EXPERIMENTS IN CONFIDENCE SCORING USING SPANISH CALLHOME DATA

Jon G. Vaver

Department of Defense
9800 Savage Road
Ft. Meade, MD 20755, USA
jgvaver@zombie.ncsc.mil

ABSTRACT

We present results relevant to tasks involved in the confidence scoring of output from a continuous speech recognition system, including the search for predictor variables and model selection. We introduce the DET curve characteristic (DCC) score, which we use along with the normalized cross entropy (NCE) score, to perform the model and predictor variable evaluation. We also show results from experiments that suggest how the NCE and DCC scores vary with recognizer performance.

1. INTRODUCTION

The output from a less-than-perfect automatic speech recognizer contains accurate as well as erroneous information about the true transcript. In almost any application, the presence of erroneous information in the recognizer output will diminish the value and utility of the accurate information. This degradation is more prominent in situations in which recognizers operate at a high error rate, as is the case when Spanish CallHome data is used as input. Therefore it is highly desirable that speech recognizers include a well developed and carefully tested confidence scoring procedure for evaluating and sorting their recognition output.

We identify two central issues associated with the formulation of a confidence scoring procedure. The first issue is finding a set of predictor variables that contain complementary information relevant to the probability that a recognized word is correct. The second issue is how to construct a function, F, which combines the information contained in the predictor variables and generates the confidence probability estimate.

In this paper we present results pertaining to both of these central issues related to confidence scoring. Model evaluation is accomplished by employing the widely used normalized cross entropy (NCE) score, as well as a score that we introduce and identify as the DET curve characteristic (DCC) score. We use these evaluation metrics to assess the development of particular confidence models on a fixed set of recognition output. It is also desirable to compare distinct confidence models that are applied to output from distinct recognition systems that have different accuracy rates. Therefore, we also show results from two simple experiments that suggest how the NCE and DCC scores vary with recognizer performance.

2. CONFIDENCE MODELS

Recently, a number of research groups have incorporated confidence scoring into their speech recognizers. A variety of methods for constructing F have been used, including generalized linear models [6] [15], generalized additive models [15], tree based regression [3], and neural networks [16]. While each of these models has been used successfully in confidence modeling, it is not clear from these applications how different confidence models perform relative to one another. We begin to address this issue by showing results for several different confidence models, including models based on logistic regression via maximum likelihood, tree based regression, and discriminant analysis. An S-Plus implementation of these methods was used in each case.

Both the GLM and GAM regression fits are found by maximizing the log-likelihood as described by McCullagh and Nelder [12], and Hastie and Tibshirani [7], respectively. The regression trees that we used were generated using binary recursive partitioning [2]. We used one set of confidence training data to generate a regression tree and a second independent set of confidence training data to generate a set of pruned trees that each contain a different number of terminal nodes. The pruned tree chosen as the final regression tree was the one that maximized the NCE score over the second training data set.

Linear discriminant analysis (LDA) is a simple prototype classifier that distinguishes two or more classes of data, which in our case is correctly and incorrectly recognized words. Bayes formula can be used to estimate the probability that an observation (a recognized word) belongs to a given class (is correct or incorrect) [see, for example, [10]]. LDA classification uses linear decision boundaries, which do not always adequately separate classes. An alternative to LDA is flexible discriminant analysis (FDA) [8], which utilizes nonparametric regression techniques to enable the use of nonlinear decision boundaries.

3. PREDICTOR VARIABLE SOURCES

3.1 Recognizer Output

The confidence score of a recognized word is determined using information from the recognition process. As is the case for existing confidence models, the major source of predictor variables used in our confidence models was the output of the automatic speech recognizer. We used a Dragon Systems based recognizer [1] that performs at a 69.3% word error rate when applied to Spanish CallHome data.

We used one occurrence independent predictor variable (i.e., predictor that does not vary with each occurrence of a recognized word), the number of phonemes in the recognized word. The only source of recognizer output used in this predictor is the recognized word itself.

We used several occurrence dependent predictor variables stemming from recognizer output, and we divide these into word based and utterance based predictors. An example of an occurrence dependent, word based, predictor variable is the word language model score. This predictor also has an utterance based counter part; the utterance language model score. The utterance based language model score is the average word language model score for all of the words in the utterance.

3.2 Recognition Training Data

Recognition training data have not been widely discussed as sources of predictor variables. Our confidence models included two predictors that fall into this category. The first predictor, the overall word frequency, is the number of times a recognized word occurs in the acoustic recognition training data. Word frequency in recognition training was used by Eide et al., [4] although the primary focus of the study was the diagnostic analysis of recognition performance rather than confidence scoring. The second predictor, the conversation word frequency, is the number of different speakers who uttered the word during recognition training.

In addition to these two predictors, we also experimented with a normalized phoneme duration (NPD) predictor variable that uti-

lizes information from the recognition training data as well as the recognition output. The word duration is a weak predictor for confidence scoring; the NPD predictor was an attempt to refine word duration at the phoneme level in order to create a stronger confidence predictor. This first step in computing the NPD predictor is to use the recognition training data to create frequency distributions for the duration of each phoneme. The NPD predictor is then defined as

$$NPD(w) = \frac{1}{n_\phi} \sum_{i=1}^{n_\phi} \frac{|\phi_i - \bar{\phi}_i|}{\bar{\phi}_i} \quad (1)$$

where n_ϕ is the number of phonemes in word w, ϕ_i is the duration of the i^{th} phoneme in w, and $\bar{\phi}_i$ is the mode of the frequency distribution for the phoneme ϕ_i.

3.3 Confidence Training Data

Our final source of predictor variables was the confidence training data. We used one predictor of this type, the previous word performance (PWP) predictor, which is the probability that a word was recognized correctly in the confidence training data. This predictor could come from any previously scored recognition run, but we assume that all of the previously scored recognition data will be used for recognizer or confidence training.

Not all of the recognized words occur with sufficient frequency in the confidence training data to give good statistics for the PWP predictor. Therefore, we generated a preliminary confidence model based on all of the predictor variables except the PWP predictor for use in the following procedure. If the recognized word, w, occurs with sufficient frequency in the confidence training data then the value assigned to the PWP predictor is the probability that the word was recognized correctly in the confidence training data. If the recognized word does not occur frequently in the confidence training data then the value assigned to the PWP predictor is the confidence probability generated by the preliminary confidence model. After the PWP predictor is generated, another confidence model that includes all previously used predictors plus the PWP predictor is constructed. This preliminary-final confidence model set up that we used is similar to the one employed by Siu, Gish, and Richardson [15] (hereafter referred to as SGR).

The transition between input sources for the PWP predictor was done smoothly. The PWP predictor is defined by the following equation,

$$PWP(w) = \beta(f_{CT})\ p_p(w) + [1-\beta(f_{CT})]\ p_{CT}(w) \quad (2)$$

where f_{CT} is the frequency of w in the confidence training data, $p_p(w)$ is the confidence probability of w generated by the preliminary model, $p_{CT}(w)$ is the probability that w was recognized correctly in the confidence training data, and

$$\beta(f_{CT}) = e^{-\alpha f_{CT}} \quad (3)$$

determines how the transition between the input sources for PWP construction occurs. The free parameter α was determined experimentally using confidence training data.

4. EVALUATION METRICS

One measure we use to evaluate our confidence models is the normalized cross entropy (NCE) score, defined as

$$NCE(\boldsymbol{p}, \boldsymbol{c}) = \frac{H(\boldsymbol{p}_o, \boldsymbol{c}) - H(\boldsymbol{p}, \boldsymbol{c})}{H(\boldsymbol{p}_o, \boldsymbol{c})} \quad (4)$$

where p_o is the percentage of correctly recognized words and $H(\boldsymbol{p}, \boldsymbol{c})$ is the entropy for recognized words that are assigned probability estimates contained in $\boldsymbol{p}$ and have 0/1 truth $\boldsymbol{c}$. The NCE score is 0 if all words are assigned a confidence score of p_o and 1 if all words are assigned a probability of c_i.

In addition to the NCE score, we also compute the DET curve characteristic (DCC) score. The confidence model assigns a probability of correctness to each recognized word, and these assignments provide a means of ranking recognized words. In order to extract the recognized words that are most likely to contain accurate information about the true transcript, we start at the top of the ranked list of recognized words and work our way down. The 50% DCC score is defined as the percentage of incorrectly recognized words removed from the list before 50% of the correct words are removed.

We identify the DCC score as such because it characterizes the information contained in the DET curve [11] with a single number. The DCC score is useful because it is easily interpreted, and it gives a clear indication of how well a confidence model sorts the recognizer output. The 50% DCC score could be replaced by, for example, a 75% (25%) DCC score, for recognizers that operate at a higher (lower) word error rate.

One characteristic that distinguishes the DCC score from the NCE score is that it remains constant if the confidence probabilities are remapped in such a way that the ranked list of words remains constant, (e.g., all confidence probabilities are halved). The DCC score is then, in a limited sense, independent of the specific confidence probabilities assigned to the recognized words, whereas the NCE score is not. We do not advocate the use of the DCC score in place of the NCE score, unless the specific application warrants such a switch. We do advocate the use of both scores for generic applications and the evaluation of confidence models.

5. RESULTS

The Dragon Systems based recognition system referred to in section 3.1 was applied to the 1995 and 1996 Spanish CallHome evaluation data sets in which p_o is 41.4% and 41.1%, respectively. The entire 1995 evaluation set was used for training the confidence models, and the entire 1996 evaluation set was used for testing the confidence models, with the exception of tree based regression. The tree models were constructed using the entire 1995 data set, trimmed using the first half of the 1996 data set, and tested using the second half of the 1996 evaluation data. The full confidence model included a total of 17 predictor variables. Several of these predictor variables are particularly weak and could have been removed from the model without significant degradation in prediction performance. We did not focus on the task of finding the optimal set of predictor variables because the conclusions of this paper do not depend on doing so.

5.1 Individual Predictor Performances

We begin by showing some results that indicate the strength of individual predictors. The results in Table 1 were generated by training and testing the confidence model (GAM) using a single predictor variable.

Table 1: Individual Performance of Predictor

Predictor Variable	NCE
Nbest Occurrences (weighted)	10.2
Normalized Acoustic Score	5.8
Language Model Score	3.8
Word Freq. in Rec.Training (conversation side)	2.2
Word Freq. in Rec.Training (overall)	1.8
Normalized Phoneme Duration Score	0.4

The predictor variables shown in the top three rows of table 1 are the best individually scoring predictor variables. The best individually scoring predictor variable is the weighted N-best predictor with a 10.2% NCE score. The predictors shown in the bottom three rows of the table are three of the four new predictors described above. Both of the word frequency in recognition training predictors contain a modest amount of information regarding confidence probabilities. The last entry in Table 1 indicates that the normalized phoneme duration score defined in (1) is a very weak predictor variable.

5.2 Preliminary Model Results

Table 2 shows NCE and DCC score results from the preliminary confidence model, which includes all predictor variables except the PWP predictor. The best performing method according to the NCE score is flexible discriminant analysis with (depth 2) interactions, followed by the generalized additive model and flexible discriminant analysis without interactions. The poorest performing method is the tree based regression, which is clearly overtrained due to a possible S-plus bug that does not allow the tree to be pruned in an optimal manner. However, even the performance of the tree regression on the training data does not match the performance of the FDA and GAM methods on the evaluation data.

Table 2: Preliminary Model Results

Model	NCE (%)		50% DCC	
	Train	Test	Train	Test
FDA[1]	19.5	18.8	13.5	13.2
GAM	18.4	18.4	14.3	13.0
FDA	18.0	18.4	14.4	13.1
GLM[1]	18.7	17.3	14.1	14.4
GLM	14.5	15.0	16.8	16.5
LDA	14.5	15.0	17.0	16.9
Tree	17.8	13.1	14.9	18.1

1. Includes interactions between predictor variables.

5.3 Full Model Results

The results for the confidence models that use the entire set of predictor variables are shown in Table 3. The best performing full model is the GAM, followed closely by FDA with interactions. Two additional columns indicate the gain achieved by including the PWP predictor. Adding the PWP predictor results in a 0 to 1% improvement in the NCE and DCC scores.

Table 3: Full Model Results

Model	NCE (%)			DCC (%)		
	Train	Test	Gain	Train	Test	Gain
GAM optimal α	(19.7)	(19.2)	(0.8)	(13.6)	(12.3)	(0.7)
GAM	21.3	18.9	0.5	12.3	12.9	0.1
FDA[1]	19.5	18.8	0.0	13.5	13.2	0.0
FDA	20.5	18.6	0.2	12.7	12.6	0.5
GLM[1]	21.6	17.3	0.0	14.1	14.4	0.0
LDA	17.4	16.0	1.0	15.3	16.5	0.4

1. Includes interactions between predictor variables.

Ideally, the parameter α should be determined using a substantial amount of independent confidence training data. We trained the set of confidence models on 75% of the 1995 data set and used the remaining 25% of this data set to determine an estimate for α. Because this 25% may not be sufficient to provide a good estimate for α, we include the first line in Table 3 to indicate how the GAM model performs with the optimal choice for α. With enough training data, the preliminary to full model gain for the GAM would likely be somewhere between 0.5 and 0.8%

6. SENSITIVITY OF EVALUATION METRICS

Figure 1 shows the PDF's for the estimated confidence probabilities (generated by the preliminary GAM model) after separating the correctly recognized words from the incorrectly recognized words. In experiment 1, first performed by SGR, the number of recognized words is held fixed and a set of words with the lowest confidence probabilities are assigned zero confidence. At the same time, the correct words from this set are marked as incorrect. The net result is a set of incorrectly recognized words that are assigned a confidence probability of zero.

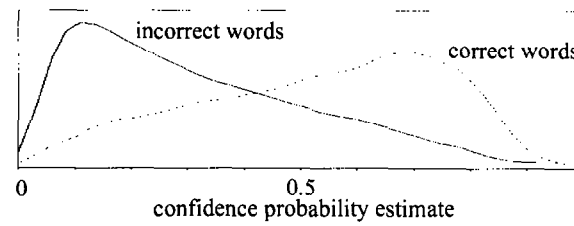

Figure 1. PDF's for the confidence probability estimates of correctly (dotted) and incorrectly (solid) recognized words.

Not surprisingly, figure 2 shows that the NCE score increases dramatically as the size of the converted word set increases. The increase occurs because the experiment introduces perfect knowledge of word correctness where there was none previously. If this experiment is run in reverse (i.e., incorrect words are marked correct), the NCE score also increases dramatically. Taken to the extreme, all words could be marked incorrect and assigned zero confidence to yield a perfect NCE score of 100%. A recognizer with an inherently lower performance level does not have this advantage of perfect word knowledge and, therefore, this experiment does not give an accurate indication of how the NCE score varies with recognizer performance.

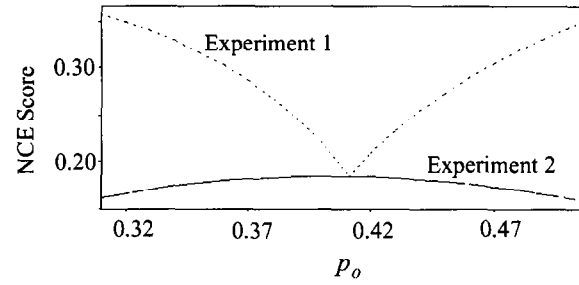

Figure 2. Dependence of the NCE score on recognizer performance for experiments 1 and 2 (described in the text).

In experiment 2, words are randomly removed from the PDF of correctly recognized words, marked as incorrect, and assigned a confidence probability chosen from the PDF of incorrectly recognized words. In this experiment, the number of recognized words remains constant, there is no introduction of perfect knowledge about word correctness, and the shapes of the PDF's in figure 1 remain unchanged. This experiment more accurately resembles the situation in which the same confidence model is applied to recognizers with different performance levels. The approximation is that the confidence model determines the shape of the PDF's, and the accuracy of the recognizer determines what percentage of recognized words end up in each PDF. Figure 2 shows that the resulting NCE dependence is fundamentally different in experiment 2; the NCE score decreases modestly as p_o varies

from its original value of 41%, in contrast to the dramatic increase in experiment 1. Figure 3 shows the result of performing the same experiments using the DCC score. The lack of variation in the DCC score in experiment 2 provides additional support for including the DCC score in confidence model evaluation.

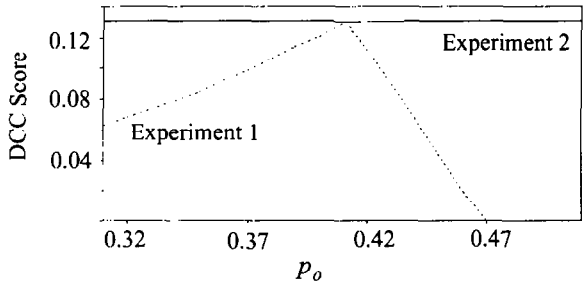

Figure 3. Dependence of the NCE score on recognizer performance for experiments 1 and 2.

We also consider a set of idealized PDF's in which correctly and incorrectly recognized words have fixed confidence probabilities of μ_r and μ_w, respectively. Considering such PDF's eliminates variability in the shape of the PDF as a factor in confidence model performance thereby isolating the effect of varying the mode locations. Moving words from one PDF to the other still idealizes a shift in recognizer performance, while changing the values of μ_r and μ_w idealizes a shift in confidence model performance, which was not considered in experiments 1 and 2.

Figure 4 (asterisks) shows the result of setting μ_r and μ_w to the approximate mode locations in figure 1 and computing the NCE score for different values of p_o. The curve is qualitatively similar to the experiment 2 curve in figure 2, and it has a maximum shifted to the left of center. Figure 5 shows how the NCE score varies (solid lines) as words are shifted from one PDF to the other while, at the same time, the mode locations (μ_w and μ_r) are shifted. These curves have a maximum at $p_o = 0.5$ because the mode locations are shifted simultaneously. These results suggest how the NCE score might vary with recognition accuracy, but a better, and far more computationally expensive experiment, would be to perform successive recognition runs in which an increasing number of performance enhancing features are turned off or on (e.g., vocal tract length normalization, MLLR, rapid adaptation, etc...), thereby naturally degrading or enhancing the recognizer performance.

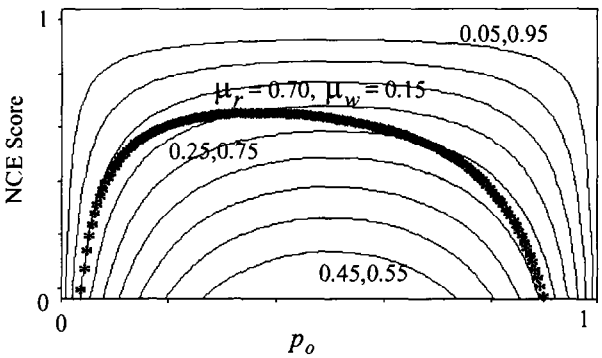

Figure 4. Dependence of NCE score on recognizer performance for μ_r=0.70, μ_w=0.15 (asterisks), and different combinations of μ_r, μ_w ranging from 0.45, 0.55 to 0.05, 0.95 (solid lines).

7. SUMMARY AND CONCLUSIONS

We found that the normalized phoneme duration is not a good confidence model predictor. Nevertheless, the performance of several other predictors indicate that it is worthwhile to consider recognition and confidence training data as sources for confidence model predictors.

The best performing confidence models considered were the GAM and the FDA models, with no decisive advantage in performance for either model. Better performance could potentially come from using even more general statistical models including projection pursuit models [5][14] and mixture discriminant analysis [9]. Neural Networks are an additional area that we have not investigated in the context of confidence modeling.

The DCC score appears to be much less sensitive to recognition accuracy than the NCE score. However, the sensitivity of the NCC score to recognition performance is not as dramatic as suggested by SGR.

8. REFERENCES

[1] Barnett J., Corrada G., Gao G., Gillick L., Ito Y., Lowe S., Manganaro L., Peskin B., **Multilingual Speech Recognition At Dragon Systems,** *Proc. ICSLP'96*, Philadelphia, PA, vol. 4, pp. 2191-2194, October 1996.

[2] Breiman L., Friedman J., Olshen R., and Stone C., **Classification and Regression Trees,** Wadsworth International Group, Belmont, CA, 1984.

[3] Chase L., **Error-Responsive Feedback Mechanisms for Speech Recognizers,** Technical Report No. CMU-RI-TR-97-18, Carnegie Mellon University, The Robotics Institute, 1997.

[4] Eide E., Gish H., Jeanrenaud P., and Mielke A., **Understanding and Improving Speech Recognition Performance Through the Use of Diagnostic Tools,** *Proc. ICASSP-95*, 1995.

[5] Friedman J., and Stuetzle W., **Projection Pursuit Regression,** *J. Amer. Statist. Ass.*, vol. 76, 817-823, 1981.

[6] Gillick L., Ito Y., and Young J., **A Probabilistic Approach to Confidence Estimation and Evaluation,** *Proc. ICASSP-97*, Munich, Germany, vol. 2, pp. 879-882, April 1997.

[7] Hastie T., and Tibshirani R., **Generalized Additive Models,** Chapman and Hall, London, 1990.

[8] Hastie T., Tibshirani R., and Buja A., **Flexible Discriminant Analysis By Optimal Scoring,** *Journal of the American Statistical Association*, vol. 89, 1255-1270, 1994.

[9] Hastie T., Tibshirani R., **Discriminant Analysis By Gaussian Mixtures,** *J. Royal Statist. Soc. (Series B)*, Vol. 58, pp. 155-176, 1996.

[10] Lindeman R., Merenda P., Gold R., **Introduction to Bivariate and Multivariate Analysis,** Scott, Foresman and Company, 1980.

[11] Martin A., Doddington G., Kamm T., Ordowski M., Przybocki M., **The DET Curve in Assessment of Detection Task Performance,** to appear in *Proc. ESCA Eurospeech*, Rhodes, Greece, September 1997.

[12] McCullagh P., and Nelder J., **Generalized Linear Models,** Chapman and Hall, 1983.

[13] Peskin B., Gillick L., Liberman N., Newman M., van Mulbregt P., Wegman S., **Progress in Recognizing Conversational Telephone Speech,** *Proc. ICASSP-97*, Munich, Germany, April 1997.

[14] Roosen C., and Hastie T., **Logistic Response Projection Pursuit,** AT&T Bell Laboratories, Document No. BL011214-930806-09TM, 1993.

[15] Siu M., Gish H., and Richardson F., **Improved Estimation, Evaluation and Applications of Confidence Measures For Speech Recognition,** to appear in *Proc. ESCA Eurospeech*, 1997.

[16] Weintraub M., Beaufays F., Rivlin Z., Konig Y., and Stolcke A., **Neural - Network Based Measures of Confidence For Word Recognition,** *Proc. ICASSP-97*, Munich, Germany, vol. 2, pp. 887-890, April 1997.