ESTIMATING CONFIDENCE USING WORD LATTICES

Thomas Kemp *Thomas Schaaf*

Interactive Systems Laboratories, ILKD
University of Karlsruhe
76128 Karlsruhe, Germany

ABSTRACT

For many practical applications of speech recognition systems, it is desirable to have an estimate of confidence for each hypothesized word, i.e. to have an estimate which words of the speech recognizer's output are likely to be correct and which are not reliable.

Many of today's speech recognition systems use word lattices as a compact representation of a set of alternative hypothesis. We exploit the use of such word lattices as information sources for the measure-of-confidence tagger JANKA [1]. In experiments on spontaneous human-to-human speech data the use of word lattice related information significantly improves the tagging accuracy.

1. INTRODUCTION

Current speech recognition systems are far from perfect. Unfortunately, number and location of the errors in their output is usually unknown. However, this information could be used in a number of applications. Examples are word selection for unsupervised adaptation schemes like MLLR [5], automatic weighting of additional, non-speech knowledge sources like lip-reading, or aiding a NLP system towards generating repair dialogs in case a semantically important word has a low confidence.

Consider the sentence "Mary loves her little child" and the corresponding speech recognizer output "Eight Mary loves her brittle child". Then, the desired output of a measure of confidence (MOC) tagger would be "0.0, 1.0, 1.0, 1.0, 0.0, 1.0" where "0.0" stands for a recognition error and "1.0" for a correctly recognized word.

Previous work has shown [1] [2] [3] that the representation of alternative hypothesis, like N-best-lists or word lattices, can be used estimate word-level confidence. Many state-of-the-art speech recognition systems output their result in the form of word lattices anyway. Therefore, it would be convenient if a MOC tagger could be built on this type of output alone. In this work, we describe several different features which can be extracted from word lattices alone. The correlation of the features with the actual error rate on an independent test set is measured. In experiments carried out on spontaneous speech data we show that a high-accuracy MOC tagger can be built basing only on the word lattice. In an additional experiment, we compare the results of the purely lattice-based confidence tagger with the performance of a confidence tagging system that uses a combination of the lattice-based features with a large set of non-lattice related knowledge sources.

2. EVALUATING CONFIDENCE TAGGER

Different methods for the evaluation of confidence measuring systems have been proposed [8] [7] [10]. However, the best method for scoring depends on the application for the confidence tags. In this work, *confidence accuracy* CA, defined as

$$CA = \frac{\text{Number of correctly assigned tags}}{\text{total number of tags}} \qquad (1)$$

is used.

Another measure, which can only be used for continuously valued confidence tags, is the plot of precision (PRC) and recall (RCL) over decision threshold. PRC and RCL are defined as

$$PRC_X = \frac{\text{Number of correctly assigned tags for class X}}{\text{Number of total tags for class X}} \qquad (2)$$

$$RCL_X = \frac{\text{Number of correctly assigned tags for class X}}{\text{total number of elements in class X}} \qquad (3)$$

where $X \in \{correct, false\}$.

A single metric for confidence scores, which can be viewed as normalized cross entropy, has been proposed by NIST as

$$S = \frac{H(C) + \frac{1}{N}\left(\sum_{correct} \log_2(P_c) + \sum_{incorrect} \log_2(1 - P_c)\right)}{H(C)} \qquad (4)$$

where P_c is the output of the MOC tagger for the a-posteriori probability that word c has been correctly recognized. $H(C)$ is the base entropy $H(C) = -(p \log p + (1-p)\log(1-p))$ and p the a-priori probability that a hypothesis word is correct.

3. DERIVING KNOWLEDGE SOURCES FROM WORD LATTICES

In many applications of speech recognition it is desirable to have more than one hypothesis for a given utterance. In such cases, many existing speech recognition systems use *word lattices* as output format. Through the use of word lattices a very large number of alternative hypotheses can be stored with a small amount of memory.

In our system, a word lattice is a directed graph, where the nodes are associated with words and the links represent the possible succession of words in the different hypothesis. As the same word may have a different number of frames when followed by a different successor, the acoustic word scores must be stored in the links rather than in the nodes of the lattice.

3.1. Link probability

For a given word lattice, the probability of any link may be computed in very much the same way as in the standard forward-backward algorithm [12] for HMMs. Here, the lattice nodes can be viewed as HMM states, and the links of the lattice give the possible transitions. As the nodes are associated with the words in the hypothesis, the emission probability of a node is the acoustical score of this word at this time segment. The transition probability can be taken from the statistical language model which has been used in the decoding process. As a result of the forward-backward algorithm, the probability of each link in the lattice is available. These probabilities can be directly interpreted as a-posteriori probabilities for words (the start nodes of the links) occurring in the time segment of the link.

The plot of recognition error probability over **gamma** is shown in figure 1.

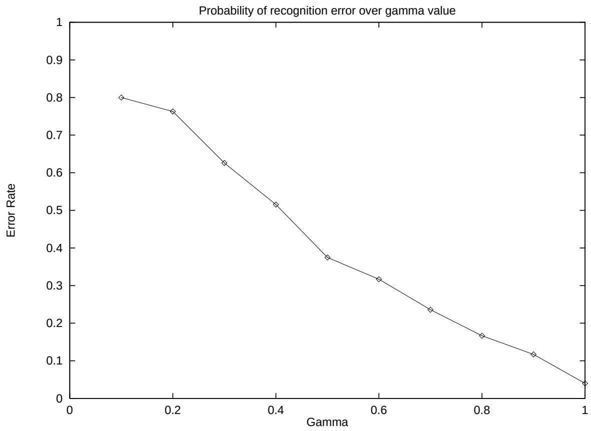

Figure 1. Error rate over feature value for feature **gamma**

3.2. Hypothesis density

In the decoder of a speech recognizer for large vocabularies, unlikely hypos must be pruned from the search space. In time segments where the probability for a word W_i is very much higher than the probability of all other words, most of the competing other words are pruned. As the word lattice can be viewed as the compact representation of the decoder search space, the number of links that span such a time segment should be low. If, on the other hand, a great number of words has similar likelihood in a time segment, no effective pruning can take place, and hence the number of links in this time segment will be high. As a high number of hypos with similar likelihood implies a higher probability of error, the number of links that span the time segment of a word in the most likely hypothesis should be strongly correlated with the word error. This number can be easily computed for each frame of an utterance. For each word, we computed three numbers of competing links: at the word beginning, at the word end, and the average number averaged over the time segment into which the word was aligned. The resulting features are named **nTa**, **nTe** and **nAverage**, respectively. To capture the effects of high or low confidence of the neighbouring words, we also computed the hypothesis density at the last frame of the predecessor word and the first frame of the successor word. This two features were named **nPre** and **nAfter**.

3.3. Acoustic stability

For this feature [6] [7], a number (typically 100) of alternative hypotheses with different weighting between acoustic scores and language model scores is computed. Each of these hypotheses is aligned against the reference output of the recognizer, where the reference output is defined as the output with the (assumedly) best weighting between acoustics and language model. For each word of the reference output, the number of times the same word occurs in the set of alternative hypotheses, normalized by the number of alternative hypotheses, is taken as feature value.

3.4. Correlation results

To exploit the usefulness of the new features, we computed the (linear) correlation of the feature values with the likelihood of a recognition error. A high correlation indicates a useful feature. As a comparison, a 'classical' feature for confidence evaluation, the normalized acoustic score per frame, is shown. The correlation coefficients are summarized in table 1.

Feature	correlation
gamma	0.520
A-stabil	0.481
nPre	-0.401
nAfter	-0.231
nTa	-0.388
nTe	-0.335
nAve	-0.377
normScore	-0.171

Table 1. Correlation coefficients to c/f tag

4. EXPERIMENTAL

4.1. Database

For all described experiments we used the GSST database, which has been collected simultaneously at four different sites in Germany. It consists of high-quality recordings of human-to-human spontaneous German dialogs in the appointment scheduling domain, i.e. two persons try to schedule a meeting within the next month.
A more detailed description of the database is given in [1].

For the evaluation of the word lattice features described in this paper, we used a subset of 1251 utterances from the GSST database. None of the speakers of this subset was used for the training of the acoustic models and the language model of the recognizer. The subset data was divided into a training, crossvalidation and test set. Table 2 shows the composition of the subset of the database used for training and evaluation of the measure of confidence classificator.

set	speakers	utterances	words	duration (min)
Training	46	785	14906	101
Crossvalid.	6	134	3063	22
Test	20	332	5940	39
Total	72	1251	23909	162

Table 2. Database composition

4.2. The JANUS-3 system

The speech-to-speech translation system JANUS-3 [9] is a joint effort of the Interactive Systems Labs at Carnegie Mellon University, Pittsburgh, and the University of Karlsruhe, Germany. The baseline speech recognition component of JANUS-3 uses mixture-gaussian, continuous density HMMs with a scalable amount of parameter tying as acoustic model. A standard statistical trigram-backoff language model is used. In the preprocessing stage, mel-cepstral LDA-transformed coefficients are computed with a frame rate of 10 ms. After the initial recognition run, vocal tract length normalization [4] is employed.

The JANUS-3 decoder achieved a word error rate of 13.2% in the 1996 VERBMOBIL evaluation. This was the lowest error rate of the five participating institutions.

In the experiments described, we evaluated the system that was used for the required test of the 1996 VERBMOBIL evaluation. The baseline confidence accuracy on the MOC test set, when tagging all words with 'correct', was 85.3%.

A detailed description of the JANUS-3 recognizer can be found in [1] [4].

5. RESULTS

5.1. Evaluation of the new features

To evaluate the performance of the six new features (**nTa**, **nTe**, **nAve**, **nPre**, **nAfter** and the forward-backward probability **gamma**), we built a set of linear classifiers basing on different combinations of the input features. The linear classifiers made use of an LDA transformation based on the classes [correctly recognized, recognition error]. A more detailed description is given in [1].

The baseline confidence accuracy CA on the evaluation set was 85.3%. The results are summarized in table 3. As a comparison, the result achieved with 11 not lattice related features [1] is given.

The classifier relying solely on the output lattice performed very well in comparison to a classifier that made use of the full set of 18 features.

5.2. Classifier design

The transformation based approach described in the previous section works well for linearly separable classes. However, on many data sets it does not yield satisfying results. Therefore, we compared two additional classifiers with the linear classifier: a 3-layer neural network (described in detail in [1]) and a decision tree based classifier, as described in [11]. The results are summarized in table 4. As can be seen, the neural net classifier yields slightly better results, than the linear classifier.

Features	CA	error reduction
baseline	85.3%	-
gamma alone	88.0%	18.3%
nTa + nTe + nAve	87.5%	14.9%
plus nPre and nAfter	87.9%	17.7%
plus gamma	88.4%	21.1%
plus acoustic stability	88.9%	24.5%
acoustic stability alone	87.4%	14.3%
11 non-lattice features	87.3%	13.6%
all features combined	90.0%	29.9%

Table 3. Performance of different feature sets

Features	linear	tree	neural net
lattice only	88.9%	88.5%	88.9%
all	90.0%	89.6%	90.1%

Table 4. Performance of different classifiers

5.3. Adding contextual information

It has been shown [1], that the use of contextual information, i.e. the neighbouring words, improves recognition performance. Therefore, we added the feature vector of the left and the right neighbour of each word to the input of the neural net. As some of the lattice related features contain contextual information, the additional gain of the context is expected to be smaller for the lattice based classifier than for the full system. The result is shown in table 5.

Features	context	CA	S
lattice only	no	88.9%	0.326
lattice only	yes	89.1%	0.340
all	no	90.1%	0.398
all	yes	90.6%	0.416

Table 5. Influence of contextual information

The result in terms of PRC and RCL are shown in figure 2. For a recall rate of 90%, i.e. 90% of the correctly recognized words are spotted as such, a precision of more than 95% can be achieved.

6. SUMMARY

We have shown, that the word lattice that is the output of many speech recognition systems, contains useful information which allows to estimate the likelihood of a misrecognition of every word of the recognizer's output. The performance of a confidence tagger which relied solely on the lattice was higher than that of a classical approach using 11 non-lattice related features [1], which included normalized word scores, language model backoffs, word lengths, speaking rate, and others.

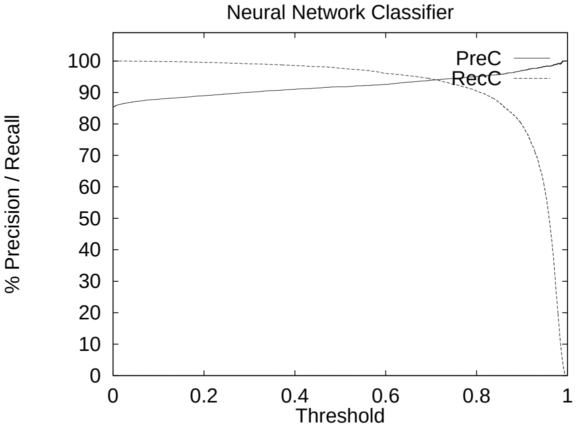

Figure 2. Precision and recall of best system

7. ACKNOWLEDGEMENTS

This research was partly funded by grant 01 IV 701 U7 from the German Ministry of Science and Technology (BMBF) as a part of the VERBMOBIL project. The JANUS project was supported in part by the Advanced Research Project Agency and the US Department of Defense. The views and conclusions contained in this document are those of the authors.

The authors wish to thank all members of the Interactive Systems Labs, especially Michael Finke, for useful discussions and active support.

REFERENCES

[1] T. Schaaf, T. Kemp: *Confidence measures for spontaneous speech recognition*, in Proc. ICASSP 1997, Vol 2, pp. 875 ff, Munich, April 1997

[2] L. Gillick, Y. Ito, J. Young: *A probabilistic approach to confidence estimation and evaluation*, in Proc. ICASSP 1997, Vol 2, pp. 879 ff, Munich, April 1997

[3] M. Weintraub, F. Beaufays, Z. Rivlin, Y. Konig, A. Stolcke: *Neural-network based measures of confidence for word recognition*, in Proc. ICASSP 1997, Vol 2, pp. 887 ff, Munich, April 1997

[4] M. Finke, P. Geutner, H. Hild, T. Kemp, K. Ries, M. Westphal: *The Karlsruhe-Verbmobil speech recognition engine*, in Proc. ICASSP 1997, Vol 1, pp. 83 ff, Munich, April 1997

[5] C.J. Legetter, P.C. Woodland: *Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models*, Computer Speech and Language **9** (1995), 171-185

[6] M. Finke, T. Zeppenfeld, M. Maier, L. Mayfield, K. Ries, P. Zhan, J. Lafferty, A. Waibel: *Switchboard April 1996 Evaluation Report*, DARPA, April 1996

[7] Haitao Qiu: *Confidence Measure for Speech Recognition Systems*, Masters Thesis, Carnegie Mellon University Computational Linguistics Philosophy Department, Pittsburgh, PA, April 1996

[8] S. Cox, R. Rose: *Confidence Measures for the Switchboard Database*, in Proc. ICASSP-96, pp 511 ff, Atlanta, May 1996, ISBN 0-7803-3192-3

[9] M. Woszczyna, M.Finke, D.Gates, M.Gavalda, T.Kemp, A.Lavie, A.McNair, L.Mayfield, M.Maier, I.Rogina, K.Shima, T.Sloboda, A.Waibel, P.Zhan, T.Zeppenfeld: *Janus II - advances in spontaneous speech translation*, in Proc. ICASSP-96, pp 409 ff, Atlanta, May 1996, ISBN 0-7803-3192-3

[10] Sheryl Young: *Detecting misrecognitions and out-of-vocabulary words*, in Proc. ICASSP-94, pp. II-21 ff., Adelaide, Australia, April 1994

[11] J. R. Quinlan: *C4.5: programs for machine learning*, Morgan Kaufman publishers Inc, San Mateo, CA, ISBN 1-55860-238-0 (1993)

[12] L. Rabiner: *A tutorial on Hidden Markov Models and selected applications in speech recognition*, in Readings in Speech Recognition, Kai-Fu Lee and Alex Waibel (edts), pp 267 ff, Morgan Kaufman Publishers, ISBN 1-55860-124-4