Word Graph Rescoring Using Confidence Measures

Pablo Fetter[1] *Frédéric Dandurand*[2] *Peter Regel-Brietzmann*[1]

[1] Daimler-Benz AG, Research and Technology, Wilhelm-Runge-Str. 11,
D-89081 Ulm, Germany
[2] McGill University, Department of Electrical Engineering, 3480 University Street
Montreal, H3A 2A7, Canada
e-mail: fetter@dbag.ulm.daimlerbenz.com

ABSTRACT

This paper presents a novel approach to using confidence scores for word graph rescoring. For each word in the system's vocabulary, we computed the probability that the observation is correct given its acoustic score. Afterwards, we used these probabilities for rescoring word graphs outputted by the recognizer. We will present some implementation details as well as accuracy improvements obtained using this method.

1. INTRODUCTION

Most speech recognition systems are based on the maximization of the Bayes' rule:

$$P(W|A) = \frac{P(A|W)P(W)}{P(A)} \quad (1)$$

where

$P(W|A)$ is the probability of the word string W given the acoustic signal A;
$P(A|W)$ is the probability that the acoustic signal A was produced by the word string W and is usually given by the Hidden Markov Models;
$P(W)$ is the *a priori* probability of the word string W (given by the language model); and
$P(A)$ is the probability of the acoustic signal A.

Within this framework a speech recognizer can compute the most likely word string W for a given acoustic signal A using approximations of $P(W|A)$. These approximations are often referred to as *scores*.

In [6] and [7] the concept of *confidence mapping* is introduced for pattern recognition problems. Confidence represents the probability that an event is correctly classified. In the simplest approach, confidence is computed from the so-called *eigen* and *fremd* distributions, which contain the observations of a particular event correctly and falsely classified, respectively.

In [9] Young uses confidence measures for detecting unknown words in the ATIS task, which are based on score normalization. A very similar technique was implemented for German in [3]. Rivlin [5] also used correct and incorrect distributions in order to compute phoneme-based confidence measures. In [2] Jeanrenaud et al. presented two different approaches for estimating confidence scores: using the Bayes' rule on correct and incorrect distributions, and using a rank of hypotheses in a large collection of putative hits. In the current paper we extend the Bayesian approach and present some implementation details as well as accuracy improvements obtained using this method.

Our work is based on the Verbmobil database [8] (also called the German Spontaneous Scheduling Task). At the time when this study began, about 400 human-to-human dialogs had been collected and transcribed at various German universities. The complete corpus contains over 200,000 word entries, including phenomena such as pronunciation variations, word fragments, disfluencies, repairs, etc.

We will first describe the training method we used and then the testing conditions for word graph rescoring. Finally we will present some optimization methods for our test.

2. TRAINING

Our HMM-based speech recognition system [1] was trained with the material available for the Evaluation '95 within the Verbmobil project. In order to achieve realistic testing conditions, a disjunct set of dialogs (approximately 200) was used for confidence training. Hypotheses for these 200 dialogs were generated and aligned to the (spoken) reference text strings. Figure 1 shows an aligned utterance. The second column shows the spoken string, and the third shows the string hypothesized by the recognizer.

Every hypothesized word is then marked as being correct or incorrect. The first column of Figure 1 shows the tags **COR**rect or, if incorrect, **SUB**stitution, **DEL**etion, or **INS**ertion for the spoken words. For each word, two score histograms are then generated, one for each category (correct and incorrect). Figures 2 and 3 show examples of well

COR	SCH"ONEN	SCH"ONEN	127.11	12	41
DEL	GUTEN	***	0.00	0	0
COR	TAG	TAG	57.86	43	62
COR	HERR	HERR	43.45	66	76
INS	***	WEITER	96.73	75	104
SUB	<UNK>	LEID	118.95	110	145

Figure 1: Example of alignment.

and badly separated distributions respectively.

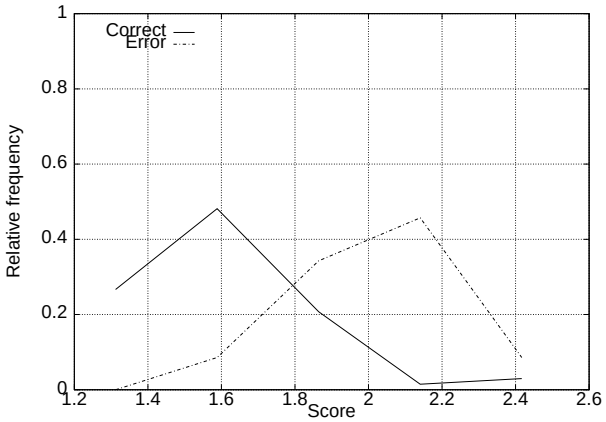

Figure 2: The word *denke* exhibits two well separated distributions. Scores are frame-based.

These histograms represent, respectively, the probability of a word w having a particular score s_w, given that the word was correctly or incorrectly recognized, that is $P_w(s_w|C)$ or $P_w(s_w|E)$. Finally, for each word w, the *a posteriori* confidence[1] $P_w(C|s_w)$ is computed using Bayes' Rule:

$$\begin{aligned} P_w(C|s_w) &= \frac{P_w(s_w|C)\,P_w(C)}{P_w(s_w)} \\ &= \frac{P_w(s_w|C)\,P_w(C)}{P_w(s_w|C)\,P_w(C) + P_w(s_w|E)\,P_w(E)} \\ &= \frac{1}{1 + \frac{P_w(s_w|E)}{P_w(s_w|C)}\,\frac{P_w(E)}{P_w(C)}} \end{aligned} \quad (2)$$

This confidence represents the probability that an observation is correct given its acoustic score. The confidence $P_w(C|s_w)$ is a central issue in this work, since it forms the basis for word graph rescoring.

[1] *a posteriori* confidence, or simply confidence.

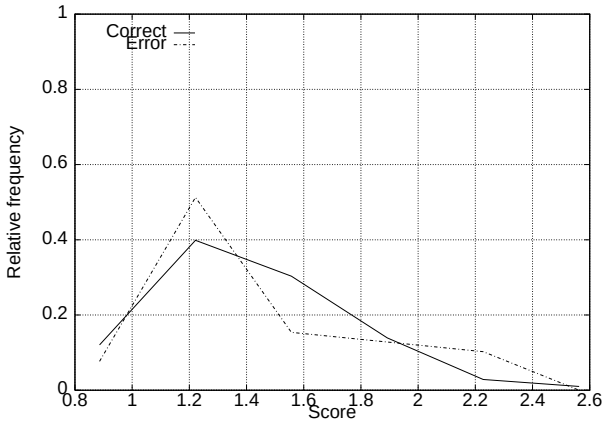

Figure 3: The word *noch* exhibits overlapped correct and incorrect distributions. Scores are again frame-based.

3. WORD GRAPH RESCORING

A word graph is a sextuple (a, e, w, s_w, t_a, t_e), where a and e are the logical start and end nodes of word w; t_a and t_e are the corresponding time frames of the start and end nodes; and s_w is the log HMM score of the word w in the region $[t_a, t_e]$. Figure 4 shows an example of such a word graph.

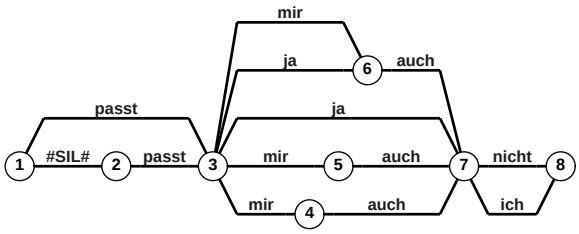

Figure 4: Example of a word graph.

A simple rescoring method was implemented. Each word hypothesis w in the word graph was rescored using its weigthed log confidence, as follows:

$$S_w = s_w + k \ln(P_w(C|s_w)) \quad (3)$$

where S_w is the updated acoustic score of word w, s_w its original acoustic score, k the empirically optimized rescoring factor, and $P_w(C|s_w)$ the confidence.

We could not compute confidence for all the words in the recognizer's vocabulary (more on this in Section 4.1.), so we used for those words (instead of $P_w(C|s_w)$) what we dubbed *a priori confidence*:

$$APC = \frac{Count(COR)}{Count(COR) + Count(SUB) + Count(INS)}$$

where $Count(\ldots)$ represents the number of correct (COR), substituted (SUB), and inserted (INS) words in the training set. Note that since deletions are not considered here, *a priori* confidence is a concept distinct from word accuracy.

All experiments were carried out using unseen material during the training phase. This is official test set for the Evaluation '95 within the project Verbmobil. It contains 329 utterances and 7204 words. The recognition lexicon comprises approximately 3500 words[2]. We used a bigram language model with a test set perplexity of 95 (116 when transitions involving out-of-vocabulary words are included). Our baseline system achieved a word accuracy of 62.2%. This performance is suboptimal, since word graphs of a low density ($\approx$ 15 hyp/word) were produced by the recognition engine and used in the second stage of our speech recognition system [4]. We chose this system for efficiency reasons.

4. OPTIMIZATION

Given the training and testing framework presented in the previous sections, we ran experiments to find out the influence of different parameters on overall performance. In the next subsections we will present detailed results.

4.1. Minimum Number of Observations

The first problem we encountered when computing histograms of correct and incorrect distributions was the lack of training material. Many words appeared only once (either correctly or incorrectly) in our training set. But since we have to compute standard deviation for further processing, a theoretical minimum number of observations per distribution $MinObs = 2$ is required. There is a compromise between the coverage of the test set (which is high for a low minimum number of observations) and the quality of the estimated distributions (which is low for a low minimum number of observations). In order to find an optimum, we ran rescoring experiments for different values of $MinObs$. The number of words having a $MinObs = \{10, 15, 20\}$ in our training set was 309, 231, and 196 respectively. The coverage of the test set was 83%, 78%, and 75%. We found that $MinObs = 15$ was a good compromise, and chose it for subsequent experiments.

4.2. Frame Scores

In [9] and [3] confidence is computed using frame scores (that is, word scores are divided by the word duration). We found that for frame scores, (correct and incorrect) histograms are much more separated than for word scores. This is probably because frame scores reduce the influence of speaking rate. However, this could not be confirmed by rescoring experiments, as can be seen in Figure 5. The curves show word accuracy results for various rescoring factors (k in Equation 3).

[2] For this work we used a lexicon slightly larger than the official lexicon of the Evaluation '95.

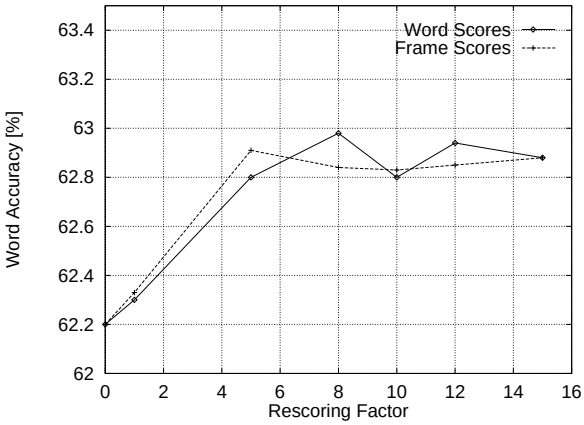

Figure 5: Word scores vs. frame scores.

When rescoring word graphs that have higher densities, frame scores do seem to have better discrimination power, but the gain is minimal. We could not reach general conclusions on this topic and will use word scores for subsequent experiments because they can be computed faster.

4.3. Score Normalization

In [9], Young introduces the concept of score normalization. She uses a phoneme recognizer working in parallel with a lexically constrained (word) recognizer in order to estimate $P(A)$ (see Equation 1). This estimated probability is then used to normalize the scores of the word recognizer. Using this method, very positive results were achieved.

We implemented a similar approach and obtained word accuracy improvements compared to the baseline system, but not compared to the systems introduced in the previous sections. Since the estimation of $P(A)$ is computationally expensive, and the results do not seem to justify the effort, we decided not to integrate this topic in our system.

4.4. Discrete and Continuous Distributions

As mentioned above, one of the main implementation problems in computing confidence for words is the lack of data. Due to this, the curves in Figures 2 and 3 have a rough appearence and are only a coarse approximation to the actual (correct and incorrect) distributions. In order to cope with this, we computed continuous distributions using maximum likelihood estimation, and reran the rescoring experiments. The results are shown in Figure 6. Word graph rescoring using continuous distributions yields slightly, but consistently better results than when using discrete distributions. This is very positive, since continuous distributions have the additional advantage of requiring less memory—only two parameters per distribution are necessary. The conclusions arrived

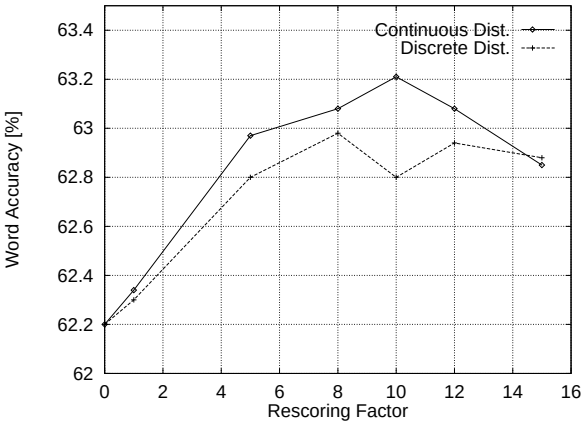

Figure 6: Discrete vs. continuous distributions

above for discrete distributions (Sections 4.1., 4.2., and 4.3.) are also valid for continuous distributions.

5. CONCLUSIONS

We computed confidence measures for every word in the recognition vocabulary using normal distributions for the correct and incorrect histograms, which were extracted from data hypothesized by the recognizer. We achieved an improvement of approximately 1% word accuracy (from 62.20% to 63.21%) using confidence for word graph rescoring. Confidence can be very easily integrated into our two-stage speech recognition system [4], as an additional knowledge source.

ACKNOWLEDGMENTS

This work was partially funded by the German Federal Ministry of Education, Science, Research and Technology (BMBF). Any opinions and conclusions expressed in this paper are those of the authors.

The authors gratefully thank all colleagues that contributed to our large vocabulary speech recognition system, without which this work could not have been possible: Fritz Class, Alfred Kaltenmeier and Thomas Kuhn. We also thank Udo Haiber and Prof. Jürgen Schürmann for his invaluable advice and ideas, and David Stall for converting this manuscript into correct English. Finally, thanks to Monika Woszczyna and Thomas Kemp for sharing the experiences gained in the master's thesis [3].

This work was inspired by the interesting ideas of Sheryl Young [9] and we would like to dedicate it to her memory.

REFERENCES

[1] F. Class, A. Kaltenmeier, and P. Regel-Brietzmann. Optimization of an HMM-Based Continuous Speech Recognizer. In *Proc. EUROSPEECH'93*, pages 803–806, Berlin, Germany, 1993.

[2] P. Jeanrenaud, M. Siu, and H. Gish. Large vocabulary word scoring as a basis for transcription generation. In *Proc. EUROSPEECH'95*, pages 2149–2152, Madrid, Spain, 1995.

[3] Stefan Kopp. Sicherheitsmasse für die Erkennung spontaner Sprache. Master's thesis, University of Karlsruhe, Karlsruhe, Germany, 1995.

[4] Thomas Kuhn, Pablo Fetter, Alfred Kaltenmeier, and Peter Regel-Brietzmann. DP-Based Wordgraph Pruning. In *Proc. ICASSP'96*, Atlanta, USA, 1996.

[5] Ze'ev Rivlin. A confidence measure for acoustic likelyhood scores. In *Proc. EUROSPEECH'95*, pages 523–526, Madrid, Spain, 1995.

[6] Jürgen Schürmann. *Polynomklassifikatoren.* Oldenbourg, 1977.

[7] Jürgen Schürmann. *Pattern Classification: a unified view of statistical and neural approaches.* Wiley-Interscience, 1996.

[8] W. Wahlster. Verbmobil - Translation of Face-to-Face Dialogs. In *Proc. EUROSPEECH'95*, volume "Opening and Plenary Sessions", pages 29–38, Berlin, Germany, 1993.

[9] Sheryl Young. Recognition confidence measures: Detection of misrecognitions and out-of-vocabulary words. Technical Report CMU–CS–94–157, Carnagie Mellon University, Pittsburg, USA, 1994.