Review on "Large Vocabulary Decoding and Confidence Estimation Using Word Posterior Probabilities"

Feng Zheng

Institute for Signal and Information Processing
Mississippi State University
Mississippi State, MS 39762 USA
email: zheng@isip.mstate.edu

ABSTRACT

"Large Vocabulary Decoding and Confidence Estimation Using Word Posterior Probabilities" [1] by *Evermann et al.* investigates the estimation of word posterior probabilities based on word lattices and presents applications of these posteriors in a large vocabulary speech recognition system. Two approach to estimate the word posterior probabilities and incorporate into the decoder. A method based on decision trees is suggested to solve the problem of overestimation by word posteriors according to the idea of confidence scores.

In our review, the theory behind the technique looks sound, the technique seems promising to improve the performance in word error rate of speech recognition, and the experiment results nicely verify the improvement achieved by this word posterior technique.

1. INTRODUCTION

In Large Vocabulary speech recognition, the conventional Viterbi decoder typically generates a word lattice which contains a large number of competing word hypotheses and their associated likelihood scores. And then the recognizer is used to rank these competing hypotheses and select the 1-best hypothesis as the final output. However, a Viterbi decoder only considers the state level path with the best scoring and ignores the influence of likelihoods of all other paths, i.e. different time segmentations of the same word sequence, pronunciation variants, or competing word hypotheses. The problem of this approach is that it would completely disregard word sequence constraints by treating all frames independently and would, for example, not be able to detect when the same word is spoken twice in sequence. Thus, Evermann et al in their paper investigate techniques to augment these likelihood with estimates of word level posterior probabilities that allow these alternative paths to be incorporated into decoding.

Two approaches were proposed to estimate the word level posterior probabilities as: 1) *time dependent posterior rescoring,* which takes the sum of all word hypothesis probabilities which represent the occurrence of the same word in more or less the same segment of time; 2) *confusion network,* in which the posteriors of time overlapping links corresponding to the same word are added to yield word posterior estimates. The novel contribution of the paper is that Evermann et al apply the method of confidence scores to help trace the problem of overestimation by word posteriors, and then a method that maps posteriors based on decision trees was applied to compensate for the effects of the lattice size and the resulting overestimation of the posteriors.

Therefore, the review has been organized as the following: section 2 will describe the theory behind the word posterior technique, and procedures to estimate the word posteriors; section 3 will analynize the experimental results achieved by Evermann et al; section 4 will summarize the conclusion that we should recommend this paper.

This paper is well worthy of recommending due to the following nice features: (1) This technique really help improve the performance in word error rate according to the experiments of authors, for example, resulting in >2.5% relative reduction on triphone systems with both hub4 and hub5 data sets. (2) Nice and insightful analysis about the experimental comparison not only confirms readers the correction of the technique but also conveys much beneficial information to readers.

2. METHODOLOGY

2.1. Motivation:

Most speech recognition systems are based on the maximization of the Bayes' rule:

$$P(W|X) = \frac{P(X|W)P(W)}{P(X)}$$

Within this framework a speech recogniser can compute the most likely word string W for a given acoustic signal X using approximations of P(W|A). These approximations are often referred to as *scores*. And Viterbi decoding finds state-level path with maximum a-posteriori probability P(W|X) (sentence level MAP). However, decision based on sentence-level MAP leads to minimum sentence error rate -- sub-optimal for word error rate. Thus, we are thinking that we can use word posteriors p(w|X) instead.

2.2. Estimating Word Level Posterior Probabilities:

The estimation of word level posterior probabilities is based on the scores contained in a word lattice, which associates with the following information:

- Lattice contain information needed to estimate word posteriors;
- represent relevant (most likely) part of search space;
- each link has a word label and acoustic pronunciation and LM score;
- each node corresponds to point in time;
- duplicate links to keep LM history and acoustic context unique.

Step 1: Use forward-backward algorithm to calculate a link posterior probability for each link in the lattice. The link posterior p(l|X) is defined as the sum of the probabilities of all paths q passing through the link l normalized by the probability of the signal p(X):

$$p(l|X) = \frac{\sum_{Ql} p(q, X)}{p(X)}$$

where p(X) is approximated by the sum over all paths through the lattice, that is

$$P(X) = \sum_{Q} p(q, X)$$

The probability of a path p(q, X) is composed from the acoustic likelihood $P_{acc}(X|q)$ and the language model likelihood $P_{lm}(W)$:

$$p(q, X) = p_{acc}(X|q)^{\frac{1}{\gamma}} p_{lm}(W)$$

where $\frac{1}{\gamma}$ is acoustic model scale factor, which was taken as the reciprocal of the standard language model factor.

Step 2: Estimation of word posterior:

Approach 1: Posterior Rescoring:
The word posteriors were added as an additional score to the acoustic and language model scores and the search was performed based on the resulting new decoder objective function:

$$f(W) = p(X, \hat{q}|W)^{\frac{1}{\gamma}} p(W) \rho^{\frac{\|W\|}{\gamma}} \prod_{t=0}^{T} p(w((\hat{q}, t), t|X$$

where γ is the language model weight, ρ is the word insertion penalty and ||W|| is the number of words in the hypothesis. Here, $p(w(\hat{q}, t), t|X)$ is the word posterior probability of the word hypothesised at time t in path $\hat{q}$.

Thus, the word posteriors act as a local consistency measure, if one link hypothesis is supported by many high scoring alternatives then its likelihood is increased.

Approach 2: *Confusion Network Clustering:*
As shown in figure 1, the idea behind the algorithm is to 1) combine overlapping links corresponding to the same word in the same time segment to obtain a word posterior estimate (adding link posteriors); 2) cluster phonetically similar words to form a linear graph, called confusion network, in which a clustering procedure is performed to achieve a total order of the links, i.e. two links are either in the same cluster or

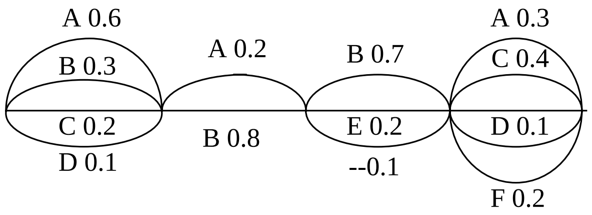

Figure 1: confusion network: pick the hypotheses with the highest posteriors from each cluster, i.e. ABBC

one precedes the other. 3) pick the hypothesis with the highest posterior probability from each cluster, by doing this, they expect to get the word sequence that minimizes the expected word error rate.

3. EXPERIMENT ANALYSIS

The experiments were conducted on the lattices generated by the HTK system used in the 1997 Hub4 (broadcast news) and the 1998 Hub5 (conversational telephone speech) evaluations using triphone and quinphone acoustic models and 4-gram language models.

	Hub4		Hub5	
	WER	SER	WER	SER
baseline	17.4	92.0	42.6	80.2
post	17.0	92.0	42.5	80.5
confnet	16.9	92.3	41.5	80.6

Table 1: Decoding experiments on triphone lattices for HTK Hub4/Hub5 systems using time dependent posteriors (post) and confusion network clustering (confnet)

Table 1 clearly verifies Evermann et al.'s assumption that word based posteriors can be used to improve the accuracy of a Viterbi MAP decoder. The results also exhibit this improvement might come from tradeoff between word and sentence error rates, i.e. a decrease in WER but an increase in SER. The confusion network approach proves to be more robust, as we can see it yields similar improvements on both corpora while the posterior rescoring technique works well on Hub4 corpus but gives no significant improvement on the Hub5 data. Then Evermann et al. explain that hypotheses on the Hub5 data tend to be rather poor, which greatly deteriorates the accuracy of posterior rescoring technique because it only considers links covering the same frame, while the confusion network clustering performs on group time overlapping links into clusters to find the optimal alignment thus compensating for the poor segmentation by the acoustic models.

	triphone	quinphone
baseline	17.4	16.2
post-dec	17.0	16.1
confnet	16.9	16.0

Table 2: WER for Hub4 eval'97 decoding experiments

	triphone	quinphone	Rover
baseline	42.6(0.182)	40.3(0.170)	39.5(0.145)
post-dec	42.5(0.234)	40.0(0.188)	39.1(0.197)
confnet	41.5(0.213)	39.7(0.198)	39.1(0.186)

Table 3: WER & NCE for the Hub5 eval'98 set

Evermann et al dig this up deeply, and they claim that at lower word error rates the correlation between sentence and word error rates should be much stronger than at higher error levels, therefore, it is very interesting to find a greater potential for improvements by using a word posteriors in the decoder even at the lower word error rate level, as there is still significant improvement with Hub4 set.

Table 3, and Table 4 show a comparison of results on Hub4 and Hub5 with triphone and quinphone system. Both techniques are similarly successful in achieving a useful improvement on the triphone lattices, but the effectiveness on the quinphone system is much smaller. Evermann et al apply *Confidence Scores* to analynize this that makes sense. Word posterior probabilities can be used directly as confidence scores of the word hypotheses. If the lattices used are small and contain only a small fraction of the likely word sequences, the posteriors are therefore relatively poor confidence scores. Quinphone system has relatively smaller lattices, thus, the posterior probabilities are less useful as confidence scores than on the triphone system. Further, both the quinphone lattices and the pruning triphone lattices tend to overestimate the posterior probabilities of words. In order to compensate for the effects of the lattice size and the resulting overestimation of the posteriors, Evermann et al. propose a piecewise linear mapping function based on the step function defined by the decision tree and applied to the posterior values. They gave the results to show the improvement of the resulting confidence scores for both techniques of posterior estimation, but failed to describe how to implement this approach clearly. In addition, Evermann et al. tell us that the tree mapping is necessary for both methods, however, it is not fair that they don't tell the mapping function in detail they use, and even the references for this function.

Besides triphone and quinphone systems, Evermann et al applied this technique to their Rover system which combines triphone and quinphone, and obtain the same best number on both posterior rescoring approach and confusion network decoder in word error rate. But they simply contribute the high improvement using posterior rescoring approach to the better confidence estimates, the more compensation. I doubt it, because in the triphone system, posterior rescoring approach also has the better confidence score than confusion network approach, even better than that in Rover system, but only achieves 0.1% a not significant improvement in word error rate.

4. CONCLUSION

This paper present a review on a technique using estimation of word level posterior probabilities based on word lattices to improve the performance in word error rate in large vocabulary speech recognition. Two approaches can be applied to estimate the word posteriors, and then the word posteriors was added as an additional score to the acoustic and language model scores. A mapping function based on the decision tree was applied to the posterior values to compensate for the effects of the lattice size and the resulting overestimation of the posteriors.

The experiments nicely demonstrate the hypotheses of authors: 1) word posterior probabilities do help to improve the accuracy of a Viterbi MAP decoder; 2) the approach to estimate word posteriors using confusion network clustering is more robust than using time dependent posterior, and achieves much more improvement in word error rate on both triphone and quinphone systems. On the other side, Evermann et al dig up the reason behind the experimental results, and suggest decision trees to make the estimation more robust based on confidence scores. In this way, I strongly recommend that this paper should be published.

As we point out some weakness or suggestion in this paper, especially making confidence score related to the performance more clearly, if there is some linear relationship between confidence score and performance. We hope we can see more reports on extensive theoretical analysis of the word posterior procedure, especially how to implement the mapping function based on decision trees.

REFERENCES

[1] G. Evermann, and P. Woodland, "Large Vocabulary Decoding and Confidence Estimation Using Word Posterior Probabilities," *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pp. 1655-1658, Istanbul, Turkey, June 2000.

[2] P. Fetter, F, Daudurand, and P. Regel-Brietzmann, "Word Graph Rescoring Using Confidence Measure," *IEEE transactions on Speech and Audio Processing*, Volume 6 Issue: 6, pp. 539-548, November 1998.

[3] M. Weintraub, "LVCSR Log-Likelihood Ratio Scoring for Keyword Spotting," *IEEE Transactions on Speech and Audio Processing*, Volume: 5 Issue: 2, pp. 297-300, May 1995.

[4] F. Wessel, K, Macherey, and R. Schluter, "Using Word Probabilities as Confidence Measures," IEEE transactions on Speech and Audio Processing, Volume 6 Issue: 6, pp. 539-548, November 1998.

[5] R. Duda, R. Hart, and D. Stork, "Pattern Classification," NewYork: Wiley, 2000.