WORD-BASED CONFIDENCE MEASURES AS A GUIDE FOR STACK SEARCH IN SPEECH RECOGNITION

Chalapathy V. Neti, Salim Roukos, E. Eide *

IBM T. J. Watson Research Center
Yorktown Heights, NY 10598

ABSTRACT

The Maximum *a posteriori* hypothesis is treated as the decoded truth in speech recognition. However, since the word recognition accuracy is not 100%, it is desirable to have an independent confidence measure on how good the maximum *a posteriori* hypothesis is relative to the spoken truth for some applications. Efforts are in progress [1, 2, 3] to develop such confidence measures with the intent of applying it to assesment of confidence of whole utterances [4], rescoring of N-best lists, etc. In this paper, we explore the use of word-based confidence measures to adaptively modify the hypothesis score during search in continuous speech recognition: specifically, based on the confidence of the current sequence of hypothesized words during search, the weight of its prediction is changed as a function of the confidence. Experimental results are described for ATIS and SwitchBoard tasks. About 8% relative reduction in word error is obtained for ATIS.

1. METHOD

1.1. Assigning Confidence

Given a decoded word string $w_1, w_2, ..w_i, ..w_N$, we would like to assign a confidence score, $C(w_i)$ between 0 and 1, that the w_i is correct. Let X be a binary random variable, such that

$$X = \begin{cases} 1 & \text{if } w_i \text{ is correct} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Let y be a vector of scores for the hypothesized word w_i during search. If Y is a random vector which takes on values y, We define $C(w_i)$ as the posterior probability that the word is correct given features y of the word, $p(X = 1/Y = y(w_i))$. One could obtain the posterior probability using Bayes rule as follows:

$$C(\cdot) = p(X = 1/Y = y(\cdot)) = \frac{p(X=1)p(y(\cdot)|X=1)}{p(y(\cdot))} \tag{2}$$

by estimating the conditional densities $p(y(\cdot)|X = 1)$ and $p(y(\cdot)|X = 0)$ and prior probabilities $p(X = 1)$. However, in this paper we use a decision tree approach to estimate the posterior probability $p(X = 1/Y = y(\cdot))$, directly.

To estimate $C(w_i)$, we grow a binary decision tree which asks a series of univariate questions of the form "is $y_i > \theta_i$", about each of the components of the feature vector y. At each node of the tree, each question of the form "is $y_i > \theta_i$", leads to a binary split of the data into two regions, say L and R. Let P be the parent of Y. The decision tree is grown iteratively by finding that question at each node which minimizes the average conditional entropy of X given the split of P to L and R:

$$\begin{aligned} H(X/Y = P, "Y_i > \theta_i?") = \\ p(Y = L/Y = P)H(X/Y = L, Y = P) + \\ p(Y = R/Y = P)H(X/Y = R, Y = P) \end{aligned}$$

where $Y = P, L$, or R implies $y \in P, L$ or R, respectively. Growth of the decision tree is stopped when the decrease in conditional entropy due to the best question is less than a stopping threshold. If L(y) denotes the discrete label of Y corresponding to the leaf of the tree given $Y = y$, the conditional entropy $H_d(X/Y)$ given the decision tree is:

$$\begin{aligned} H_d(X/Y) = \\ -\Sigma_{Y=L(y)} p(L(y)) \Sigma_{x=(0,1)} p(x/L(y)) log(p(x/L(y))) \end{aligned} \tag{3}$$

After the decision tree is grown, given a decoded word w and the set of features y, the confidence that the word is correct is taken as the empirical probability that $X = 1$ at the leaf of y, i.e., $C(w) = p(X = 1/L(y))$.

1.2. Using confidence to guide search

Given acoustic data A corresponding to an utterance, the maximum *a posteriori* word sequence $W = w_1 w_2..w_n$ is that which maximizes the conditional probability P(W/A). By Bayes rule

$$P(W/A) = \frac{P(A/W)P(W)}{P(A)} \tag{4}$$

Since P(A) is independent of the hypothesized word sequence, the maximization is simply done over the numerator of (4). In practice, the two components $P(A/W)$ (acoustic match score) and prior probability of the word sequence $P(W)$ (language model score) are weighted as follows:

$$Score(W/A) = P(A/W)^{\alpha} * P(W)^{\beta} \tag{5}$$

α and β are fixed constants for the search. Given a hypothesized word sequence, $W_i = w_1 w_2..w_i$, and the corresponding

*We thank T. Ward for providing some initial code to build decision trees and L. Bahl and P. Gopalakrishnan for helpful discussions

acoustic segment A_1^i, the score of a hypothesized extension $w_1 w_2 .. w_i w_{i+1}$ is

$$\begin{aligned} & Score(W_{i+1}/A_1^{i+1}) = \\ & Score(W_i/A_1^i) * P(A_i^{i+1}/w_{i+1})^{\alpha} * P(W_{i+1})^{\beta} \end{aligned} \quad (6)$$

We will call W_i the word history for extension W_{i+1}. In general, both α and β could be functions of confidence of W_{i+1}, $C(W_{i+1})$. In this paper, We keep α constant and propose β to be a function of the confidence of the word history as follows:

$$\begin{aligned} & Score(W_{i+1}/A_1^{i+1}) = \\ & Score(W_i/A_1^i) * P(A_i^{i+1}/w_{i+1})^{\alpha} * P(W_{i+1})^{\beta(C(W_i))} \end{aligned} \quad (7)$$

One could assign various functional forms to β as a function of $C(W_i)$ and $C(W_i)$ as a function of the individual $C(w_i)$. First, we choose $C(W_i) = C(w_i)$, implying that the confidence of the word history is only a function of the most recent word in the word history, $C(W_i)$. The motivation for this is based on the following observation. Correctness of w_i has a significant impact on the correctness of w_{i+1}: for example, on a subset of SwitchBoard dev-tst data (80% word accuracy), w_{i+1} is correct about 87% of the time when w_i is correct and only 48% of the time when w_i is incorrect. Second, since $P(W_{i+1})$ is less than 1, $\beta(C(W_i))$ should be less than β_0 when $C(W_i) > C_0(W_i)$ so that predictions by more confident histories have a higher score. Thus, in this paper we use the following functional forms:

$$\begin{aligned} \beta(C(W_i)) &= \beta_0 * \frac{2}{1 + exp(-\gamma * (C_0(W_i) - C(W_i)))} && (8) \\ C(W_i) &= C(w_i) && (9) \\ C_0(W_i) &= C_0(w_i) = p(X = 1) && (10) \end{aligned}$$

The specific functional form for β was chosen so that β changes smoothly around β_0 with the gradient concentrated around the prior probability that the history is right, i.e., $p(X = 1)$. ($p(X = 1)$ is set equal to the relative frequency of correct words in the training data when $\beta = \beta_0$). The constant γ controls the rate of change around β_0. Figure 1 shows an example of the variation of β as a function of $C(W_i)$ for some values of γ.

The notion of language model weight β dependent upon the current word sequence W_{i+1} and the acoustics of the current word A_i^{i+1} has been described in the past [7]. However, they estimate β jointly with the acoustic parameters by defining a new objective function that measures the distance between the truth and alternate hypotheses in an N-Best list. While in this paper, β is a function of the Confidence of the Word history W_i and is estimated independently of the acoustic model parameters by building a decision tree. Also, the functional dependencies (8) are different in this paper.

2. RESULTS

Experiments were performed on the ATIS speech recognition task and SwitchBoard task. ATIS is a medium vocabulary (1700 words) spontaneous speech recognition task to get information about airline travel. A baseline ranks-based left-context only system [5] was trained using 16,223 training utterances from the ARPA ATIS training data with 2153 context dependent states and 28,451 Guassian prototypes. A class-based trigram model was built using the same training utterances. Test set contained 930 utterance from the ARPA ATIS test set containing 7881 words. The baseline system performance for this task was 6.27% word error. For the purpose of this paper, we picked a baseline ATIS system for which the acoustic models were not our best models.

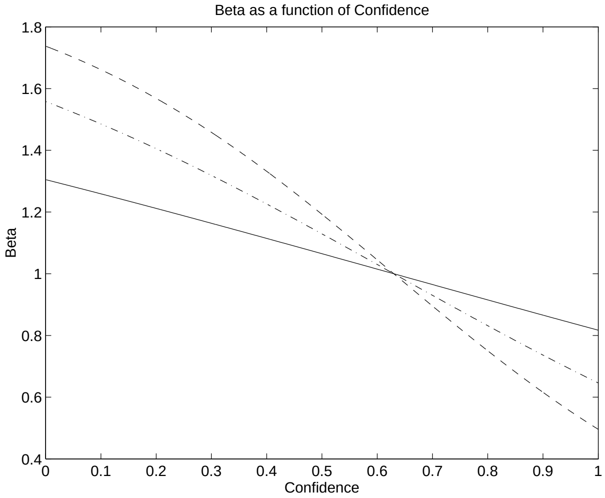

Figure 1. β as a function of $C(W_i)$, $\gamma = 1.0$ (solid); $\gamma = 2.0$ (dash-dot); $\gamma = 3.0$ (dash). $\beta_0 = 1.0$ and $C_0(W_i) = 0.63$.

To grow the decision tree, Word features were generated for about 49000 words by decoding 6000 of the 16223 training utterances using Wall Street Journal (WSJ) acoustics and a language model built using the remaining 10223 utterances to eliminate any bias. The features used to derive the questions were the average likelihood score of a word, LS, i.e.,

$$LS(w_i) = \frac{Score(W_i/A_i) - Score(W_{i-1}/A_{i-1})}{t_i - t_{i-1}} \quad (11)$$

where t_i, t_{i-1} are the most likely ending times of hypothesized word strings W_i and W_{i-1}, respectively. In our search, we do a fast match based on approximate acoustic models to get a list of candidate words (FM list), followed by a language model pruning of the fast match list to get a list of candidates to perform the detailed acoustic match (DM list). Two additional features used were the size of the fast match list (FMC) and the size of the detailed match list (DMC). FMC provides a measure of confusability of words in the match list purely based on an acoustic score, while DMC provides a measure of confusablity of words in the match list based on a combined acoustic and language model score. These features were chosen amongst others based on a normalized mutual information measure (or efficiency [1]) defined as:

$$Efficiency = \frac{H(X) - H_d(X/Y)}{H(X)} * 100 \qquad (12)$$

where $H(X)$ is the entropy at the root of the decision tree and $H_d(X/Y)$ is the conditional entropy of X given the decision tree. An example of a partial decision tree is shown in Figure 2. for the switchboard task.

Feature Set	Efficiency
LS	6.5%
LS, FMC, DMC	18.5%

Table 1. Efficiency as a function of the feature set

Table 1 shows the normalized mutual information for two sets of features measured on an independent test set (7420 words). Note that using the two additional variables FMC and DMC improves the efficiency from 6% to 18%.

Feature set	γ	Error rate
baseline	0.0	7.00%
LS	1.2	6.75%
LS, FMC, DMC	1.2	6.62%

Table 2. Error rate as a function of features: WSJ acoustics

To carry out the recognition experiments, IBM's stack search algorithm [6] was modified according to section 2.2. In the first experiment, ATIS test data was decoded using Wall street Journal acoustic models and a trigram Language model built using 10223 ATIS training utterances. Recognition results for the ATIS test data using Wall Street Journal Acoustic models are shown in Table 2. $\gamma = 0.0$ sets the value of β to the constant β_0 and thus simulates recognition without the decision tree for word confidence. Note that an improvement of about 5.4% is obtained by using a feature set with LS, FMC and DMC. Furthermore, for the same value of $\gamma = 1.2$, adding features FMC and DMC to LS yields a small gain, consistent with the Mutual Information measure. In the second experiment, ATIS test data was decoded using ATIS trained acoustic models. Although the training of the decision tree was done using Wall Street Journal acoustics, We get similar reductions (8%) in the decoding error rate when ATIS trained acoustic models are used to decode the test data as shown in Table 3. Note that the value of γ effects the error rate. The best value of γ for this experiment was 1.3. Increasing it above 1.3 degraded the performance. For the SwitchBoard task, 2800 dev-tst 94 sentences (1698 conversation sides) were used to train the decision trees and 184 held out dev-tst 94 sentences were used as test set. The recogntion experiments were carried out on the same 184 dev-tst 94 sentences containing 1258 words. The baseline system for this task was a left-right context system with 6945 context dependent states and 126,734 Guassian prototypes. A trigram LM built using 1.8 million words was used for the language model. The baseline performance on this test set is 55.01% word error.

γ	Error rate
0.0	6.27%
1.2	5.84%
1.3	5.74%

Table 3. Error rate as a function of γ : ATIS acoustics

Table 4 shows the efficiency for three sets of features. When FMC and DMC are added to LS, the efficiency is smaller (11.5%) compared to the ATIS task. Adding additional features such as Detailed match score (DM) and Language model score (LM) did not help. They degrade the efficiency on the test set. Given the small differnce between the first two sets of features we also looked at the hard classification performance for this task.

Define, % rejects (% of correct words classified as incorrect) as:

$$= (1 - \frac{\Sigma_{(x=1,y)\in T} I[p(X = 1/L(y)) \geq \theta]}{\Sigma_{(x=1,y)\in T} 1}) * 100 \qquad (13)$$

Where T is the test set and I is an indicator function. Define, % False alarms (% of incorrect words classified as correct) as:

$$= (1 - \frac{\Sigma_{(x=0,y)\in T} I[p(X = 0/L(y)) > (1-\theta)]}{\Sigma_{(x=0,y)\in T} 1}) * 100 \qquad (14)$$

Another way to evaluate the goodness of the feature sets in predicting word confidence is by plotting the classification performance (%rejects .vs. %false alarms) at various thresholds. Figure 2. shows %rejects as a function of %false alarms for various values of the hard classification threshold θ. The ideal performance is the case when %rejects can be reduced without increasing %false alarms. Thus, the closer the curve to the left and bottom coordinate axes the better the feature set. Note that for the three sets of features shown, the feature set consisting of LS, FMC and DMC is better. This is consistent with the Normalized mutual information measure. A partial decision tree using LS and FMC features is shown in Figure 3. As we had expected the Word Confidence is higher when the likelihood slope is high (0.82 when $LS > 0.05$) and a lower confidence is predicted when the likelihood slope is small and negative (0.47 when $LS < -0.08$). The decision tree with LS, FMC and DMC features was used for recognition experiments.

Feature Set	Efficiency
LS	8.25%
LS, FMC, DMC	11.5%
LS, FMC, DMC, DM, LM	9.11%

Table 4. Efficiency as a function of the feature set: Switchboard task

Table 5 shows the recognition results for the switchboard task. $\gamma = 1.0$ was the best amongst the values tried. Only two values are shown in the table. The reduction in error rate is small relative to the improvement for the ATIS task.

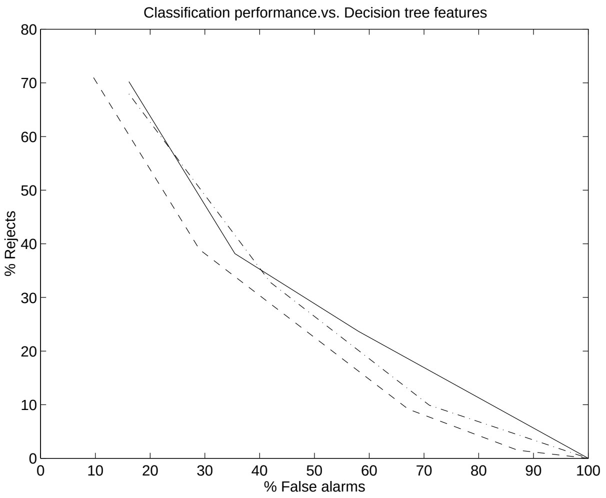

Figure 2. Classification performance for the SwitchBoard task. Solid line - LS only; Dash line - LS, FMC and DMC; dash-dot - LS, FMC, DMC, DM and LM

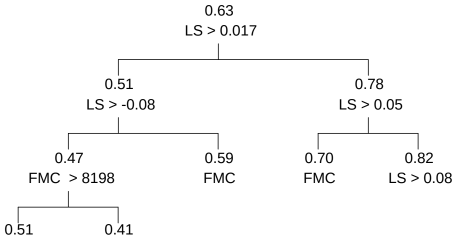

Figure 3. A partial decsion tree with LS and FMC features for the SwitchBoard task. Confidence value and the best question are shown for each node. Right branch is the yes answer to the question.

γ	Error rate
0.0	55.01%
1.0	54.37%
1.3	54.61%

Table 5. Error rate as a function of γ : SwitchBoard task

3. CONCLUSIONS AND DISCUSSION

Noting that the correctness of decoded hypothesis effects the correctness of its prediction we explore a method to assign a confidence between 0 and 1 to decoded hypothesis at a given point in search, and use the confidence to gracefuly modify the weight of its prediction as a function of its confidence. We show an improvement of about 8% relative for the ATIS task and negligible improvement for the SwitchBoard task. A possible reason for the small improvement on the SwitchBoard task is smaller training data size for the decision trees to estimate confidences. Furthermore, in the current scheme the weight of the prediction is changed by changing the weight on the Language Model $P(W_{i+1})$ in equation (7).This suggests that the Language Model is trusted for its correctness. So, another possible reason is that the Language Model is weaker for the SwitchBoard task leading to a negligible improvement. One possible mechanism to offset this weakness is to make β a function of $C(W_{i+1})$, i.e., estimate the confidence based on the predicted word w_{i+1} and include Language Model features as questions in the decision tree. An initial experiment in this direction, led to an error rate of 5.7% compared to 5.79% by using word history only.

REFERENCES

[1] S. Cox and R. Rose. Confidence Measures for the SwitchBoard Database. *Proceedings of ICASSP-96*, Atlanta, 1996.

[2] E. Eide, H. Gish, P. Jeanrenaud, A. Mielke Understanding and Improving speech recognition performance through the use of diagnostic tools. *Proceedings of ICASSP-95*, Detroit, 1995.

[3] F. Beaufays, Y. Konig, Z. Rivlin, A. Stolcke, M. Weintraub. Neural-Network based Measures of Confidence. *Proceedings of the Speech Recognition workshop*, Arden House, Hariman, NY, February, 1996.

[4] E. Lleida and R. Rose. Efficient decoding and training procedures for utterance verification in continuous speech recognition. *Proceedings of ICASSP-96*, Atlanta, 1996.

[5] L. Bahl, P.V. de Souza, P. S. Gopalakrishnan, D. Nahamoo, M.A. Picheny. Robust methods for using context-dependent features and models in a continous speech recognizer. *Proceedings of ICASSP-94*, Adelaide, Australia, 1994.

[6] P. S. Gopalakrishnan, L. R. Bahl, R.L. Mercer. A tree search strategy for large-vocabulary continuous speech recognition. *Proceedings of ICASSP-95*, Detroit, 1995.

[7] X. Huang, F. Alleva, M. Hwang and R. Rosenfeld. An overview of the SPHINX-II speech recognition system. *Human Language Technology Proceedings*, Princeton, 1993.