A PROBABILISTIC APPROACH TO CONFIDENCE ESTIMATION AND EVALUATION

Larry Gillick, Yoshiko Ito, and Jonathan Young

Dragon Systems, Inc.
320 Nevada St., Newton, MA 02160,USA

ABSTRACT

In this paper we propose a novel way of estimating confidences for words that are recognized by a speech recognition system, together with a natural methodology for evaluating the overall quality of those confidence estimates. Our approach is based on an interpretation of a confidence as the probability that the corresponding recognized word is correct, and makes use of generalized linear models as a means for combining various predictor scores so as to arrive at confidence estimates. Experimental results using these models are presented based on four different sources of speech data: Switchboard, Spanish and Mandarin CallHome, and Wall Street Journal.

1. INTRODUCTION

It is likely that for the foreseeable future automatic speech recognition will be fraught with error. While error may be inevitable, perhaps we can soften the impact of our mistakes if we have some notion of where they may lie. For example, confidence estimates may be highly useful for unsupervised adaptation algorithms, or for information retrieval applications, because of the resulting ability to focus on the parts of the recognized transcript that are more likely to be correct.

In this paper we propose to explicitly model confidence as the probability that a recognized word is correct and present a principled way to estimate the parameters of those models and to evaluate the quality of the resulting estimates. The probability model that we shall describe can easily combine an arbitrary number of predictors, so that we have a general framework for incrementally improving our model as we discover new information-bearing scores or statistics. There have been a number of previous studies aimed at addressing the issue of how to estimate confidences [7,8,9, for example] that we have built on in the present work.

2. EVALUATING CONFIDENCES

Let us suppose that a particular speech recognition system generates a transcription of a set of utterances, the transcription consisting of the sequence of n words W_1, W_2, ..., W_n. Now, over the long term, we could determine the fraction p of recognized words which are correct, and, if we had no other information bearing on the correctness of each word, we would then assess the probability of correctness for each word W as simply being the long term average p. Can we do better than simply using p as the probability of correctness for each word, and how can we measure how much better we are doing?

Let us assume for the moment that we have a confidence model that computes a probability p_i that word W_i is correct, and let us assume further that the "correctness events" are independent. How might we then assess the quality of such a model? Let c_i=1 if the recognized word W_i is correct, and c_i=0 otherwise. The log of the probability of observing c_i is then $l_i = c_i * \log p_i + (1 - c_i) * \log(1 - p_i)$, and the log of the probability of observing the whole sequence of c_i's is the sum of the l_i's.

A key quantity for us will then be L, the average loglikelihood of the observed sequence of c_i's, i.e., the average of the l_i's. In essence, we are regarding the series of observations c_i as a sequence of coin flips where the probability of heads on the ith flip is given by p_i, and we will be using L as our measure of our ability to predict the outcomes of those flips. An important drawback of this approach lies in the (somewhat unrealistic) assumption that the "coin flips" are independent, conditional on the specific probabilities p_i.

One way to assess the quality of L is by comparing it to a baseline loglikelihood L_{base} computed from the long term correctness rate p. Let n_c be the total number of recognized words that are correct, with $n-n_c$ being the total number that are incorrect. Then,

$$L_{base} = 1/n * (n_c * \log(p) + (n - n_c) * \log(1 - p))$$

If L is considerably larger than L_{base}, then we would infer that our confidence predictions are better than random. Indeed, the difference between L and L_{base} can serve as a measure of the quality of the confidence assessments being made, since it is actually equivalent to the likelihood ratio test of the null hypothesis that the confidence model is no better than the constant confidence model. One can also easily compute an estimated standard error of L, as it is an average of quantities

that are assumed independent, and assess improvements to existing confidence estimates by seeing whether L increases by a statistically significant amount. For purposes of human intuition, rather than looking at L itself we favor the use of the more intuitive $p_g = \exp(L)$, which is the geometric mean of the predicted probabilities of the observed sequence of c_i's.

Our analysis does not deal with the fact that there are different kinds of errors made by speech recognizers: deletions, substitutions, and insertions. We could modify our analysis so as to deal with these issues, but in this paper we will focus our attention on simply predicting whether there has been an error in the recognized words. In particular, we therefore address substitution and insertion errors (though we lump them together) but not deletion errors.

It is important to note that the correctness rate that we refer to is not equal to one minus the word error rate (WER), the quantity which is normally used in speech recognition as the measure of recognition quality. The correctness rate is the proportion of **recognized** words which are correct, while the WER refers to the ratio of the number of errors to the number of words in the correct transcript.

3. MODELS FOR CORRECTNESS PROBABILITIES

We suppose that there is a development test set that has been transcribed with the speech recognizer being evaluated and that for each recognized word W in the set we know whether it is correct or not. In addition we assume that we have available for each word a vector **x** of k scores that we expect would carry information concerning the hypothesized word's likelihood of being correct. For example, the vector **x** might contain such predictors as the average score per frame for W or the average "Best Score" [1], which involves using the best available output distribution for each frame. It might also include the language model score and scores from other sources.

Although there are a variety of types of models that one might use to relate the vector of k predictors to the probability of being correct for this sort of data, in this paper we will be exploring the use of generalized linear models (glm) [2].

The glm assumes that $g(p) = \mathbf{b}^T\mathbf{x} + a$, where p is the confidence, g(p) is a monotone function (called the link function) mapping the unit interval to the real line, **x** is the vector of k predictor scores, $\mathbf{b}^T$ is the transpose of **b,** a vector of unknown parameters, and "a" is the "intercept" term. In this paper we compare two commonly used link functions: the logit [$g(p) = \text{logit}(p) = \log(p/(1-p))$] and the complementary log [$g(p) = \log(-\log(1-p))$]. (The use of a glm with the logit link function is commonly known as logistic regression.) We estimate a and **b** by choosing them to maximize L on the development test set. It is also possible to estimate the standard errors of these estimates from the diagonal elements of the inverse of the matrix of the second derivatives of L. By examining which of the elements of **b** are statistically different from zero, we can infer which of the scores might actually be informative.

Our models made use of five predictors together with an intercept term: word duration (WDUR), the language model score (LM), "acoustic score minus best score" (SCR), "nbest score" (NBEST), and "active node count" (ACTV).

The "acoustic score minus best score" of a recognized word is computed as follows. First, a recognized transcription is Viterbi aligned to its utterance. Then for each word, the "acoustic score" is computed by scoring each speech frame that was assigned to the word against the output distribution of the appropriate state and by then averaging those scores. The "best score" is computed by determining for each speech frame the score of the best scoring output distribution in the acoustic model and by then averaging those scores over the word The predictor "SCR" is the difference between these two average scores – one expects that when it is near zero, the word is more likely to be correct, since the best possible acoustic match is close to the acoustic match of the word that was recognized. Our use of "best score" is similar to the acoustic score normalization that Young [7] performed in her study on misrecognized and out-of-vocabulary words except that she used the score obtained from phoneme recognition for normalization purposes.

The "nbest score" of a recognized word is the fraction of the nbest list (n=100 in this study) that contains the given word in the correct position – the idea here being that the stability of a recognized word on such a list should be a good indicator of whether it is actually correct. The "active node count" is the average over a word of the number of states (in the whole vocabulary) that were active on each frame – the intuition here is that a large number of active nodes reflects a large degree of uncertainty of the recognizer about the data it is decoding.

4. EXPERIMENTAL RESULTS AND DISCUSSION

We have carried out experimental investigations of confidence prediction on four speech corpora: Switchboard [3], Spanish and Mandarin CallHome[4], and Wall Street Journal (WSJ). The first three corpora consist of spontaneous telephone conversations. The WSJ data consists of read speech recorded over a high quality microphone.

From each corpus we chose three disjoint sets of utterances: one for training the recognition models, one for training the confidence models, and one for testing the confidence models. The speech was recognized with Dragon Systems' large vocabulary continuous speech recognizer[6]. The amount of acoustic data used for building the acoustic models was 16 hours each for CallHome Spanish and Mandarin, 170 hours for Switchboard, and 100 hours for Wall Street Journal. All acoustic models were speaker and gender-independent. All models except WSJ were speaker-normalized. The CallHome

models were also speaker-adapted (unsupervised). We used standard backoff bigram language models for all corpora. The vocabulary sizes were 11k for CallHome Spanish, 8k for CallHome Mandarin, 10k for Switchboard, and 20k for Wall Street Journal. The Spanish CallHome language model was interpolated with the language model constructed from the ECI corpus [5], consisting of transcriptions of talk radio broadcasts The five predictor scores were computed for each recognized word and used for training the models.

Table 1 lists the t-values of the model parameters for the glm using logit as the link function. The t-values are the ratios of the estimated parameter values to their estimated standard errors. The table indicates that NBEST tended to be the most informative predictor, followed by SCR and LM. ACTV, intercept, and WDUR are informative to varying degrees depending on the corpus. Our observation that NBEST was one of the most informative predictors is in agreement with that made by Eide et al. [8]. Similarly Cox and Rose [9] found that the number of competing hypotheses above the pruning threshold was a useful predictor.

Table 2 summarizes the confidence estimates from the two models that we examined, one for each of the link functions: it exhibits the geometric means (p_g's) of the estimated probabilities for the outcomes (the c_i's). When the value of p_g is larger, that is a sign that the confidences are better. We recommend the quantity p_g as a somewhat intuitive quantitative representation of the quality of a set of confidence estimates, which is, at the same time, a monotone function of the average loglikelihood L (which we maximized during training).

The table shows that both models produced confidences that were substantially better than just assigning the long-term word correct rate p to each word – p-values for this comparison are very small. The confidences estimated from the two glms are close to each other although the complementary log link function seems to fit the data somewhat better than the logit link. A paired t-test was performed on the loglikelihoods obtained with the logit and the complementary log link functions to see whether the observed differences were significant. The p-values were 0.45 for Spanish, 0.52 for Mandarin, less than .001 for Switchboard and for Wall Street Journal.

The signs of the model parameters are generally in agreement with our intuition. We would expect NBEST to have a positive sign and ACTV to have a negative sign because a larger value of NBEST and a smaller value of ACTV would both suggest that the recognizer is more "certain" of its answer. Similarly, smaller values of SCR and LM indicate a better fit between the recognition models and the word chosen by the recognizer. Because these scores are negative log probabilities, the smaller scores are better. The negative sign of WDUR for Mandarin is puzzling as longer words are normally easier to recognize.

In addition to predicting whether a recognized word is correct, confidence measures can be used for predicting the speaker ranking. Figure 1 shows the average word correctness versus average confidence (computed with the logit link function) for each test speaker from each corpus. The figure shows a very strong association between average word correctness and average confidence. The apparent strong correlation across the corpora may be somewhat of an artifact of having all corpora on the same plot. However, the correlation within each corpus is also strong. The correlation coefficients were 0.82 for Switchboard, 0.83 for WSJ, 0.82 for Mandarin CallHome, and 0.89 for Spanish CallHome. The p-values for all correlation coefficients were less than 0.01 according to Pearson's test.

We have presented an exploration of confidence models that estimate the probability that a word recognized by a speech recognition system is correct and described a way to evaluate the quality of the resulting confidence estimates. Moreover, these models allow one to combine information from an arbitrary collection of scores or statistics. We have shown on four speech corpora that our models do considerably better than chance, although there is much room for improvement. We have also shown that the confidence estimates can be used to predict the word correctness rate of a speaker with reasonable accuracy.

5. ONGOING AND FUTURE WORK

We are working on improving our confidence estimates from two directions. One aspect of our work is devoted to the search for additional informative confidence predictors. We have not, for example, explored sentence-level predictors such as the estimated speaking rate, utterance duration, and speed of decoding. The other way to improve confidences is to improve our confidence models themselves. We plan to investigate additional link functions, different types of models such as classification and regression trees, and the possibility of combining multiple sorts of models.

One of the issues not addressed here is that of comparing confidence estimates corresponding to different recognition error rates, such as comparing confidences estimated on different corpora, or confidences estimated on the same speech material but recognized with different systems. This is because in the loglikelihood-based evaluation of confidences that we have presented, the recognition error rate and the quality of confidence estimation cannot easily be separated. We plan to address this issue in future work.

6. REFERENCES

[1] S. Mendoza, et al., "Automatic Language Identification Using Large Vocabulary Continuous Speech Recognition," Proc. ICASSP 1996, pp 785-788.

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

[3] J. Godfrey, E. Holliman, and J. McDaniel: "SWITCHBOARD: Telephone Speech Corpus for Research and Development," Proc. ICASSP 1992, pp 517-520.

[4] J. Godfrey, “Multilingual Speech Databases at LDC,” Proc. ARPA Human Langage Technology Workshop, 1994, pp 23-26.
[5] D. Graff, “Multilingual Text Resources at the Linguistic Data Consortium,” ARPA Human Langage Technology Workshop, 1994, pp 18-22.
[6] B. Peskin, et al., “Progress in Recognizing Conversational Telephone Speech," Proc. ICASSP 1997.
[7] S. Young, “Detecting Misrecognitions and Out-of-vocabulary Words," Proc. ICASSP 1994, v.2, pp 21-24.
[8] E. Eide, H Gish, P. Jeanrenaud, and A. Mielke, “Understanding and Improving Speech Recognition Performance Through the Use of Diagnostic Tools," Proc. ICASSP 1995, pp 221-224.
[9] S. Cox and R. Rose, “Confidence Measures for the Switchboard Database,”, Proc. ICASSP 1996, pp. 511-514.

Table 1. Estimated model parameters for the logit link function.

Predictors	Spanish CallHome		Mandarin CallHome		Switchboard		WSJ	
	value	t-value	value	t-value	value	t-value	value	t-value
(Intercept)	1.4	14.0	0.5	5.4	1.1	7.0	0.7	2.3
WDUR	0.00064	0.5	-0.0073	-5.0	0.0068	2.9	0.046	5.7
LM	-0.0088	-22.6	-0.0056	-13.6	-0.0061	-10.1	-0.013	-13.5
SCR	-0.26	-29.4	-0.22	-25.5	-0.20	-18.6	-0.093	-9.8
NBEST	1.8	25.5	2.1	31.7	2.7	27.0	4.9	20.1
ACTV	-6.1e-05	-14.8	-2.0e-05	-11.7	-3.0e-05	-7.5	-3.1e-05	-2.8

Table 2. Predicted geometric means. p_g = geometric mean = exp(average loglikelihood)

	Spanish CallHome		Mandarin CallHome		Switchboard		WSJ	
	train	test	train	test	train	test	train	test
base	0.505	0.504	0.516	0.513	0.530	0.526	0.690	0.688
logit	0.564	0.563	0.567	0.566	0.600	0.585	0.776	0.766
clog	0.564	0.563	0.568	0.566	0.605	0.587	0.781	0.769
word correct rate	0.430	0.440	0.374	0.388	0.668	0.657	0.878	0.876
# words	17071	16169	17911	16278	8787	13161	5509	6429

Figure1. Plot by speaker of average word correctness vs. average confidence.