IMPROVED ESTIMATION, EVALUATION AND APPLICATIONS OF CONFIDENCE MEASURES FOR SPEECH RECOGNITION

Man-hung Siu, Herbert Gish, Fred Richardson

BBN Systems and Technologies
70 Fawcett St. Cambridge, MA 02138

ABSTRACT

This paper describes our approach to the estimation of confidence in the words generated by a speech recognition system. We describe the models and the features employed for confidence estimation. In addition we discuss the characteristics of an information-theoretic metric for assessing the performance of the confidence measure. We provide a simple application of confidence measures in which we rank the performance of speakers.

1. INTRODUCTION

As speech recognition technologies become increasingly useful in real-world applications, the information required of a speech recognizer will need to go beyond just providing the best transcription. In addition to the transcription, many applications need the ability to detect the breakdown of a recognizer or reject erroneous recognition answers. In our previous work, [1], we presented our approach of generating recognition confidence measures for words in the best recognition hypothesis. This confidence measure was derived solely from information generated internally to the recognizer.

In this paper we put confidence measures in a more general framework, including an information-theoretic basis for evaluating confidence measure performance. The metric for evaluating confidence measures that we discuss has been proposed by us as well as other members of the speech community [2]. In this paper, we discuss its properties and inadequacies. We also discuss the selection and incorporation of different knowledge sources to estimate word confidence measures. Finally, we consider the application of confidence measures to the speech recognition process itself, as a means of estimating recognition errors or incorporating speech knowledge that exists external to the HMM into the decoding process.

This paper is organized as follows. First, we describe an information-theoretic measure for assessing the performance of confidence measures. Second, we describe the various candidate knowledge sources and our selection algorithm. Third, we compare two models for combining these knowledge sources, the logit model and extension of it to the generalized additive model. Then, we present some experimental results on speech recognition for estimating confidence measures on the Switchboard corpus [3], employing our Byblos speech recognition system which achieves state-of-the-art recognition performance. We then proceed to discuss some applications of confidence measures.

2. EVALUATION OF CONFIDENCE MEASURES

We consider a confidence measure for a recognized word as the probability that a recognized word is correct. We associate with each recognized word, w_i, an indicator variable, c_i such that $c_i = 1$ if w_i is correctly recognized and 0 otherwise. The confidence measure for w_i, $q(c_i = 1|X)$ is an estimate of $p(c_i = 1|X)$ where X is all available information during test. For a recognizer operating with an average accuracy of p_0, the *a priori* estimate of $p(c_i = 1|X)$ is p_0. This becomes a baseline for us to compare our confidence measure performance.

One way to evaluate the goodness of $q(c_i|X)$, proposed by us and others, is the normalized mutual information given by,

$$M(q|C,X) = \frac{H(C) - H(C|X)}{H(C)} \qquad (1)$$

where $H(C)$ is the per-word entropy of the c_i sequence and is given by

$$H(C) = -p_0 \log p_0 - (1-p_0)\log(1-p_0), \qquad (2)$$

where p_0 is the accuracy of the recognizer. This assumes that the c_i's are independent Bernoulli events. $H(C|X)$ is the entropy of the c_i sequence with knowledge of all the information available and is given by

$$\begin{aligned} H(C|X) &= \frac{-1}{n}\sum_{i=1}^{n}[c_i \log q(c_i = 1|X) + \\ &\qquad (1-c_i)\log(1 - q(c_i = 1|X))]. \end{aligned} \qquad (3)$$

In essence, the $M(q|C,X)$ measures the percentage change in uncertainty of c_i after we use the confidence measure $q(c_i|X)$ as compared to the original uncertainty $H(C)$.

By measuring relative improvement rather than absolute, we bound the range of $M(q|C,X)$ to being between zero and one. If one has perfect knowledge of which words are correct, then one will have $M(q|C,X) = 1$. If the measurements provide no additional information as to which words are correct

then one will have $M(q|C,X) = 0$. We will refer to $M(q|C,X)$ as the confidence evaluation metric (CEM).

Sensitivity to operating point The CEM given as the measure of the ratio of change in entropy of the word class (correct and incorrect) to the test-set class entropy is an attempt to normalize the evaluation against recognition operating points across different systems. However, it is still sensitive to a change in recognition accuracy and cannot fairly compare against different recognizers with different operating points. Suppose we have a system that we can change its accuracy p_0 without changing $H(C|X)$. How would $M(q|C,X)$ change? The derivative $\frac{d\ M(q|C,X)}{d\ p_0}$ is given by

$$\frac{d\ M(q|C,X)}{d\ p_0} = \frac{H(C|X)}{(p_0 \log p_0 + (1-p_0)\log(1-p_0))^2} \log \frac{1-p_0}{p_0}. \quad (4)$$

Equation 4 shows that the sign of the derivative is determined by $log(1 - p_0/p_0)$. When $p_0 < 0.5$, increasing p_0 will increase M. However, when $p_0 > 0.5$, increasing p_0 will decrease M. This implies that if we can maintain $H(C|X)$ and change p_0, we can in effect change M.

In addition to comparing confidence scores between different systems, sensitivity to the operation point can be an issue for a system in which all of the information that is employed in the decoding process is not employed in estimating the word confidence scores. Thus the operating point of the system is not the same operating point at which the confidence score was computed. For example, this situation can occur in n-best rescoring where we improve our recognition by means of re-ordering n-best based on extra knowledge sources. If those knowledge sources are not effectively incorporated in the estimation of confidence measure, the confidence scoring process is operating at a different operating point than the system and also in this situation $H(C|X)$ may change very little. In Section 5. we will provide experimental results for a rescoring experiment.

In order to better understand the sensitivity of CEM to the operating point we will demonstrate that one can actually trade recognition accuracy for increases in CEM score. This trade-off can serve as a basis for further normalization of the CEM. In Section 5., we will consider several approaches to decreasing p_0 and increasing M by the technique of changing the recognition transcriptions and the assigned confidences to the changed words.

3. KNOWLEDGE SOURCES

In our previous work, we generated our confidence measures based on estimation of the posterior probabilities obtained from the likelihood measurements from the decoder. However, there are other factors that can affect the performance of recognition and may not be effectively captured by the HMM model. For our present model, in addition to using knowledge from HMM model such as acoustic scores and language modeling scores, we also utilize additional features that may affect performance. We group the features into four categories: 1) scores from the HMM such as word posterior probability estimates directly from the decoder [1], and those estimated from n-best lists, 2) language modeling information such as bigram and trigram probabilities, language model likelihood of the sentence and amount of training, 3) acoustic information such as estimated speaking rate, amount of training for words, triphones involved and their durations, signal-to-noise ratio, 4) context features such as acoustic likelihood of preceding and succeeding words.

To select the right set of features optimally requires evaluating a very large number of possibilities. Instead, we use a greedy incremental selection algorithm. This algorithm tests one feature at a time on a held-out set. Once a feature is added, there is no mechanism to remove it. It is based on techniques for feature selection in regression analysis. The algorithm works as follows,

1 initialize the selected set as empty and the candidate set to contain all features

2 for each feature in the candidate set, evaluate the goodness measure on the cross-validation test when added to the selected set

3 if the best feature improves the goodness measure, put the feature in the selection set and go back to 2

The order in which the features are selected indicates the importance of the feature as measured by the amount of extra information each feature contributes. If two features are correlated, one may not be selected if the other is already selected.

Feedback Features To better capture the context information, we use the confidence score of preceding words as features. In order to do so, we perform a two-pass training and test. During training, an initial model is built based on the selected features. The initial model is then used to generate confidence measures of all training tokens. A final model is built using the confidence measures of the preceding and succeeding words as features. During test, the initial model is used to generate an initial confidence measures of all test tokens. Then, the final model uses the initial confidence measure of neighboring words and generates a final confidence measure.

4. MODELS

We consider two models for combining features and computing confidence measures: the logit model which is a member of the family of generalized linear models (GLM) and a variant of the logit model in which features are transformed before combination. This latter variant of the logit model is a member of the family of generalized additive models (GAM). These models are particularly useful for modeling probabilities.

The Logit Model The logit model is a model for the log odds of an event which is a linear function of the features. In the case of confidence measures we have for the logit model:

$$\log \frac{q(c_i = 1|X)}{1 - q(c_i = 1|X)} = \sum_i \beta_i x_i \qquad (5)$$

where the x_i are the components of X.

Generalized additive model- The Logit Variation The GAM version of the logit model [3] extends it by allowing non-linear transformations for selected features. That is,

$$\log \frac{q(c_i = 1|X)}{1 - q(c_i = 1|X)} = \sum_i g_i(x_i) \qquad (6)$$

The g_i's provides a very powerful method for transforming features that are not truly linear to the response. Furthermore, some of the features are nonlinearly related to each other and the use of a nonlinear transformation on the feature reduces the number of features needed by making the related ones redundant. The disadvantage of this model is the number of parameters used and robustness of the model. Depending on the way g_i's are estimated, this model can require many more parameters than the conventional logit model. It also adds the complexity of deciding whether one feature should be transformed or not.

5. EXPERIMENTS

We performed our confidence measure experiments on the Switchboard Conversational speech corpus [5]. We trained our models using 25 minutes of recognition output and selected features on an held-out set of 25 minutes of speech. The models were estimated using S-Plus [4] software.

Sensitivity of the evaluation measure The last part of our recognition system re-orders n-best hypotheses based on other knowledge sources such as a CNN grammar, a grammar derived from Cable Network News transcripts [6]. However, this information source is not employed in our confidence estimates. The rescoring improves our recognition performance by about 2 percent but our CEM is worse. In Table 1, we tabulate the class entropy $H(C)$ and class conditional entropy $H(C|X)$ before and after rescoring on our test data using the best feature set. We can see that the change in $H(C)$ degrades CEM. As noted previously, this drop is due to the operating point mismatch between the final decoding and the confidence estimates.

We also ran several experiments in which we traded a decrease in recognition performance for improvements in the CEM. In one experiment we changed a tail portion of the recognized words, ranked by its confidence score, to the unknown-word symbol which decreased p_0. At the same time, we changed the confidence measure of those words to zero which decreased $H(C|X)$. This is because the confidence score of these words is around 0.1 and most of them are recognized incorrectly. The solid (top) line in Figure 1 shows the change in $M(q|C,X)$ with respect to the recognizer accuracy in our experiment. As we changed more words, the recognition accuracy decreased while the performance of confidence as measured by CEM increased dramatically.

The dotted (middle) curve shows the result of adding new unknown-words to the recognition which in effect only changes the recognition accuracy. The per word $H(C|X)$ decreases because of the increase in number of tokens, but not as dramatically as in the previous experiment. If we maintain $H(C|X)$ the same, the change in CEM is more moderate as shown in the dashed (bottom) line where $H(C|X) = 0.503$.

The above described trade-offs between recognition performance and confidence performance show that the influence of the operating point can be quite significant. Also, although we have not attempted to do so here, these trade-offs can serve as the basis for additional normalizations of the CEM.

	before re-order	after re-order
p_0	0.694	0.710
$H(C)$	0.616	0.602
$H(C\|X)$	0.498	0.498
$M(q\|C,X)$	0.192	0.173

Table 1. The effect of re-ordering n-best

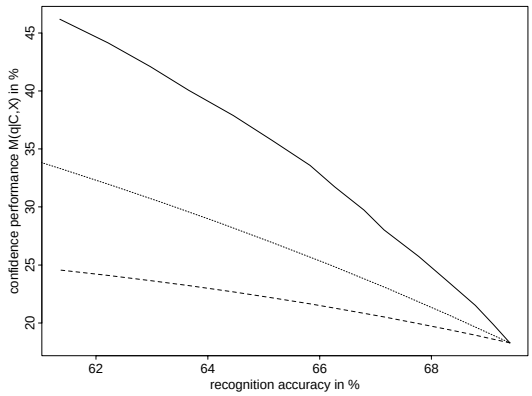

Figure 1. The effect of changing recognition accuracy on the CEM

Feature selection and combination We use the logit model, which is a special case of the GLM to estimate confidence scores. Table 2 shows the list of top features selected by the GLM model on our Viterbi recognition output (without rescoring) on our feature selection test-set. Observe that the top 5 to 6 features accounts for 90% of the performance. Furthermore, many of the less significant features are actually variants of other selected features.

The feature selection in GAM is slightly different from logit. Since not all features require a nonlinear transformation, we modified our feature selection scheme by testing whether the use of the nonlinear function improves the power of the feature sig-

Features selected	$M(q\|C,X)$
n-best based score	0.141
recognizer posterior score	0.149
word trigram	0.158
sentence LM score	0.161
number of words in sentence	0.166
word duration	0.169
all selected features	0.173

Table 2. Effects of different features added to logit model

nificantly. In our work, a spline was estimated as the non-linear transformation of the feature. When we use the GAM model, fewer features were selected while each feature contributed more as shown in Table 3. From the ranked list, similar to the logit list, the most important feature is the word-score from n-best rescoring and the posterior probability from the recognizer.

Features selected	$M(q\|C,X)$	transformed
n-best score	0.151	yes
posterior score	0.171	yes
Word grammar score	0.181	yes
word duration	0.186	yes
all selected features	0.192	
all selected + feedback	0.193	

Table 3. Effects of different features added to GAM

6. APPLICATIONS

Confidence measures have many applications such as creating partial transcriptions of the highest accuracy words, improved back-off strategies for language modeling, speaker adaptation and topic discrimination [7], among others. One application is the use of confidence measures in predicting the recognition performance of a speaker. This is particularly useful to detect sudden breakdown of recognition condition where the recognition performance degrades under adverse conditions. In Figure 2, we show the scatter plot of actual recognition word error rate against the predicted speaker word error on the DARPA Hub-5E 97 evaluation on both Switchboard and Callhome English speakers. The predicted speaker error is computed as $1 - a$, where a is the average confidence measure of all the hypothesized words. A correlation coefficient of 0.87 is obtained showing that the estimated error is linearly related with the actual error and by means of the confidence measure, the performance of each speaker can be inferred.

Another application is to employ confidence measures in the decoding process which allows information not normally employed in the HMM decoding process to be incorporated. We performed this confidence based decoding in the n-best rescoring paradigm where confidence measures on all words in our n-best hypotheses are first generated. Based on the the word confidences of each hypothesis, we rescore the top 20 n-best. Our initial experiments

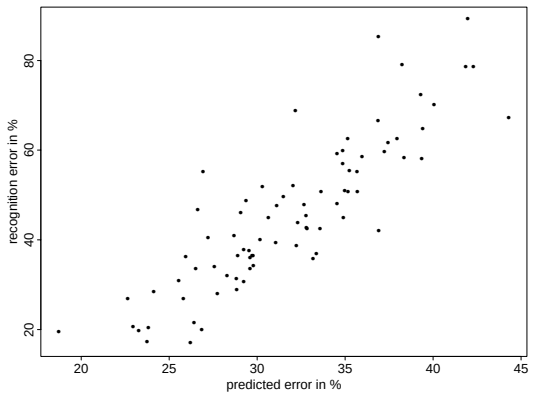

Figure 2. Scatter plot of predicted errors and word recognition errors

give a gain of 0.4% over the optimized Viterbi decoding.

7. CONCLUSIONS

We have demonstrated how we can use logit models and their generalized additive versions for estimation of word confidence as well as describing the features used and a method for their selection. We showed that using the generalized additive versions of logit model performs better at the cost of increasing model complexity. We discussed a metric for measuring the performance of confidence measures and illustrated that one needs to take into account the recognition system operating point to best appreciate what it is doing. We illustrated an application of confidence measures for predicting the performance of speakers and found a high correlation between estimated and actual performance. Finally, we showed that word confidence measures can improve recognition performance. We plan to investigate various ways of incorporating word confidence scores into the recognition process.

REFERENCES

1 P. Jeanrenaud, M. Siu, and H. Gish. "Large Vocabulary Word Scoring as a Basis for Transcription Generation" *Proc. ESCA Eurospeech*, 1995

2 DARPA Hub-5E workshop 1996.

3 T. Hastie, and R. Tibshirani. "Generalized Additiive Models." Chapman and Hall, London, 1990.

4 "S-plus: Guide to statistical and mathematical analysis", Mathsoft, 1995.

5 J. Godfrey, E. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone Speech Corpus for Research Development" *Proc. IEEE ICASSP*, 1992, pp. I-517-520

6 R. Iyer, M. Ostendorf and H. Gish, "Using Out-of-Domain Data to Improve In-Domain Language Models", Tech. Rep. ECE-97-001, Boston University, Jan. 1997. Available from `ftp://raven.bu.edu/pub/reports`.

7 J. McDonough, H. Gish, et al., U.S. Patent No. 5625748, "Topic Discriminator using Posterior Probabilities and Confidence Scores," Issued April 29, 1997.