COMBINATION OF CONFIDENCE MEASURES FOR PHRASES

Bernd Souvignier and Andreas Wendemuth

Philips Research Laboratories
Weisshausstr. 2, D-52066 Aachen, Germany
e-mail: {souvi,wendemu}@pfa.research.philips.com

ABSTRACT

Since automatic speech recognition is error prone it is highly desirable to obtain information on the reliability of a recognition result. For many applications it is interesting to know how well a sequence of words rather than a single word was understood, for example a credit card number or a phrase specifying a time or location.

In this paper we present different approaches to confidence measures for phrases. In a first step we demonstrate that combining multiple confidence features on word level gives a confidence measure superior to all single features. We then compare different methods to combine the confidence values for the words in a phrase to a compound confidence measure on phrase level. As an alternative approach we investigate an N-best list based confidence measure that can be directly applied to phrases.

Finally, experimental data gives evidence that combining the different approaches gives significantly higher performance than each approach taken singularly.

1. INTRODUCTION

Phrases are of interest in automatic speech recognition in various scenarios. For example, spoken digits can be compounded into phrases as well as regular expressions in dialogue turns such as the departure time information in a timetable request: “tomorrow morning”, “between three and five p.m.”. Typically, such phrases are concepts defined in a context free grammar. They are of variable length and can contain words from a large vocabulary under given (grammatical, language model) restrictions.

Why would one apply confidence measures in the context of phrases? Confidence measures will be used to specify a point of operation in the receiver-operating-characteristics (ROC). For example, an accepted level of phrase recognition errors can be pre-specified, and the corresponding rejection threshold of the confidence measure is then chosen accordingly. In this sense, confidence measures allow an application-specific tuning of accuracy and robustness.

The standard confidence measures available to us are computed on the word level. Many of them depend on singular words only and can be computed during recognition. Others (e.g. the “word-graph” measure) are a-posteriori measures which can only be computed after the full utterance has been recognized.

In this paper we present methods which judge the reliability of a phrase by processing word-based confidence measures. We complement the word-based approach by an N-best list based method that can be directly applied to phrases and finally combine the two approaches into a unique confidence measure.

2. COMBINATION OF CONFIDENCE FEATURES ON WORD LEVEL

Given the set X of “raw” confidence features on word level, can we define a classifier for utterance verification $f(X)$ and a threshold τ such that the recognized word will be classified as unreliable (rejection) if $f(X) \leq \tau$ and as reliable (acceptance) otherwise?

We subject $f()$ to a number of restrictions to facilitate this task. Since we want to find the form of $f()$ from training data with known classification, we choose a parametrization $f_{\underline{\mathbf{J}}}()$ with parameters $\underline{\mathbf{J}}$. The form of the parametrization must be as simple as possible to remain easily adaptable, yet cover the complexity of the problem sufficiently. It shows both from theoretical considerations and experiments (see [6]) that a *linear* parametrization is sufficient, thus $f(X) = \underline{\mathbf{J}} \cdot X$. The *weights* comprised in the vector $\underline{\mathbf{J}}$ must be found from the training data through minimization of a suitable cost function. An obvious approach, chosen for example in [2] or [5], is to apply linear discriminant analysis. We found that a two-step optimization is favourable: the weights are first adapted to minimize the *cross-entropy*, and these weights are then fine-tuned to minimize the *Gardner–Derrida* cost function measuring the number of incorrectly classified data. For details see [6]. The threshold τ is optimized together with the weights in a fully automated, efficient procedure, supplying us with a good combined confidence measure on *word level*.

3. COMPOUND CONFIDENCE MEASURES

How do we now treat phrases of variable length? In case of simple concatenation of words, a varying length of the input vector X to $f()$ must be handled. This would not be possible in a direct way but e.g. by means of recurrent procedures where the individual words are subsequently presented to $f()$.

This approach is disfavourable for computational as well as theoretical reasons. The latter are: the probability of observing a chain of words is, strictly under the assumption of stochastic independence, a product of the individual observation probabilities. If we treat $f()$ as a probability, we must achieve promising results by postprocessing the individual word's combined confidence measure into a *compound confidence measure* for the full phrase.

We can arrive at a compound confidence measure by postprocessing the word level or by going into subword or other levels ([3]). We will not follow the latter approach here.

The assumption of stochastic independence of course not being strictly valid, we can operate a number of other postprocessings. The geometric and arithmetic means of the individual word's combined confidence measures are promising as well as the minimum combined confidence measure. The latter is motivated by the triggering of misrecognitions in the neighbourhood of a wrongly recognized word. As a compromise between taking the arithmetic mean and the minimum we also looked at the average of the confidence values below some threshold.

4. N-BEST LIST BASED METHOD

An alternative to the methods based on the acoustic confidence measures is the N-best list based method introduced by B. Rueber in [4]. Given an N-best list of sentence hypotheses with scores sc_i one computes the confidence for some pattern A (e.g. a word sequence or an attribute) by accumulating the a-posteriori probabilities of the sentence hypotheses in which A occurs. Let I be the set of indices of those sentences in the N-best list that contain the pattern A. Then the confidence for A is defined as

$$C(A) := \frac{\sum_{i \in I} exp(-\alpha \cdot sc_i)}{\sum_{i=1}^{N} exp(-\alpha \cdot sc_i)}.$$

The scaling factor α distributes the probability mass over the N-best list and can be regarded as a tuneable parameter.

It is obvious that this approach can be directly applied to phrases and it will be compared with the compound confidence measures obtained from the acoustic confidence features. Moreover, following the philosophy of Section 2 we will also look at the linear combination of the N-best list based and the compound confidence measures.

5. EXPERIMENTAL RESULTS

5.1. Evaluation criteria

The performance of confidence measures will be assessed by inspecting ROC-curves. An ROC-curve is parametrized by a confidence threshold where the x-value for a point on the curve is the number of false acceptances for a certain threshold and the y-value is the number of correct acceptances. An ideal ROC-curve rises vertically from the origin up to the final y-value and then horizontally to its ending point. Confidence measures are the better the more their ROC-curve approaches this ideal curve (i.e. the upper left corner of the coordinate system).

In addition we also compared the recall rates at two specific points of operation P_1, P_2 on the ROC-curves which we defined as follows:

P_1: RCL(c) = 95%: at least 95% of the correctly recognized items have to be accepted and the recall rates of the incorrectly recognized items are compared

P_2: RCL(f) = 90%: at least 90% of the incorrectly recognized items have to be rejected and the recall rates of the correctly recognized items are compared

5.2. Corpora

We applied the different methods to two different task:

- digit strings from the SIETILL corpus (German digit strings recorded over telephone lines)
- `complete_time` concepts from the railway timetable information system TABA (cf. [1])

For both domains we created a training corpus to obtain the combination weights and a test corpus. The statistics for these corpora are given in Table 1.

corpus	#utterances	#words
SIETILL-test	11877	35631
SIETILL-train	900	4593
TABA-test	3671	11079
TABA-train	773	3709

Table 1: Corpus statistics

For the SIETILL task we use 22 gender-specific whole word models and a network accepting only strings consisting either fully of female models or fully of male models. In the TABA application the recognized sentences are parsed by a context free grammar to extract the `complete_time` phrases.

5.3. Word Level

Eight standard confidence features on word level were applied (for a description see e.g. [5], [6]). They are: two_best (tb), n_averaged_best (nab), word_end_frequency (wef), n_best_active_states (nbas), average_acoustic (aa), word_graph (wg) (see [7]), active_state_count (asc), speaking_rate (sr).

Firstly, we will compare the individual confidence features with the combined measure on word level.

Figure 1 shows the ROC-curves for the individual confidence features as well as that for the combined confidence measure on the TABA-test corpus. The corresponding results on the SIETILL corpus are not displayed since they are very similar.

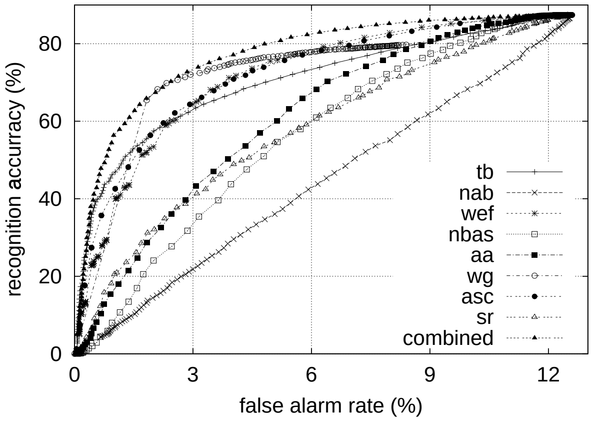

Figure 1: ROC-curves for individual and combined confidence measures on word level on TABA

As a result, combining all eight confidence features is the method of choice, since it produces an ROC-curve enveloping all the single feature curves. We want to stress that one can win a lot from features that have a poor performance when taken singularly.

5.4. Phrase Level

In the next step we combine the confidence values of single words to a confidence value for a phrase. Several methods are suggested and compared in [3] which we took as an inspiration for our considerations. In the SIETILL task we want to reject digit strings that contain an error, hence we aim at a confidence for the full strings. Since the test corpus contains only digit strings of the same length, we restrict ourselves to the following methods, where c_i are the confidence values for the n words in the phrase and C is the resulting value for the phrase:

(1) minimum: $C := \min_{i=1\ldots n} c_i$

(2) sum: $C := \sum_{i=1}^{n} c_i$

(3) product: $C := \prod_{i=1}^{n} 1/(1 - e^{-c_i})$

(4) cut_mean: $C := (\sum_{c_i<0} c_i)/(\sum_{c_i<0} 1)$

Figure 2 shows ROC-curves for the different methods. Included is also the ROC-curve obtained by applying the N-best list based confidence measure.

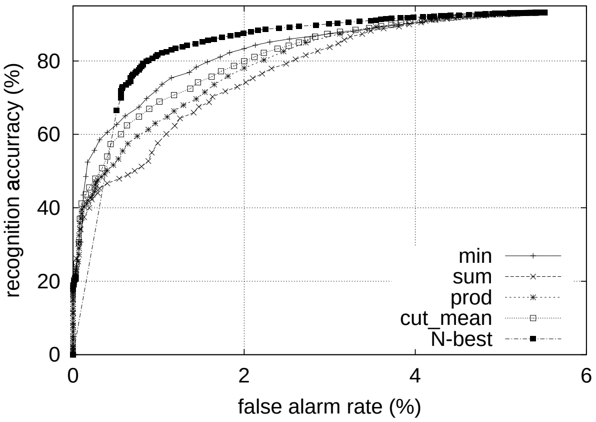

Figure 2: ROC-curves for different compound and the N-best confidence measures for digit strings on SIETILL

One clearly sees, that using the minimum outperforms the other compound confidence measures, which is a quite satisfactory result, as it agrees with the assumption that recognition errors trigger errors in their neighbourhood. However, the N-best list method performs much better than all the compound methods which is not surprising, since it exploits the a-posteriori probabilities of the full digit strings.

In the TABA application the crucial question is whether the information derived from a phrase is correct, in the case of the `complete_time` concept these are a start time and an end time for a time interval. Therefore, in this case the confidence of the phrase is used to either accept or reject the information items (attributes). Since the lengths of the `complete_time` phrases vary, we additionally investigated the following averaging methods:

(5) arith_mean: $C := 1/n \cdot \sum_{i=1}^{n} c_i$

(6) geom_mean: $C := (\prod_{i=1}^{n} 1/(1 - e^{-c_i}))^{1/n}$

They turned out to be comparable to the sum and product methods as can be seen from Table 2 in which we give the recall rates for the chosen points of operation as defined in Section 5.1.

From Table 2 one concludes that again using the minimum gives the best compound confidence measure and that the average of the negative confidence values is the follower-up. In contrast to the SIETILL task, the N-best list based method is not better than the best compound method, but it still has similar performance.

Method	RCL(f) at P_1	RCL(c) at P_2
minimum	60.06%	77.34%
sum	44.13%	61.16%
arith_mean	39.94%	63.24%
product	47.77%	71.60%
geom_mean	46.51%	71.56%
cut_mean	48.32%	75.04%
N-best	54.27%	77.20%

Table 2: Recall rates for different compound and the N-best confidence measures for phrases on TABA

5.5. Combination of the different approaches

In a final experiment we combined the N-best list based approach with the best compound confidence measures, again using the combination method described in Section 2.

We were able to improve the results obtained so far significantly, as can be concluded from Table 3 and Figure 3. Note that on the SIETILL task an improvement is achieved although the N-best list method by far outperforms the compound confidence measure, whereas in the TABA application the single approaches perform comparably but their combination is highly superior, giving recall rates of $RCL(f) = 66.29\%$ at P_1 and of $RCL(c) = 82.35\%$ at P_2 (which are to be compared with the values in Table 2).

Method	RCL(f) at P_1	RCL(c) at P_2
compound	49.94%	70.96%
N-best	65.51%	80.41%
compound + N-best	68.73%	84.84%

Table 3: Recall rates for the different approaches and their combination on SIETILL

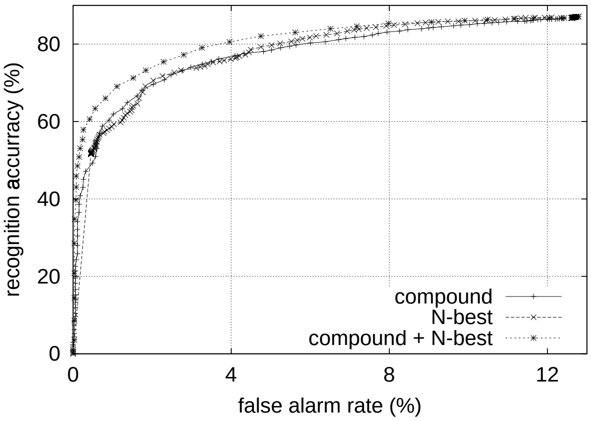

Figure 3: ROC-curves for the different approaches and their combination on TABA

6. DISCUSSION

We demonstrated how individual confidence measures can be combined on the word and phrase level. Linear combination gave satisfactory results and could efficiently be optimized. The resulting confidence measure for phrases can very favourably be used to improve receiver-operating-characteristics in both SIETILL and TABA tasks.

We could show that the best characteristics for a compound confidence measure are obtained by **(a)** combining all confidence features on the word level and **(b)** choosing the minimum of these combined confidence measures on the phrase level. The latter confirms the (pessimistic) theoretical assumption that misrecognitions trigger their environment and that phrase probabilities behave like channel pipelines, depending on the worst channel (word).

We also found that using the N-best list based method from [4] gives very satisfactory results when applied to phrases.

We finally showed that the best overall performance is achieved by combining the strengths of the different approaches into a single confidence measure for phrases by applying the combination method from [6] to the N-best list based and the minimum compound confidence measures.

7. REFERENCES

[1] Aust, H., Oerder, M., Seide, F., and Steinbiss, V. “The Philips automatic train timetable information system”, *Speech Communication*, vol. 17, pp. 249–262, 1995.

[2] Caminero, J., de la Torre, C., Villarrubia, L., Martín, C., and Hernández, L. “On-line garbage modelling with discriminant analysis for utterance verification”, *ICSLP 1996*, Philadelphia, USA, pp. 2111–2114.

[3] Kawahara, T., Lee, C.-H., and Juang, B.-H. “Flexible speech understanding based on combined key-phrase detection and verification”, *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 6, pp. 558–568, 1998.

[4] Rueber, B. “Obtaining Confidence Measures from Sentence Probabilities”, *EuroSpeech 1997*, Rhodes, Greece, pp. 739–740.

[5] Schaaf, T. and Kemp, T. “Confidence measures for spontaneous speech recognition”, *ICASSP 1997*, Munich, Germany, pp. 887–890.

[6] Wendemuth, A., Rose, G., and Dolfing, J.G.A. “Advances in Confidence Measures for Large Vocabulary”, *ICASSP 1999*, Phoenix, USA, pp. 705–708.

[7] Wessel, F., Macherey, K., and Schlueter, R. “Using word probabilities as confidence measures”, *ICASSP 1998*, Seattle, USA, pp. 225–228.