of recognition outputs with attached (true) posterior probabilities.

w_1	w_2	$P(w_1w_2\|X)$	$P(w_1\|X)$	$P(w_2\|X)$	E[correct]
a	d	.0	.44	.4	.84
a	e	.24	.44	.34	.78
a	f	.2	.44	.26	.7
b	d	.2	.26	.4	.66
b	e	.05	.26	.34	.6
b	f	.01	.26	.26	.52
c	d	.2	.3	.4	.7
c	e	.05	.3	.34	.64
c	f	.05	.3	.26	.56

For simplicity we assume that all hypotheses consist of exactly two words, w_1 and w_2, shown in the first two columns. The third column shows the assumed joint posterior probabilities $P(w_1w_2|X)$ for these hypotheses. Columns 4 and 5 give the posterior probabilities $P(w_1|X)$ and $P(w_2|X)$ for individual words. These posterior word probabilities follow from the joint posteriors but summing over all hypotheses that share a word in a given position. For example, the posterior $P(w_1 = a|X)$ is obtained by summing

$P(w_1 w_2|X)$ of all hypotheses such that $w_1 = \text{a}$. Column 6 shows the expected number of correct words $E[\text{correct}]$ in each hypothesis, under the assumed posterior distribution. This is simply the sum of $P(w_1|X)$ and $P(w_2|X)$, since

$$\begin{aligned} &E[\text{words correct}(w_1 w_2)|X] \\ &= E[\text{correct}(w_1)|X] + E[\text{correct}(w_2)|X] \\ &= P(w_1|X) + P(w_2|X) \end{aligned}$$

As can be seen, although the first hypothesis ("a d") has posterior 0, it has the highest expected number of words correct, i.e., the minimum expected word error. Thus, we have shown by construction that optimizing overall posterior probability (sentence error) does not always minimize expected word error. Of course the example was constructed such that two words that each have high posterior probability happen to have low (i.e., zero) probability when combined. Note that this is not unrealistic: for example, the language model could all but "prohibit" certain word combinations.

Furthermore, we can expect the discrepancy between word and sentence error to occur more at high error rates. When error rates are low, i.e., when there are at most one of two word errors per sentence, each word error corresponds to a sentence error and vice-versa. Thus, if we had an algorithm to optimize the expected word error directly, we would expect to see its benefits mostly at high error rates.

3. THE ALGORITHM

We now give an algorithm that minimizes the expected word error rate (WER) in the N-best rescoring paradigm [5]. The algorithm has two components: (1) approximating the posterior distribution over hypotheses and (2) computing the expected WER for N-best hypotheses (and picking the one with lowest expected WER).

3.1. Approximating posterior probabilities

An estimate of the posterior probability $P(W|X)$ of a hypothesis W can be derived from Equation (1), with modifications to account for practical limitations:

- The true distributions $P(W)$ and $P(X|M)$ are replaced by their imperfect counterparts, the language model probability $P_{\text{LM}}(W)$ and the acoustic model likelihood $P_{\text{AC}}(X|W)$.
- The dynamic range of the acoustic model, due to unwarranted independence assumptions, needs to be attenuated by an exponent $1/\lambda$ (λ is the language model weight commonly used in speech recognizers, and optimized empirically).
- The normalization term

  $$P(X) = \sum_W P(W)P(X|W)$$

  is replaced by a finite sum over all the hypotheses in the N-best list. This is not strictly necessary for the algorithm since it is invariant to constant factors on the posterior estimates, but it conveniently makes these estimates sum to 1.

Let W_i be the ith hypothesis in the N-best list; the posterior estimate is thus

$$P(W_i|X) \approx \frac{P_{\text{LM}}(W_i)P_{\text{AC}}(W_i|X)^{\frac{1}{\lambda}}}{\sum_{k=1}^{N} P_{\text{LM}}(W_k)P_{\text{AC}}(W_k|X)^{\frac{1}{\lambda}}}$$

This N-best approximation to the posterior has previously been used, e.g., in the computation of posterior word probabilities for keyword spotting [7].

3.2. Computing expected WER

Given a list of N-best hypotheses and their posterior probability estimates, we approximate the expected WER as the weighted average word error relative to all the hypotheses in the N-best list. That is, we consider each of the N hypotheses in turn as the "truth" and weight the word error counts from them with the corresponding posterior probability:

$$E[\text{WE}(W)|X] \approx \sum_{i=1}^{N} P(W_i|X)\text{WE}(W|W_i) \qquad (3)$$

where $\text{WE}(W|W_i)$ denotes the word error of W using W_i as the reference string (computed in the standard way using dynamic programming string alignment).

3.3. Computational Complexity

Rescoring N hypotheses requires N^2 word error computations, which can become quite expensive for N-best lists of 1000 or more hypotheses. We found empirically that the algorithm very rarely picks a hypothesis that is not within the top 10 according to posterior probability. This suggests a shortcut version of the algorithm that only computes expected word error for the top K hypotheses, where $K \ll N$. Note that we still need to consider all N hypotheses to compute the expected word error as in Equation (3), otherwise these estimates become very poor and affect the final result noticeably. The practical version of our algorithm thus has complexity $O(KN)$.

3.4. Other knowledge sources and weight optimization

Often other knowledge sources are added to the standard language model and acoustic scores to improve recognition, such as word transition penalties or scores expressing syntactic or semantic well-formedness (e.g., [4]). Even though these additional scores cannot always be interpreted as probabilities, they can still be combined with exponential weights; the weights are then optimized on a held-out set to minimize WER [5].

This weight optimization should not be confused with the word error minimization discussed here; instead, the two methods complement each other. The additional knowledge sources can be used to yield improved posterior probability estimates, based on which the algorithm described here can be applied. In this scheme, one should first optimize the language model and other knowledge source weights to achieve the best posterior probability estimates (e.g., by minimizing empirical *sentence* error).

	WER	SER
Switchboard		
Standard rescoring	52.7	84.0
WER minimization	52.2	84.4
CallHome Spanish		
Standard rescoring	68.4	80.9
WER minimization	67.8	81.2

Table 1. Word (WER) and Sentence error rates (SER) of standard and word-error-minimizing rescoring methods

So far, we have not implemented combined weight and word error optimization. The experiments reported below used standard language model weights and word transition penalties that had previously been determined as near-optimal in the standard recognition paradigm.

4. EXPERIMENTS

We tested the new rescoring algorithm on 2000-best lists for two test sets taken from spontaneous speech corpora. Test set 1 consisted of 25 conversations from the Switchboard corpus [3]. Test set 2 were 25 conversations from the Spanish CallHome corpus collected by the Linguistic Data Consortium. Due to the properties of spontaneous speech, error rates are relative high on these data, making word error minimization more promising, as discussed earlier.

The results for both standard rescoring and WER minimization are shown in Table 1. On both test sets the WER was reduced by about 0.5% (absolute) using the word error minimization method. A per-sentence analysis of the differences in word error show that the improvement is highly significant in both cases (Sign test $p < 0.0005$). Note that, as expected, the sentence error rate (SER) increased slightly, since we no longer were trying to optimize that criterion.

For comparison, we also applied our algorithm to the 1995 ARPA Hub3 development test set. This data yields much lower word error rates, between 10% and 30%. In this case the algorithm invariably picked the hypothesis with the highest posterior probability estimate, confirming our earlier reasoning that word error minimization was less likely to make a difference at lower error rates.

5. DISCUSSION AND CONCLUSION

We have shown a discrepancy between the classical hypothesis selection method for speech recognizers and the goal of minimizing word error. A new N-best rescoring algorithm has been proposed that corrects this discrepancy by explicitly minimizing expected word error (as opposed to sentence error) according to the posterior distribution of hypotheses. Experiments show that the new algorithm results in small, but consistent (and statistically significant) reductions in word error under high error rate conditions.

In our experiments so far, the improvement in WER is small. However, the experiments confirm that the theoretical possibility of suboptimal WER using the standard rescoring approach is manifest in practice. An important aspect of the WER minimization algorithm is that it can use other, more sophisticated posterior probability estimators, with the potential for larger improvements. Our experiments so far have been based on the commonly used acoustic and language model scores, but we are already experimenting with more complex posterior estimator methods based on neural network models [6].

REFERENCES

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

[2] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. John Wiley, New York, 1973.

[3] J. J. Godfrey, E. C. Holliman, and J. McDaniel. SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing*, vol. I, pp. 517–520, San Francisco, 1992.

[4] R. Moore, D. Appelt, J. Dowding, J. M. Gawron, and D. Moran. Combining linguistic and statistical knowledge sources in natural language processing for ATIS. In *Proceedings ARPA Spoken Language Systems Technology Workshop*, pp. 261–264, Austin, Texas, 1995.

[5] M. Ostendorf, A. Kannan, S. Austin, O. Kimball, R. Schwartz, and J. R. Rohlicek. Integration of diverse recognition methodologies through reevaluation of N-best sentence hypotheses. In *Proceedings DARPA Speech and Natural Language Processing Workshop*, pp. 83–87, Pacific Grove, CA, 1991. Defense Advanced Research Projects Agency, Information Science and Technology Office.

[6] M. Weintraub, F. Beaufays, Z. Rivlin, Y. Konig, and A. Stolcke. Neural-network based measures of confidence for word recognition. In *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing*, vol. II, pp. 887–890, Munich, 1987.

[7] M. Weintraub. LVCSR log-likelihood ratio rescoring for keyword spotting. In *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 297–300, Detroit, 1995.