Posterior Probability Based Confidence Measures Applied to a Chidren's Speech Reading Tracking System

Daniel Bolanos
HCTLab-EPS. Universidad Autónoma
de Madrid, SPAIN
daniel.bolanos@uam.es

Wayne H. Ward
CSLR, University of Colorado at
Boulder USA
whw@cslr.colorado.edu

Abstract

In this paper, we present improved word-level confidence measures based on posterior probabilities for children's oral reading continuous speech recognition. Initially we compute posterior probability based confidence measures on word graphs using a forward-backward algorithm. We study how an increase of the word graph density affects the quality of these confidence measures. For this purpose we merge word graphs obtained using three different language models and compute the previous confidence measures over the resulting word graph. This produces a relative error reduction of 8% in Confidence Error Rate compared to the baseline confidence measure. Moreover the system operating range is increased significantly.

1 Introduction

When dealing with children's continuous speech recognition, it is difficult to obtain satisfactory acoustic models due to the great variability of children's speech. Oral reading tracking systems use a speech recognizer to determine whether a child has read a known passage correctly. Such systems often cope with lack of adequate acoustic models by taking advantage of very tight language models that reflect what the child is supposed to be reading. A recognizer for a reading tracking system was developed in this context (Hagen, 2006) in which the single best scoring hypothesis from the recognizer is used as the hypothesis for what the child read. Comparing these hypotheses against the hand transcriptions for the speech yields a Word Error Rate around 10% when tested on 3rd, 4th and 5th grade children. However, the use of this kind of restrictive language model can make rejection of errors difficult and leads the system to consider misread words as correct. We apply confidence measures to the recognized words as a basis for detecting words that have been misread or skipped.

Previous work has shown (Wessel, 2001) that confidence measures based on word posterior probabilities estimated over word graphs outperform alternative confidence measures (Kemp, 1997) such as acoustic stability and hypothesis density. In the following discussion we will take advantage of this technique in order to obtain word level confidence estimates in the context of children's speech reading tracking.

2 Posterior Probability Based Confidence Measures

SONIC (Pellom, 2001), the continuous speech recognizer used in this work, is able to output the results of the first-pass decoding process in the form of word lattices. Each of these lattices can be considered as an acyclic, directed, weighted word graph, and used (Hacioglu, 2002) during the decoding process to calculate word posterior probabilities. This calculation is carried out by the forward-backward algorithm considering edges as HMM-like states, where emission probabilities are the acoustic models scores and transition probabilities between links are obtained from the trigram language model used. Taking these posterior probabilities attached to each word on the graph we estimate the following confidence measures, where $[w;s,e]$ is a word hypothesis starting at time s and ending at time e and $p([w;s,e]|x_1^T)$ is the posterior probability for the hypothesis word w for the acoustic observation sequence x_1^T.

$$C([w;s,e]) = p([w;s,e] \mid x_1^T) \tag{1}$$

$$C_{sec}([w;s,e]) = \sum_{\substack{[w;s',e']: \\ \{s,\ldots,e\}\cap\{s',\ldots,e'\}\neq 0}} p([w;s',e'] \mid x_1^T) \tag{2}$$

$$C_{med}([w;s,e]) = \sum_{[w;s',e']:s'\leq\left\lceil {}^{s+e}\!/_{2}\right\rceil\leq e'} p([w;s',e'] \mid x_1^T) \tag{3}$$

$$C_{med'}([w;s,e]) = \sum_{\substack{[w;s',e']: \\ s'\leq\left\lceil {}^{s+e}\!/_{2}\right\rceil\leq e'\wedge(s=s'\vee e=e')}} p([w;s',e'] \mid x_1^T) \tag{4}$$

$$C_{max}([w;s,e]) = \max_{e_{\max}\in\{s,\ldots,e\}} \sum_{[w;s',e']:s'\leq e_{\max}\leq e'} p([w;s',e'] \mid x_1^T) \tag{5}$$

In (1) posterior probabilities are taken directly as a confidence measure for a word hypothesis. However, previous work (Wessel, 2001) has demonstrated that this measure of confidence does not give satisfactory results. The reason is that the fixed starting and ending time frames of a hypothesis word strongly determine the paths involved in the calculation of the forward-backward probabilities. The following confidence measures calculated (2), (3), (4) and (5) take advantage of the fact that, usually, word hypotheses with similar starting and ending time frames represent the same word and therefore makes sense to consider the summation of the posterior probabilities of these words as a confidence measure. The differences between them consist basically of how word hypotheses are selected to be used in the posterior probability accumulation process. In (2) word hypotheses that overlap in time are considered, this procedure has been shown to perform very well as a confidence measure but suffers from a lack of normalization since the summation of the accumulated posterior probabilities over all different words on a single time frame no longer sums to one. To cope with this drawback (3), (4) and (5) are used. In all of them, the posterior probability accumulation process is carried out over all different hypotheses of a word with at least one time frame in common.

Note that while Csec, Cmed and Cmax were proposed on (Wessel, 2001) Cmed’ is a variation of Cmed in which only words with the same first or last frame are taken into account in the posterior probability accumulation process. Cmed’ performs slightly better than Cmed in terms of Confidence Error Rate as can be seen in Table 1.

2.1 Score normalization

One of the characteristics of children’s oral reading is the propagation of errors due to repetitions of text segments, self-corrections and other kinds of disfluencies. We compensate for these events by doing the following normalization over the confidence measures applied to each word. The parameters μ and λ are estimated on a development set distinct from the testing set to avoid over-adaptation.

$$C_{norm}([w_i,s_i,e_i]) = \mu C_{\max}([w_{i-1},s_{i-1},e_{i-1}]) + \lambda C_{\max}([w_i,s_i,e_i]) + (1-\mu-\lambda)C_{\max}([w_{i+1},s_{i+1},e_{i+1}]) \tag{6}$$

2.2 Experimental results

We present experimental results on a corpus composed of the CU Prompted and Read Children’s Speech Corpus (Cole, 2006), the OGI Kid’s speech corpus (Shobaki, 2000) and the CU Read and Summarized Story Corpus (Cole, 2006). Children’s acoustic models are estimated from over 62 hours of audio from the CU Prompted and Read Children’s Speech Corpus, the OGI Kids’ speech corpus grade K through 5, and data from 1st and 2nd graders found in the CU Read and Summarized Story Corpus. Confidence measures are evaluated on the 106 3rd, 4th and 5th graders from the CU Read and Summarized Story Corpus.

To evaluate the performance of the confidence measures applied we use the confidence error rate (CER), defined as the number of incorrectly assigned tags divided by the total number of recognized words. The CER of the baseline system is calculated by dividing the number of insertions and substitutions by the number of recognized words. Since the CER depends on the tagging threshold selected, as well as the acoustic and language model scaling factors, these parameters are adjusted not on the test corpus but on a different cross-validation corpus.

Confidence Measure	**CER**	**Error reduction**
Baseline	9.70%	0.00%
C	9.24%	4.74%
C_{sec}	8.05%	17.01%
C_{med}	8.11%	16.39%
$C_{med'}$	8.08%	16.70%
C_{max}	8.05%	17.01%
C_{norm}	7.93%	18.25%

Table 1. Confidence error rates and error reduction

Table 1 summarizes the CER for the confidence measures applied. It can be seen that while word posterior probabilities used directly as a confidence measure don’t perform well, the normalized version of C_{max} performs better than the others.

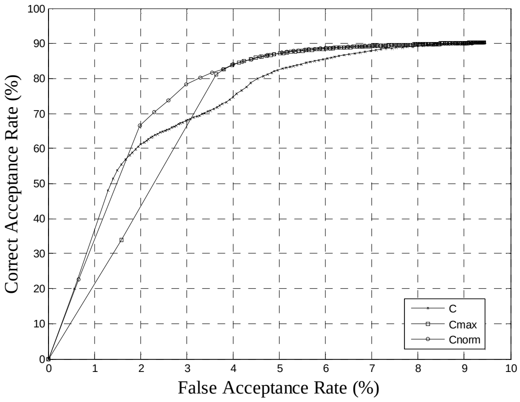

Figure 1. Receiver operating characteristic (ROC) curves.

As can be seen in Figure 1, the C_{norm} measure also has the best performance for the ROC. Note that correct acceptance is tagging a correct word as correct and false acceptance is tagging an incorrect word as correct.

3 Increasing Word Graph Density to Improve the Quality of Confidence Measures

During the decoding process an adaptive language model is used due to the fact that words that are likely to be spoken next can be anticipated based upon the words in the text that are currently being read. For this purpose position-sensitive trigram language models are obtained (Hagen, 2006) partitioning the training text into overlapping regions. After decoding each utterance, the position-sensitive language model that gives a higher probability to the last recognized words is selected for the first pass decoding of the subsequent utterance. This results in a very low perplexity language model.

The main problem when estimating posterior probability based confidence measures over a word graph generated using a very tight language model is the low word graph density, defined as the word hypotheses per spoken word. Previous work (Wessel, 2001; Fabian, 2003) has shown that the word graph density has a clear impact on the quality of confidence measures and therefore it is necessary to adjust the WGD in order to get the best confidence error rates.

To cope with this problem we generate word graphs using the following language models.

1) The original trigram adapted language model that produces the best output in terms of WER.

2) A trigram language model without adaptation.

3) A bigram language model without adaptation.

For each utterance we take the three word graphs generated and merge them into one graph, and then we use it to estimate the confidence measures over the hypothesis generated. During the posterior probability accumulation process, hypotheses coming from different graphs are weighted differently. From now we refer to this confidence measure as C_{merge}.

$$C_{merge}([w,s,e]) = \alpha C_{\max(adapted)}([w,s,e]) + \\ \beta C_{\max(trigram)}([w,s,e]) + (1-\alpha-\beta) C_{\max(bigram)}([w,s,e]) \qquad (7)$$

After doing the merging process we also do a score normalization as described in 2.1.

3.1 Experimental results

We conducted experiments with two configurations, in the first one we build a word graph merging word graphs obtained with language models 1 and 2. In the second configuration we build a word graph that merges all three language models in order to increase the graph density further. The results are shown in Table 2. Increasing the word graph density does provide better performance of the confidence estimation measure. The ROC curves shown in Figure 2 also demonstrate that the confidence measures generated from the more dense graphs perform better.

Confidence Measure	WGD	CER	Error Reduction
Baseline	6.45	9.70%	0.00%
C_{norm}	6.45	7.93%	18.25%
C_{merge} (config. 1)	16.64	7.51%	22.58%
C_{merge} (config. 2)	32.79	7.44%	23.30%

Table 2. Word graph densitiy (WGD), confidence error rate (CER) and error reduction respect the baseline for the confidence measures applied using the different configurations.

In the first configuration, the value for α, i.e., the weight applied to the hypotheses coming from the word graph obtained with language model 1, that yields the best performance is in the range of (0.7-0.77), while the weighting value for hypotheses coming from the graph obtained with language model 2 is in the range of (0.23-0.3). In the second configuration, the values for α, β and (1-*α*-*β*), that yield the best performance are in the range of (0.65-0.7), (0.15-0.2) and (0.1-0.15). These values show that hypotheses coming from graphs generated with smoother languages models must be weighted less during the posterior probability accumulation process in order to obtain satisfactory results.

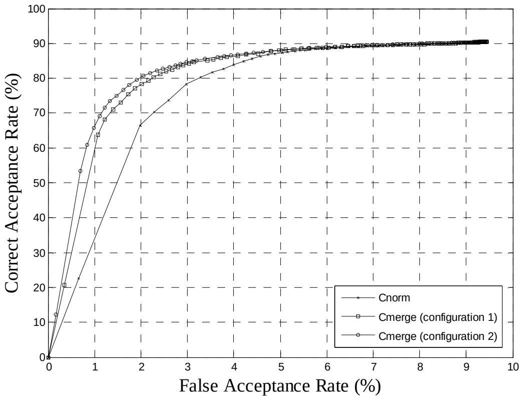

Figure 2. Receiver operating characteristics (ROC) curve.

4 Conclusions

We have evaluated the performance of using confidence measures based on word posterior probabilities to reject misrecognized words in hypotheses generated by a speech recognizer in a reading tracker task. While this technique has been shown to work relatively well for large vocabulary speech recognition, the task of a reading tracker presents a special case. A very tight language model produces the best word error rate, but does not produce a dense enough graph to provide good confidence estimates. We have shown that, adding hypotheses generated from more smoothed language models to increase the word graph density and doing a score normalization based on word context information, the performance of the confidence measures is improved significantly.

5 Future Work

The current work uses posterior probabilities of words to generate confidence scores that are used to make accept/reject decisions on the words in a hypothesis produced by the recognizer. In a reading tracker, the final goal is to estimate whether words in the reference string were read correctly. We will apply the confidence measures estimated for the words in the speech recognition output as features to make a classification as to whether words in the reference string were read correctly.

References

R. Cole, P. Hossom, and B. Pellom. University of Colorado prompted and read children's speech corpus. Technical Report TR-CSLR-2006-02, University of Colorado, 2006.

R. Cole and B. Pellom. University of Colorado read and summarized story corpus. Technical Report TR-CSLR-2006-03, University of Colorado, 2006.

T. Fabian, R. Lieb, G. Ruske and M. Thomae. "Impact of Word Graph Density on the Quality of Posterior Probability Based Confidence Measures," in Proc. 8th Eur. Conf. Speech, Communication, Technology, Geneva, Switzerland September 1-4, 2003, pp. 917–920.

K. Hacioglu and W. Ward, "A Concept Graph Based Confidence Measure", in ICASSP, Orlando-Florida, USA, 2002.

A. Hagen, "Advances in Children's Speech Recognition with Application to Interactive Literacy Tutors", *Ph.D. thesis*, University of Colorado, Dept. of Computer Science, 2006.

T. Kemp and T. Schaaf, "Estimating confidence using word latices," in Proc. 5th Eur. Conf. Speech, Communication, Technology 1997, Rhodes, Greece, Sept. 1997, pp. 827–830.

B. Pellom, "Sonic: The University of Colorado Continuous Speech Recognizer", Technical Report TR-CSLR-2001-01, CSLR, University of Colorado, March 2001.

K. Shobaki, J.-P. Hosom, and R. Cole. The OGI kids' speech corpus and recognizers. In 6th ICSLP, Beijing, China, 2000.

F. Wessel, R. Schlüter, K. Macherey, and H. Ney, "Confidence Measures for Large Vocabulary Continuous Speech Recognition," IEEE Transactions on Speech and Audio Processing, vol. 9, pp. 288–298, March 2001.