ON THE INFLUENCE OF FRAME-ASYNCHRONOUS GRAMMAR SCORING IN A CSR SYSTEM

A. J. Rubio, J. E. Díaz-Verdejo, P. García and J. C. Segura

E-mail rubio@hal.ugr.es
Dpto. de Electrónica y Tecnología de Computadores
Universidad de Granada, 18071 GRANADA (Spain)

ABSTRACT

It is usually assumed that grammar probabilities and acoustic probabilities in a Continuous Speech Recognition system have to be incorporated to the general score with different weights. This is an experimental fact and there is no generally accepted theoretical explanation.

In this paper we propose an explanation to this fact, related to the way grammar scoring is incorporated in the searching procedure. Accordingly to this explanation, we perform a set of experiments to test our hypothesis.

We are also proposing a new way of introducing grammar probabilities in a tree-based vocabulary search strategy, where systems are usually bound to use the worst strategy.

To apply our ideas to unigrams is rather simple. For more complex language models like bigrams we have to implement a new procedure.

1. INTRODUCTION

The Automatic Speech Recognition problem can be summarized as finding the sequence of words W that maximizes the conditional probability $P(W|X)$, where X is the input to the recognizer.

Unfortunately, this probability is not directly measurable and we have to use Bayes' rule

$$P(W|X) = \frac{P(X|W)\ P(W)}{P(X)}$$

Now, $P(X|W)$ is the probability of the acoustics given the sequence of words, which can be estimated by using a statistical model, $P(W)$ is the probability of the sequence of words, estimated from a stochastic language model, and finally $P(X)$ is the probability of the acoustics, which is not generally considered, as results do not actually depend on it.

Therefore, two probabilities have to be included in the scoring to be summarized in practice

$$S(W|X) = P(X|W)\ P(W)$$

Experimental results show that the best performance of the recognition systems is obtained when the contribution of the language model is modified by introducing an experimental parameter e in the following way

$$S(W|X) = P(X|W)\ P(W)^e$$

There is no satisfactory theoretical explanation for this empirical fact, as Bayes' rule predicts that $e = 1$ should be the correct solution.

Some authors suggest that the need of this experimental compensation for the grammar contribution to the general score is due to bad models estimation [1], in such a way that the decision procedure is mainly driven by the grammar, with little influence of the acoustics. Values of e greater that 1 could diminish the importance of grammar in favor of acoustics.

2. FRAME-ASYNCHRONOUS SCORING

Our suggestion is that the root of the problem is a different one, related to the lack of synchronization between the ways in which both acoustical and grammar probabilities are incorporated in most Automatic Speech Recognition systems. The same problem has been observed when trying to apply SLHMM to CSR, using an asynchronous scoring for acoustic probabilities [2].

Most systems implement a pruning technique to the search space when looking for the best sequence of words according to the input. This pruning process is necessary because of the size of memory and time a system would need otherwise. It is implemented in

0-8186-7919-0/97 $10.00

a way that reduces the search space without introducing too much degradation in the performance of the system.

In principle, without pruning, the order in which all probabilities are scored is irrelevant, as the total final product remains unchanged. But the pruning algorithm introduces some distortions to this commutative property. When several states are compared to prune those with lower local score, we are usually comparing scores obtained in different situations.

In our recognition system, for example, we score the grammar probability at the beginning of a word, when there is an expansion from the last state of the last phoneme of a word to the first state of the first phoneme of the following word. Thus, an unlikely word (according to the grammar) could be eliminated in the pruning process before its acoustical evidence is checked.

Other possibility widely used is to score the grammar probability only when a given word is finished. In this case, a word could be eliminated by the pruning process because its bad acoustic evidence even if grammar would indicate that that is the best candidate. This is the case of systems with tree-structured vocabularies, since words (and their probabilities) are unknown until reaching the end of the branch (a leaf, corresponding to a word).

Therefore we suggest that the intimate reason for artificially weighting grammar probabilities against acoustic evidence could be that while acoustical evidence is estimated and scored in a frame-synchronous manner, grammar probabilities are scored only in a word-synchronous way.

3. SOME EXPERIMENTAL EVIDENCE

To test our hypothesis, we have performed a set of experiments, aimed to show the influence of the asynchronous scoring of grammar probabilities. For this purpose, the grammar probabilities have been accumulated at the beginning of words, at the beginning of the i-th phoneme (assuming the word is long enough) and at the end of the words. The experiments have been developed in a DHMM-based system using a 256 centroids codebook.

The database used is a part of EUROM1 [2], [3], composed by 40 speakers for training and testing, using 803 sentences for training and 403 new sentences for testing in multispeaker mode. The vocabulary size is 1103 and the bigram perplexity is 7.0. Table 1 summarizes the important characteristics of the database.

Experimental results (table 2) show that the best approach is to accumulate the probability at the beginning of the words. It is worth to mention that one

Database	EUROM-1
Language	Spanish
Number of speakers	40
Training sentences	803
Test sentences	403
Vocabulary size	1103
Bigram perplexity	7.0

Table 1: Database parameters

of the worst possibilities is to accumulate at the end of words (without using grammar weighting), as it is usually done when a tree-based vocabulary is used [4].

Table 2 shows the word error-rate (WER) for two series of experiments: the first one according to the Bayes' rule ($e = 1$) and the second one by using the experimental exponent $e > 1$. First column indicates the place (phoneme) in which the probability has been accumulated, the second column gives the WER for no-weighting experiments and column number three contains the WER for the experiments that were carried out with a weight for grammar, which can be found in column four.

Optimal grammar weights varies across the experiments. We interpret this variation as a confirmation of the fact that grammar weight is influenced by the way the probabilities are accumulated.

To verify this suggestion an additional experiment in which probabilities have been equally distributed among the phonemes in a word has been carried out. In table 3 results for this experiment can be found. Line starting "B" corresponds to the situation in which all grammar probabilities are scored at the beginning of words (which is the best option, according to the experiment mentioned before), and line starting "S" cor-

	WER		
Phoneme Index	e=1	Best	e
Beginning	17.3	8.9	5
1st	19.5	14.1	3
2nd	19.6	13.3	3
3rd	19.0	13.1	4
4th	19.0	13.4	4
5th	18.6	12.8	3
6th	19.0	11.1	4
7th	18.4	11.7	4
8th	18.9	11.0	4
End	19.5	10.4	5

Table 2: Experimental results obtained without grammar weighting (e=1) and for the best weight (e value)

e	1	2	3	4	5	6	7
B	17.3	12.1	10.6	9.1	8.9	9.3	10.7
S	18.1	12.6	11.3	10.1	9.7	9.7	10.4
E	19.5	13.1	11.6	11.2	10.4	11.3	12.4

Table 3: Results obtained for distributed probabilities (Split) compared to experiment "Beginning"

responds to the situation in which probabilities have been distributed along the word. Finally, the line starting with "E" corresponds to accumulating probabilities at the end of words. Results on this table show a performance not so good as the one obtained when grammar probabilities are accumulated at the beginning of words, but better than the performance obtained for other situations. Thus, tree-based systems can take advantage of this fact.

4. TREE-BASED GRAMMAR

We are extending this idea of synchronization to the tree-based large vocabularies search strategy.

When the vocabulary size is large enough, the simple organization of the vocabulary as a list is not valid, due to memory and time requirements of the search procedure.

Thus, most large-vocabulary systems use a tree strategy for vocabulary structure. For tree-based vocabulary structures, the only practical solution until now is to score grammar probabilities at the end of a word, because only at this moment the appropriate probability can be accessed. This is the worst situation related to the scoring strategy (with no grammar weight), according to the experiments commented in Section 3. Even though grammar probabilities can not be scored at the beginning of the words, a better solution can be to distribute the influence of these probabilities along the word.

4.1. Unigrams

We have designed the vocabulary as a regular grammar of phonemes, whose transitions will be affected by "local" probabilities associated to transition between phonemes. We hope that this change from word-synchronous scoring to phoneme-synchronous scoring will bring better recognition results, and it will support our proposition. Anyway, the final goal is to achieve frame-synchronous scoring for both grammar and acoustics probabilities.

In Table 4, a simple example of a grammar can be seen. It corresponds to a lexicon composed by the

Entry	Probability
/a/ /m/ /o/ /r/	0.3
/a/ /l/	0.2
/a/ /m/ /o/	0.2
/o/ /s/	0.2
/o/ /s/ /o/	0.1

Table 4: A simple grammar

phonematic transcriptions given in the table and the probabilities given in the table are unigram probabilities.

Figure 1 shows the phoneme-level finite-state automaton constructed for grammar in table 4. In this figure, IS represents the beginning of every word (Initial State). Every arrow represents a transition to a new state, which produces a phoneme. Dashed arrows produce a NULL phoneme and represent a transition to the Initial State (IS), corresponding to the end of a word of the vocabulary.

Unigram probabilities (table 4) are now distributed along the words, especially in those parts that can be confusing because the existence of others words with the same starting phoneme sequence. Every probability refers to the transition from one phoneme to the following one, according to the grammar and the phoneme-history. This seems to be the best strategy, according to the experiments presented before.

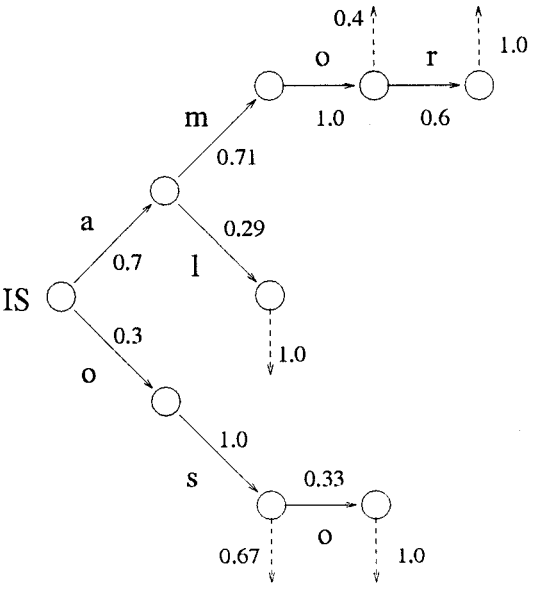

Figure 1: Tree corresponding to the grammar of the example

4.2. N-grams

The computation of this tree-based automaton for regular grammars is very simple. The real problem arises when we try to implement a more complex grammar like a bigram. In this case all probabilities corresponding to transitions between phonemes in the tree-structured finite-state automaton depend on the previous word and, therefore, the system needs a huge amount of memory. This is the same reason why tree-structured vocabularies are associated to a scoring-after-the-word strategy.

We propose to use unigram probabilities inside the tree-automaton and to compensate at the end of the word (dashed arrows) with an additional factor F given by

$$F = \frac{P(w|p)}{P(w)}$$

where w is the present word, p is the previous one, $P(w|p)$ is the bigram probability and $P(w)$ is the unigram probability.

This is a good approximation to the ideal situation of distributed probabilities only if, on average, unigram phoneme probabilities are not significantly different from bigram phoneme probabilities, which is in practice a normal situation when vocabulary size is large enough.

5. CONCLUSIONS

We have carried out a set of experiments to check our suggestion about the experimentally demonstrated necessity for weighting grammar probabilities against acoustical probabilities when determining the best word sequence in a Continuous Speech Automatic Recognition System.

We suggest that this necessity arises from the fact that both probabilities are scored in a unsynchronized manner. The lack of synchronism affects differently to both probabilities when we have to use a pruning algorithm, which is the usual situation.

Our experiments show that a distribution of the grammar probabilities along the word is a good solution or, at least, it is a better solution than accumulate grammar probabilities at the end of the words, as it is the usual procedure when using tree-structured lexicons.

6. ACKNOWLEDGEMENTS

This work is being funded by CICYT (Spanish Governmental Research Agency) under project TIC96-0956-C04-04.

7. REFERENCES

[1] K.-F. Lee and F. Alleva, *Continuous speech recognition*, ch. Advances in speech signal processing. Dekker, 1991.

[2] J. Díaz-Verdejo, *Reconocimiento de voz continua mediante una aproximación híbrida basada en SLHMM.* PhD thesis, Universidad de Granada, 1995.

[3] "Spanish EUROM-1: Phonetic contents," tech. rep., SAM-A/UPC/002, Universidad Autónoma de Barcelona, 1993.

[4] H. Ney, "Improvements in beam-search for 10.000-word continuous speech recognition," in *Proc. ICASSP'92*, vol. 1, pp. 9–12, 1992.