CONTEXTUAL CONFIDENCE MEASURES FOR CONTINUOUS SPEECH RECOGNITION

Gustavo Hernández-Ábrego and José B. Mariño

TALP Research Center
Dept Teoria del Senyal i Comunicacions, Universitat Politècnica de Catalunya
Jordi Girona 1-3, Campus Nord D-5, Barcelona 08034, Spain
(abrego/canton)@gps.tsc.upc.es

ABSTRACT

This paper explores the repercussion of contextual information into confidence measuring for continuous speech recognition results. Our approach comprises three steps: to extract confidence predictors out of recognition results, to compile those predictors into confidence measures by means of a fuzzy inference systems whose parameters have been estimated, directly from examples, with an evolutionary strategy and, finally, to upgrade the confidence measures by the inclusion of contextual information. Through experimentation with two different continuous speech application tasks, results show that the context re-scoring procedure improves the capabilities of confidence measures to discriminate between correct and incorrect recognition results for every level of thresholding, even when a rather simple method to add contextual information is considered.

1. INTRODUCTION

In speech technology, continuous speech recognition represents a big challenge where the inclusion of language modeling (LM) has proven to rise the accuracy of the developed systems. To employ LM in speech recognition is to take profit of the redundancy that the context of every word contains in order to better aim the decoding process. Nevertheless, LM may induce errors in the system forcing the presence of word sequences plausible for it but which were not present in the original utterance. On a previous work [1], we have constructed fuzzy confidence measures (CM's) as a feasible way to discriminate between correct and incorrect recognition hypotheses for a number of speech recognition applications. It seems natural to expect that the inclusion of contextual information in the calculation of confidence scores will increase their capabilities for continuous speech applications. Recently, two perspectives to combine LM probabilities and CM's for continuous speech have been tried: to use LM as another knowledge source for confidence measuring [2] and to include CM's in the LM [3] employed to recognize. Our approach takes profit of the context as a rescaling factor in order to post-process the CM's that were generated through a feature compilation procedure. Thus, it comprises three steps: feature extraction, feature compilation and CM re-scoring. The first two are described in section 2 while confidence re-scoring is detailed in section 3. The framework of our experimentation is depicted in section 4 and results are discussed in section 5. Conclusions are enumerated in section 6.

2. CONFIDENCE EVALUATION

Confidence measures can be generated by combining information about the recognition system in a feature-compilation fashion [4]. This approach has proven to rise the discriminative power of CM's when the features are extracted from the comparison of alternative recognition hypotheses or from multiple hypotheses recognition schemes [5]. Through careful study of the nature of the continuous speech recognition process, we have formulated three features to be the basis of our experimentation:

2.1. Features for confidence scoring

Our first feature is the *likelihood score ratio* (LSR). For its calculation, the likelihood score of the recognition hypothesis is normalized by the score of an alternative recognition network:

$$LSR = \log L(\vec{X}|\Lambda_p) - \log L(\vec{X}|\Lambda_a). \quad (1)$$

$\vec{X}$ is the vector of acoustic features related to the actual input utterance and Λ_p and Λ_a are the sets of hidden Markov models (HMM's) of the "principal" and "alternative" recognition networks respectively. There are two recognizers involved in our calculation of CM's: the principal, from which the hypotheses are taken, and the alternative, a reference used as *second opinion* [4]. Due to its unconstrained (and inaccurate) nature, the alternative network is capable to detect any sort of speech event although its results cannot be considered as recognition hypotheses. To use its score as normalization factor helps to verify the presence of an acoustic event in the input utterance. Because of its simplicity and high performance [3], we consider this feature, in isolation, as our baseline.

Our second feature is what we call *sequence alignment score* (SAS). This feature is intended to express the resemblance of two independent recognition hypotheses [4]. Its calculation is done by comparing both (principal and alternative) decoded strings through time alignment. The result is a sequence of "confusion pairs". A sequence of such pairs

This research was supported by CONACyT and by CICYT under contract TIC98-0423-C06-01

0-7803-6293-4/00/$10.00 ©2000 IEEE.

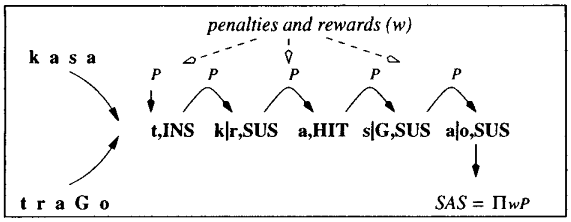

Figure 1: Calculation of SAS

appears on figure 1, every pair contains the unit (or units) it comprises and the result of the comparison (hit, substitution, insertion or deletion). To quantify the comparison, it is possible to compute the score of the alignment by means of a confusion matrix [4], calculated from the results of previous recognition experiments, but, going further, we have built a bigram of confusions, also trained with previous material, that expresses the probability of having a confusion pair given another one. The overall alignment score is the product of all "confusion bigram" probabilities for the sequence that results from the comparison. In figure 1 the bigram probabilities are represented by arcs between pairs. Depending on the nature of every pair (type of unit and alignment result), the bigram probabilities can be weighted by external penalties (for the case of non-matching units) or rewards (when units coincide). A further refinement is to categorize errors in terms of the units involved, for instance the confusion of two voiced units shows more evidence of mismatch between the principal and alternative strings than the confusion of two nasals. With this criterion, further weighting is applied to bigram probabilities.

Our third feature, that we call *relative speaking rate* (RSR), is conceived to handle the insertion and deletion errors. Within a given utterance, it may be expected that the speaking rate is maintained by the speaker between certain margins. Whenever in the utterance appears an abrupt change of speaking rate, the presence of insertions or deletions in the recognized phrase can be suspected. RSR is calculated according to:

$$RSR = \log \frac{N_{i,f}/t_{i,f}}{N_T/t_T}. \qquad (2)$$

N_T is the total of speech units detected in the whole hypothesis and $N_{i,f}$ are the units detected in the time interval $t_{i,f}$ where the actual recognition hypotheses (e.g.. word) is located. t_T is the duration of the whole utterance. Since not every unit lasts the same, the lengths (N_T and $N_{i,f}$) usually do not correspond to the number of units found. Lengths are calculated according to a table of normalized relative durations typical for each unit.

2.2. Feature compilation engine

To build CM's in a "feature-compilation" fashion has recently become a common procedure. Among the several combination schemes studied, neural networks [5] and fuzzy logic systems [4] present the best performance. Due to its versatility, efficiency and good performance [1], we have chosen a fuzzy inference system as a feature compilation engine. A Sugeno-type FIS is chosen due to its good behavior as classifier and its simplicity. One principal disadvantage that fuzzy logic systems present compared to neural networks is the need of expert knowledge for their design. To alleviate this drawback, based on the illustrative work of Shi et al [6], we have implemented an evolutionary procedure (founded on genetic algorithms) to train the parameters of the fuzzy system from examples. The goal of training is to maximize the performance of CM's as correctness predictors. In a result classification task, performance can be evaluated through ROC (receiver operation characteristics) curves. The tradeoff between the two kinds of classification errors (false alarms and false rejections) is graphically represented in a ROC (see figures 2 and 3). A useful summary for ROC's is the normalized area below the curve that indicates the average level of correct detections for the whole range of operating points (the normalized area of an ideal classifier would be 1). Our training procedure is suited to maximize the area below the ROC. For the feature compiler system, input variables are the three features extracted from recognition and the output is the fuzzy value (between 0 and 1) of confidence. The population for the evolutionary procedure can be initialized randomly (learning from scratch) or with a working system that is to be optimized. With this procedure we have finely tuned the parameters of the fuzzy systems presented in [1].

3. CONTEXTUAL CONFIDENCE RE-SCORING

Intuition suggests that correct and incorrect hypotheses do not appear in isolation, one error may lead to another and the same may happen with correct results. If the confidence labeler works efficiently, a very low value of CM denotes the presence of a wrong result. A wrong word may suggest that the surrounding ones are also wrong even if their confidence scores are high. Having a sequence of recognition hypotheses, $W_0^m = w_0, w_1, \ldots, w_m$, along with a corresponding sequence of confidence scores calculated after feature compilation, $CF_0^m = c_f(w_0), c_f(w_1), \ldots, c_f(w_m)$, the context-rescaled CM's, $CC_0^m = c_c(w_0), c_c(w_1), \ldots, c_c(w_m)$, can be calculated as:

$$c_c(w_n) = s(w_n) \cdot c_f(w_n), \qquad (3)$$

being $s(w_n)$ the scaling factor that depends on the information of the context of every word w_n. We have decided just to consider the immediate context (w_{n-1} and w_{n+1}) of every hypothesis, although the results here obtained can be easily extended to larger considerations. The contextual information that we have taken into account are the adjacent confidence values and the probabilities of the LM used for recognition (i.e., an application specific bigram). It has been shown [2] that the solely use of LM probabilities adds little (if any) information and that it is only useful when used simultaneously with the surrounding CM's. Essentially, CM's and LM probabilities are measures of uncertainty that can be treated in the fuzzy logic framework. To combine them, we propose to use another fuzzy inference system with two inputs (LM probabilities and CM's) and the scaling factor $s(w_n)$ as output. The magnitude difference between the values of inputs is compensated by non-linearly transforming the probabilities with

• Inputs		
1. LM probs.	=	$\sqrt{P(w_n \mid w_{n-1})P(w_{n-1} \mid w_n)}$
2. adjacent CM's	=	$(C_f(w_{n-1}) + C_f(w_{n+1}))/2$
• Output		
rescaling factor	=	$s(w_n)$
• Maximization		
ROC area $C_c(w_n)$	=	$s(w_n) \cdot C_f(w_n)$

Table 1: configuration of post-processor fuzzy system

	Time	GDB
voc. size	59	1215
bigram perplexity	4.71	20.04
utterances	995	535
word detection	97.04 %	95.98 %
phrase detection	85.23 %	79.88 %

Table 2: Tasks configuration and performance

μ-law. Once again, the evolutionary procedure described on section 2.2 is used to estimate the parameters of this "post-processor" fuzzy system. The objective of the training is, again, to maximize the ROC area for the re-scored CM's. The description of inputs, outputs and maximization target is shown on table 1. The first input represents the geometric mean of forward and backward bigram probabilities and the second is the arithmetic mean of surrounding CM's. The use of this second fuzzy system turns confidence measuring into a cascade procedure whose final product is an upgraded predictor of correctness. However, for the cases where LM probabilities provide little information (e.g. vocabularies with equiprobable instances), the former proposal can be reduced to the solely consideration of CM's, with no need of fuzzy combination step; thus, the re-scoring factor s might be represented by the arithmetic mean of the adjacent CM's. This "simplified" scheme is also tried in our experimentation.

4. EXPERIMENTAL FRAMEWORK

Experimental work has been carried out in two different continuous speech recognition tasks:

Time, taken from the Spanish Speechdat [7] database. Speech was collected through the fixed telephone network, recorded under several acoustic environments and sampled at 8 kHz. Close to 1000 speakers utter prompted date and time phrases.

GDB (geographical database) is part of the Albayzin [8] database. Originally, speech was collected under laboratory conditions and sampled at 16 kHz. For the sake of consistence, signals were passed through the telephonic channel and down-sampled at 8kHz. More than 130 speakers utter queries to a geographical database that contains information about cities, population, rivers, mountains, etc of the Spanish geography.

From every task database, training and test parts are separated with no signal overlap between them. Training parts are used to estimate the parameters of the specific fuzzy systems for each application. Speech was parameterized with mel-cepstrum coefficients and their cepstral means were subtracted. First and second order differential parameters plus the differential energy were used. The recognition system models the phonetic units by Gaussian semi-continuous HMM's with quantization to the 6 (2 for the energy) closest codewords. The codebook size was 128 (32 for the differential energy). For acoustic modeling of the principal recognizer, a set of high-performance sub-lexical phonetic units (Demiphones) trained under a discriminative framework [9] is used. Phonetic units were combined into vocabulary instances by means of finite-state automata as LM. Configuration and recognition performance of principal systems for the test parts of both applications are described in table 2.

On the other hand, the alternative recognizer was equipped with Phonemes discriminatively trained for acoustic modeling and loose language modeling restrictions, represented by a bigram that modeled the configuration of Spanish language. Experience indicate us, as a rule of thumb, that the higher the correct detection of the alternative system is, the better CM's work. However, one must be careful to avoid forcing false detections in the alternative recognition with restrictive LM's.

Generation of contextual confidence measures passes through the calculation of features for each word detected, then the features are compiled into a fuzzy CM. Next, CM's are re-scored according to their surrounding contexts. To avoid direct thresholding, performance of CM's is evaluated at several operating points.

5. DISCUSSION

As previously stated (section 2.2), our measures of performance are ROC's and the areas below the curves. Figure 2 shows the ROC's for the "Time" recognition task.

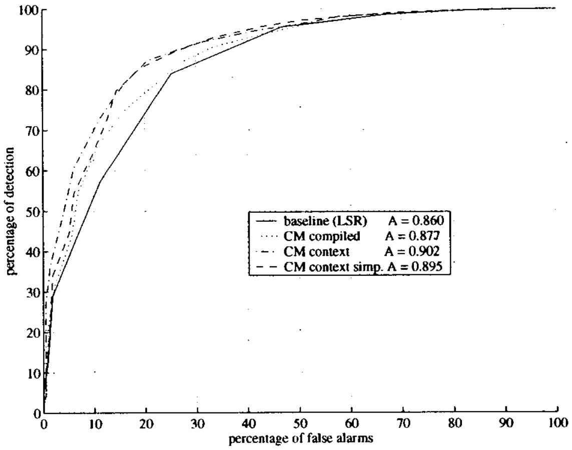

Figure 2: ROC's and areas for the "Time" database

Fuzzy-compiled CM's (which contains information from the three features) perform better than the best feature in

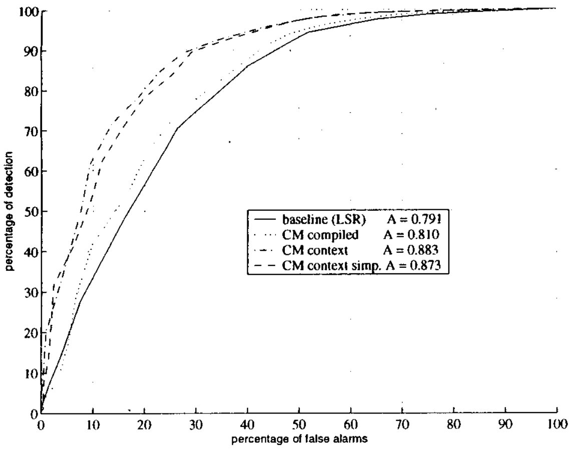

Figure 3: ROC's and areas for the "GDB" database

isolation (dotted against solid lines) and further improvement is achieved after contextual re-scoring (dashed and dash-dot lines). Between the two approaches for context inclusion (fuzzy combination and simplified method) the best performance is reached by the combination of LM and CM's but the simplified scheme is not far beyond it in terms of area below ROC's.

For the GDB task, results are plotted on figure 3. As expected, feature-compiled CM's perform better than baseline. For re-scoring, the fuzzy combination performs better than the simplified method but, again, they are not far from each other. The enhancement for rescaled CM's is noticeable. In the critical region of low false alarms (between 0 and 20 percent) a relevant increment of 20 points of detection is achieved. Overall, the increment of performance is noticed by the area below the re-scored curve: an average improvement of 7 points of detection is reached for the whole range of operating points. A comparison of the curves for both applications shows that the final performances of the re-scored CM's are similar, even though the capabilities of the baseline experiments are very different. The drop of performance of the compiled CM's in the GDB application is substantially alleviated by its richer context. Although the simplified method performs worse in both applications, its use is justified (and preferred) when design and computational efforts are issues to be considered.

Qualitative analysis of results at hypotheses level shows that, overall, the values of confidence for wrong hypotheses are reduced while those for correct ones are maintained or augmented in most of the cases. The kind of wrong hypotheses that are affected the most are errors that appear surrounded by other errors. Also wrong beginning and ending hypotheses that lead or follow further errors are better labeled. At a syntactic level, short words as articles, prepositions and conjunctions that follow correctly tagged words such as nouns or proper nouns are the hypotheses better re-scored. This procedure is particularly suited for applications where the construction of hypotheses largely depends on this kind of particles.

6. CONCLUSIONS

In this paper we have demonstrated that the implementation of a rescaling procedure, that conveys contextual information for every recognition hypothesis, provides great enhancement in confidence measuring for continuous speech recognition. Re-scoring results in a considerable reduction of CM values for consecutive wrong recognition results and for wrong beginnings and endings. It has been found that contextual CM's are specially helpful for short words. When CM's are used to discriminate between correct and incorrect results, considerable improvement is achieved for every operating point tried. Particularly noticeable is the increment of detection for low levels of false alarms. These conditions stand even when contextual information is added by means of a quite simple procedure. On the other hand, the application of an evolutionary fuzzy system as feature and information compiler introduces an efficient tool for the fields of recognition results evaluation and utterance verification.

7. REFERENCES

[1] G. Hernández-Ábrego and J. B. Mariño, "Fuzzy reasoning in confidence evaluation of speech recognition", in *Proceedings of WISP'99*, Budapest, September 1999, IEEE, pp. 221–226.

[2] D. Willet, A. Worm, C. Neukirchen, and G. Rigoll, "Confidence measures for HMM-based speech recognition", in *Proceedings of ICSLP'98*, Sydney, November 1998, vol. VII, pp. 3241–3244.

[3] R. C. Rose, H. Yao, G. Riccardi, and J. Wright, "Integration of utterance verification with statistical language modeling and spoken language understanding", in *Proceedings of 1998 ICASSP*, Seattle, May 1998, vol. I, pp. 237–240.

[4] G. Hernández-Ábrego and J. B. Mariño, "A second opinion approach for speech recognition verification", in *Proceedings of the VIII SNRFAI*, Bilbao, May 1999, vol. I, pp. 85–92.

[5] T. Schaaf and T. Kemp, "Confidence measures for spontaneous speech recognition", in *Proceedings of 1997 ICASSP*, Munich, April 1997, vol. II, pp. 875–878.

[6] Y. Shi, R. Eberhart, and Y. Chen, "Implementation of evolutionary fuzzy systems", *IEEE Transactions on fuzzy systems*, vol. 7, no. 2, pp. 109–119, April 1999.

[7] A. Moreno and R. Winsky, "Spanish fixed network speech corpus", Tech. Rep., SpeechDat Project LRE-63314, 1997.

[8] F. Casacuberta, "Development of Spanish corpora for speech recognition research", in *Proc. of Workshop on International Cooperation and Standarization of Speech Databases and Speech I/O Assesment Methods*, 1991.

[9] A. Nogueiras and J. B. Mariño, "Minimum confusibility training of context dependent demiphones", in *Proceedings of EUROSPEECH'99*, Budapest, September 1999, vol. VI, pp. 2741–2744.