OBTAINING CONFIDENCE MEASURES FROM SENTENCE PROBABILITIES

Bernhard Rueber

Philips GmbH Forschungslaboratorien Aachen, P.O. Box 50 01 45, D-52085 Aachen, Germany
E-mail: rueber@pfa.research.philips.com

ABSTRACT

The paper addresses the issue whether the "probabilities" delivered by a speech recognizer can be directly used as a measure for the confidence of the recognition. As current recognizers have to commit a lot of modelling assumptions and because of estimation problems due to sparse data this certainly is questionable. Nevertheless, this investigation shows, in the framework of recognizing semantic items in the Philips automatic telephone exchange board system PADIS, that there exists a useful correlation between probabilities and confidences.
The method proposed works out as a generalization of the more standard method of using likelihood ratios between the first- and second-best recognition path. It offers as distinct advantages a) the integration of all available knowledge sources, and b) the direct and theoretically sound computation of confidence measures on all levels of interest.

1 INTRODUCTION

Confidence measures basically serve a twofold purpose: They are most often used for rejecting complete (supposedly out-of-domain) utterances (so-called utterance verification) [1, 2, 3]. Alternatively, they may be employed for lowering the error rate at least for substitutions and insertions (of supposedly misrecognized or out-of-vocabulary (OOV) words) [2, 4, 5].

In the context of an automatic inquiry system, confidence measures may further be employed to control the amount and manner of verification needed for a successful dialogue. As a simple example, consider the case of a train-timetable inquiry system. If there is enough confidence in the query's parameters there is no need for an explicit confirmation question. The system may just present the connection information to the user, waiting for an interruption (barge-in), in case it misunderstood. Such a kind of dialogue resembles much more human behaviour than today's IVR schemes.

The paper is organized as follows: After reviewing previous literature work on confidence measures, the approach proposed in this paper is presented in sections 3–5. This includes experimental results on the task of verifying semantic items in the Philips automatic telephone exchange board system PADIS [6]. Section 6 concludes with outlining possible directions for future work. Details on the speech recognizer, the application, and the corpora used to test this approach are given in the appendices.

2 REVIEW OF PREVIOUS WORK

Of course, to give a detailled literature survey is out of scope for a conference paper (but see [7] for a fairly complete account on keyword detection and modelling of out-of-vocabulary words). Instead, we concentrate on the main lines.

One may separate the task of finding a confidence measure in a) the search for useful features and b) the way of transforming these features into a confidence measure. In literature, a large set of features was tried [8], the most useful of which are probably the number and scores of the competing hypotheses during search [1, 4, 2, 3]. Several authors examined (discriminatively) trained verification acoustic models [9, 2] while others argued not to split the training set but to use what they termed "strictly lexical fillers" [5].

The simplest way to transform the features into a confidence measure is either to do some simple thresholding [4, 10] or to do a statistical hypothesis test by calculating the likelihood ratio between the recognized item and its alternative, followed by thresholding [1, 9]. More recently, some groups have posed the problem of calculating a confidence measure in the framework of classification, i.e. they trained a classifier on the input features to arrive at an optimal reliable/unrealiable decision. Popular classifiers have been linear discriminant analysis (LDA) [2, 8, 3], the multi-layer perceptron (MLP) [8], or a simple Bayesian classifier [1].

Another new interesting approach is to directly model the statistical correlation between the input features and the likelihood of an item to be correctly recognized. Gillick et al. in [11] use generalized linear regression to predict correctness of recognition from the values of the input features.

3 MAIN POINTS OF THIS APPROACH

The contribution of this paper may be viewed in some sense as an extension of the n-best (more precisely: second-best) approach of [2]. But there are also strong

relations to [5] as only lexical entries are used as alternatives, and to [11] in the strive to directly model the correctness likelihood of a recognized item.

Instead of working with different kinds of duration normalization [10] or confidence measure combinations [9], we directly use the recognizer probabilities in a way as close to theory as possible: The probabilities delivered by the recognizer for a (fairly large) n-best list (resp. a word graph) are directly re-normalized to sum up to 1.

This approach gives us the following advantages:

1. We naturally make use of the full knowledge sources exploited for recognition.
2. Confidence measures are obtainable on all levels. E.g, we get a direct confidence measure for each recognized semantic item while still using the most complex constraints, which are only applicable for the complete sentences.
3. No additional filler models are employed. Instead, computing the confidence measures is a simple post-processing on the n-best list (or the word graph) being computationally cheap as compared to the recognition step.

4 PROBABILITY INTERPRETATION OF THE SCORES OF N-BEST SENTENCES

From a theoretical point of view the (re-normalized) probabilities obtained from the (suitably scaled) sentence scores delivered by the recognizer should be a direct measure for the reliability of the recognition. In this section, the experiments needed to assess the practical relevance of this view are presented.

In doing that, the following procedure was applied:

1. A probability for each recognized semantic item, i.e. attribute, is computed as follows:
   (a) Scale the recognizer scores of the n-best sentences, which theoretically should correspond to the negative logarithms of the probabilities, by a suitable scaling factor:
   $$sc_{\text{scaled}} = \alpha \cdot sc. \tag{1}$$
   (b) Transform the scores for the sentences s into probabilities by re-normalizing the corresponding exponentials:
   $$p_s = \beta \cdot \exp(-\alpha \cdot sc), \text{ such that } \sum_{s=1}^{n} p_s = 1. \tag{2}$$
   (c) Compute the probability for all attributes a contained in the first-best sentence by summing over the probabilities of all sentences s containing a:
   $$p_a = \sum_{s=1}^{n} p_s \cdot \delta_{a,s}, \tag{3}$$
   with $\delta_{a,s}$ being 1 for all sentences containing a, and 0 else.
2. All attributes with similar probability values are clustered in corresponding sets, i.e. a suitable probability histogram is constructed.

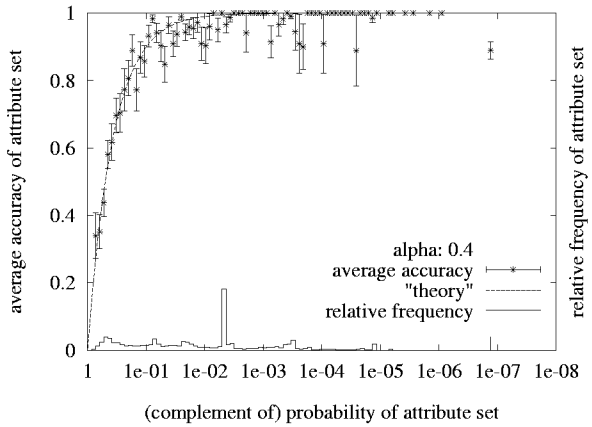

Figure 1: Accuracy distribution (for $\alpha = 0.4$ on test2). ($1-$"probability of attribute set" is plotted on the x-axis. But the orientation is reversed such that the probability increases to the right (which appears to be more natural).)

3. For each attribute in each such set its correctness is assessed, i.e. it is examined whether this attribute was actually spoken. By this procedure an average accuracy or correctness for each attribute set is established.

If the theoretical claim of a direct correlation between scores and reliability of recognition would hold, the above computed attribute probabilities should more or less match the average attribute set accuracy.

Figure 1 shows the accuracy distribution for scaling factor $\alpha = 0.4$ on a half-logarithmic scale. To allow comparison to the theoretical claim that the recognizer probability should directly correspond to the confidence level the corresponding line "theory" is also included in the diagram. Furthermore, an indication for the statistical significance of the correctness data points is given by plotting some error bars.[1]

Considering also the second curve in figure 1, it can be seen that in the regions where the most attribute sets occur, the correlation between probability and accuracy is acceptable. This indicates that probabilities can be successfully used as a direct measure for the confidence of the recognition. [image A0091G01.GIF]

5 MINIMUM PROBABILITY

The probability interpretation of the preceding section can be used for classifying the attributes as reliable resp. unreliable by a simple comparison to a minimum required probability threshold. Figure 2 shows recognition accuracies versus false-alarm rates in dependence on the number n of paths considered in the n-best list. Here, especially the case $n = 2$ is interesting as this reduces to the more standard method of using likelihood ratios between the first- and second-best recognition path [2].

[1] A simple binomial distribution (with symmetrical confidence intervals) for the correctness is assumed for that, i.e. the half width of the error bar is assumed to be $\sqrt{c(1-c)/N}$ (c: correctness value, N: number of samples falling in the attribute set delivering correctness value c).

n	corp.	p_b	$\exp(L_b)$	$\exp(L)$	S^a	${}^0\overline{\text{ra}}^{10}$
40	test1	0.80	0.60	0.71	0.33±3	65
2				0.70	0.30±3	57
40	test2	0.83	0.63	0.75	0.38±3	67
2				0.73	0.31±4	58

[a] All standard deviations are equal to ±1 in the last digit except where explicitly stated.

Table 1: Figures of merit: Influence of number n of paths in n-best list (for $\alpha = 0.4$):

p_b**:** baseline, i.e. average, probability that a recognized item is correct [11],

$\exp(L_b)$**:** exponent of (negative) baseline entropy (cf.b.),

$\exp(L)$**:** exponent of (negative) entropy attained by the proposed method [11]:

$L = \frac{1}{M}\sum_{i=1}^{M}[c_i \log(p_i) + (1 - c_i)\log(1 - p_i)]$,

$c_i = 1$ or 0 according to whether attribute i is correctly recognized or not,

p_i**:** confidence level of attribute i,

M**:** total number of attributes,

S**:** NIST normalized cross entropy [8]:
$S = ((-L_b) - (-L))/(-L_b)$,

${}^0\overline{\text{ra}}^{10}$**:** recognition accuracy averaged over false-alarm rates between 0 and 10 [12].

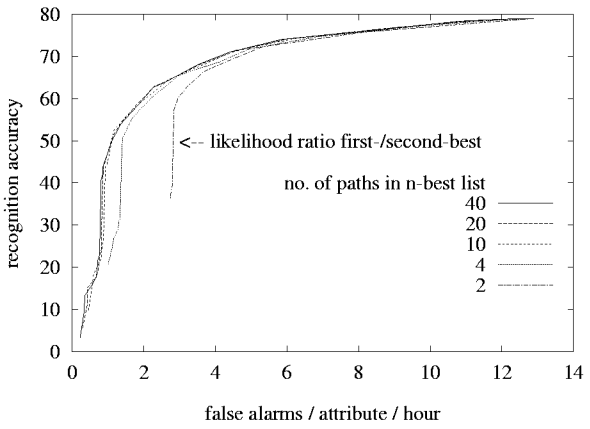

Figure 2: "Receiver Operating Characteristics" (ROC): Influence of number n of paths in n-best list (for $\alpha = 0.4$ on test1)
(Please note that $n = 2$ corresponds to the standard "second-best method" [2].)

The curves in figure 2 more or less coincide for high accuracies which is a consequence of the following: High accuracies correspond to high probability values which in turn requires that the attribute occurs in nearly all of the first paths in the n-best list. Therefore, not considering the high order (i.e. low probability) paths of the n-best list does not matter. On the contrary, clear differences show up in the low accuracy portions of the curves as there the contribution of the low probability paths is crucial. Thus, in order to obtain low false-alarm rates at still acceptable accuracies more than 2 paths are required. This is also obvious from looking at the figures of merit collected in table 1. [image A0091G02.GIF]

6 CONCLUSIONS AND OUTLOOK

We demonstrated that the probabilities delivered by the recognizer can be taken as a direct measure for the confidence of recognition. This can be considered as a theoretically sound generalization of the well-known likelihood ratio approach using the first- and second-best in an n-best list [2]. However, the way the method is developed here, it offers distinct advantages, the main of which are a) integrating all (even long-range) knowledge sources, and b) obtaining confidence measures directly on the level of the semantic items.

Natural extensions of the proposed method are: a) using non-lexical filler models to enhance the space of alternative hypotheses especially for out-of-vocabulary (OOV) words, and b) taking the recognizer probabilities as one of several features for a subsequent confidence classifier [8] or resp. a regression computation [11] (cf. section 2). Incorporating non-lexical fillers will require a careful language modelling for the OOV words (cf. [5]).

7 ACKNOWLEDGEMENTS

The author wishes to thank all members of the Philips speech recognition group for sustaining our challenging research environment. Special thanks to Andreas Kellner and Frank Seide for their uncountable hours of work building up the PADIS system, and to them and Volker Steinbiss for the stimulating discussions on the subject of confidence measures.

8 REFERENCES

1. S. Cox and R. C. Rose, "Confidence Measures for the Switchboard Database", *Proc. ICASSP*, vol. I, pp. 511–514, Atlanta, GA, May 1996.
2. A. R. Setlur, R. A. Sukkar, and J. Jacob, "Correcting Recognition Errors via Discriminative Utterance Verification", *Proc. ICSLP*, vol. II, pp. 602–605, Philadelphia, PA, Oct. 1996.
3. J. Caminero, L. Hernandez, C. de la Torre, and C. Martín, "Improving Utterance Verification Using Hierarchical Confidence Measures in Continuous Natural Numbers Recognition", *Proc. ICASSP*, vol. II, pp. 891–894, Munich, Germany, Apr. 1997.
4. R. Lacouture and Y. Normandin, "Detection of Ambiguous Portions of Signal Corresponding to OOV Words or Misrecognized Portions of Input", *Proc. ICSLP*, vol. IV, pp. 2071–2074, Philadelphia, PA, Oct. 1996.
5. R. E. Méliani and D. O'Shaughnessy, "Accurate Keyword Spotting Using Strictly Lexical Fillers", *Proc. ICASSP*, vol. II, pp. 907–910, Munich, Germany, Apr. 1997.
6. A. Kellner, B. Rueber, F. Seide, and B.-H. Tran, "PADIS – An Automatic Telephone Switchboard and Directory Information System", *Speech Communication*, to appear.
7. R. C. Rose, "Keyword Detection in Conversational Speech Utterances Using Hidden Markov Model Based Continuous Speech Recognition", *Computer Speech and Language*, 9:309–333, 1995.

language	German
bandwidth	300–3400 Hz (telephone)
features	12 MFCC + deltas
acoustic models	4461 strongly tied (intra-word) triphones sharing 703 states
acoustic training	on 12h speech from the train-timetable inquiry task
language models	mixed class and word bigrams
search	Viterbi approximation
interface to speech understanding	word graphs [14]

Table 2: Recognizer characteristics.

8. T. Schaaf and T. Kemp, "Confidence Measures for Spontaneous Speech Recognition", *Proc. ICASSP*, vol. II, pp. 875–878, Munich, Germany, Apr. 1997.
9. E. Lleida and R. C. Rose, "Likelihood Ratio Decoding and Confidence Measures for Continuous Speech Recognition", *Proc. ICSLP*, vol. I, pp. 478–481, Philadelphia, PA, Oct. 1996.
10. Z. Rivlin, M. Cohen, V. Abrash, and T. Chung, "A Phone-Dependent Confidence Measure for Utterance Rejection", *Proc. ICASSP*, vol. I, pp. 515–517, Atlanta, GA, May 1996.
11. L. Gillick, Y. Ito, and J. Young, "A Probabilistic Approach to Confidence Estimation and Evaluation", *Proc. ICASSP*, vol. II, pp. 879–882, Munich, Germany, Apr. 1997.
12. A. S. Manos and V. W. Zue, "A Segment-Based Wordspotter Using Phonetic Filler Models", *Proc. ICASSP*, vol. II, pp. 899–902, Munich, Germany, Apr. 1997.
13. H. Ney, V. Steinbiss, X. Aubert, and R. Haeb-Umbach, "Progress in Large–Vocabulary, Continuous Speech Recognition", *Proc. in Artificial Intelligence, Progress and Prospects of Speech Research and Technology*, pp. 75–92, Munich, Germany, Sep. 1994.
14. M. Oerder and H. Ney, "Word Graphs: An Efficient Interface between Continuous-Speech Recognition and Language Understanding", *Proc. ICASSP*, vol. II, pp. 119–122, Minneapolis, April 1993.

App. A RECOGNIZER CHARACTERISTICS

A (more or less) state-of-the-art HMM recognizer is employed whose main data can be found in table 2. For details, the interested reader is referred to [13, 6].

App. B THE PADIS APPLICATION

The application domain considered in this paper is the Philips automatic telephone exchange board system PADIS [6]. This is a system providing our laboratories with the possibility to get directory information and direct call completion by using natural specification of names, titles, functions, or similar. Of course, the most common use of the service is name dialling (i.e. the auto attendant functionality).

A typical dialogue is presented in figure 3. For more details, the interested reader is referred to [6].

System:	*Hello, this is PADIS. How can I help you?*
Caller:	*I'd like to talk to Mike.*
System:	*So, you want to talk to Mike Anderson?*
Caller:	*Yes, please.*
System:	*The extension 506 is busy right now. Can I do anything else for you?*
Caller:	*Oh yes. Can you give me the email address of Mary, please?*
System:	*Which Mary do you want the email address of?*
Caller:	*Uh, she is in the speech recognition group.*
System:	*So you want the email address of Mary White from the speech recognition group?*
Caller:	*Exactly.*
System:	*The email address is ...*

Figure 3: Example dialogue (translated).

word category	# words	percentage
last names	687	38%
first names	304	17%
genders	2	0%
groups	121	7%
sites	4	0%
titles	2	0%
$\sum$ database terminals	1120	62%
grammar words	139	8%
filler words	534	30%
total	1793	100%

Table 3: Composition of vocabulary.

	test1	test2	train
no. of turns	2496	2495	3381
no. of words	6899	6571	10922
no. of attributes	4183	4115	6659
approx. duration	1h	1h	2h
no. of out-of-vocabulary words	149 (2.2%)	147 (2.2%)	—
perplexity	13.2	12.3	—
best path AERs[a]	29.93%	27.78%	—
CG[b] AERs[a,c]	6.60%	7.75%	—
CG[b] density	33.4	31.3	—

[a] AER: attribute error rate, i.e. error rate of semantic items.
[b] CG: concept graph, i.e. ("word") graph after syntactic parsing [6].
[c] The numbers reported here are (close) upper bounds of the true attribute error rates of the concept graphs.

Table 4: Characteristics of the PADIS corpora and recognizer performance.

App. C LEXICON AND CORPORA OF PADIS

The telephone directory database underlying the PADIS application consists of 592 entries. Table 3 gives the lexicon composition and table 4 the sizes of the corpora used for training the recognizer and testing the confidence measures.