DYNAMIC SCHEDULING OF DECODING PROCESSES FOR DIRECTORY ASSISTANCE

Renato de Mori, Frédéric Béchet, Gérard Subsol, Dominique Massonié

LIA - University of Avignon, France
frederic.bechet,renato.demori,gerard.subsol,dominique.massonie@lia.univ-avignon.fr

ABSTRACT

This paper deals with the difficult task of recognition of a large vocabulary of proper names in a directory assistance application. Research on the European project SMADA has shown that there is a need of an elaborate and effective decision strategy that limits the risk of false automation. This paper proposes a new strategy which integrates, as well as a general decoder, a set of decoders specialized in some specific situations. Specialized recognition processes do not need to be applicable for every input, but they have to be scheduled and performed only under certain conditions. A first implementation of such a model is proposed here, through a rejection strategy of the hypotheses output by a general decoder. This strategy leads to a very significant improvement over the results obtained by a standard rejection method based on acoustic confidence scores only.

1. INTRODUCTION

Recognition of a large vocabulary of proper names is a difficult task of a very high perplexity. Moreover, practical applications require a low false automation rate, while, in many cases a certain amount of false rejections can be tolerated. A suitable dialog strategy can substantially reduce the false automation rate if the Word Error Rate (WER) on proper name recognition is kept low [1].

Actual state of the art Automatic Speech Recognition (ASR) systems show an increase in WER with task perplexity. Such an increase goes beyond acceptability thresholds when the size of the lexicon is that of the set of proper names of a big city or even a country.

Research on the European project SMADA [2] has shown that there is a need of an elaborate and effective decision strategy that limits the risk of false automation. In principle, a good strategy should evaluate an input with an initial set of ASR systems and produce an indication of acceptance or rejection. If no commit can be reached, suitable new processes which may involve specialized discriminative recognizers should be executed for refining the confidence.

Specialized recognition processes do not need to be applicable for every input, but they have to be scheduled and perform well only under certain conditions. Furthermore, they have to satisfy scheduling constraints, for example in time and space complexities.

These processes may use different acoustic features, different knowledge sources, different search algorithms, different scores and different models and each process can make an optimal set of decisions according to a given decision theory. It is important to stress the fact that performance of each process should not be evaluated on an entire test set, but only on the cases on which the process is applicable.

As decoders may use models of different precision, the decision strategy may consider combinations of hypothesis scores obtained by different decoders, but it can also reason about ranking of decoder outputs and their performance statistics. This approach follows the results obtained in the NIST evaluation programs where the composite ASR output of different systems has lower error rate that any of the individual systems [3].

The idea of scheduling processes based on preconditions was proposed with the blackboard model which was applied to ASR [4] without good results on limited tasks because the paradigm was developed for every step of the recognition process, including feature extraction. It is difficult, in this way, to model all the processes involved in ASR with precondition-action rules. This paper proposes a more realistic approach consisting in using the paradigm only for reasoning about scheduling of recognition processes which integrate models and local decision rules which have been already determined by an optimization procedure for that specific process.

2. DECODING ARCHITECTURE

2.1. Multiple decoder architecture

In principle, different decoders may use models of different units attempting to capture types of regularities in phoneme strings, phonotactics, environment knowledge. Different acoustic parameters and recognition paradigms (Hidden Markov Models (HMM), Artificial Neural Networks (ANN), Support Vector Machines (SVM)) can also be considered, as well as, different acoustic features with, for example, variable time-frequency resolutions.

In order to keep the computational complexity within acceptable values, the architecture having the scheme in figure 1 is proposed. This architecture can evolve into one in which different types of features are extracted by different front-ends.

Initially two decoders sharing feature extraction, phone models and canonical pronunciation models are used. The first decoder, indicated as D_1, is based on word models and generates an N-best list of hypotheses. The second decoder, indicated as D_2, generates a lattice of phoneme hypotheses and an initial N-best list of word hypotheses obtained using performance models applied to the most likely phoneme string.

Each of the first two decoders generates an N-best list of candidates and place them into a blackboard. When the correct hypothesis is not in either of the N-best list of candidates, the decision strategy should reject both lists and ask for a repetition. Even if with this type of rejection the number of recognition errors is reduced, it is possible to have a new type of errors: when the correct hypothesis is in one or both of the N-best lists and is wrongly rejected. This type of risk has to be taken into account together

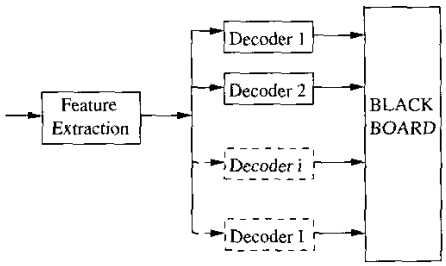

Fig. 1. Multiple decoder scheme

with word errors and a decision strategy has to be conceived which leads to the lowest overall risk.

The other decoders are executed only when certain situations described by logical expressions appear to be true based on the blackboard content.

A reasonable approach to the choice of decoders consists in focusing on models and procedures which address problems which are not solved by the available decoders and are known to be important. Furthermore, the complexity of new decoders must be such that resource and processing time constraints for a given application are satisfied. In the case of DA, the lexical models are of central importance, because a very large set of proper names has to be recognized. For such reason, investigation on pronunciation models has been given the highest priority [5, 6].

2.2. Automatic generation of pronunciation variants

Lexical models are of fundamental importance in proper name recognition. Each word in the lexicon can be represented by its canonical pronunciation generated by a Text-To-Speech (TTS) system, as for the experiments described in this paper.

Different distortions may be applied to the canonical form to produce the surface form $s(W)$ of a word W. These distortions may be produced by the speaker or perceived by the recognizer front-end. The distinction between these two cases is difficult to perform in practice. However, considering all the surface forms of each word which may arise from the imprecision of the recognizer knowledge may increase the confusion among word models and degrade recognition performance. This problem will be considered in this section by introducing decoders that use the same fast $A*$ search algorithm on a lattice of phone hypotheses but with different types of word models allowing for phone insertion, deletion and substitution. More details about the dynamic generation of pronunciation variants can be found in [7]. Three kind of decoders are built following this method:

- D_I be the decoder that considers the insertion of one phone,
- D_D be the decoder that considers the deletion of one phone,
- D_S be the decoder that considers the substitution of one phone.

3. BLACKBOARD BASED DECISION STRATEGY

3.1. Description of the strategy

Decision strategies are treated in this paper according to the following definitions:

Let $A = \{a_i\}$ be a set of actions. The decision strategy proposed in this paper is based on a sequence of actions executed when certain preconditions are met. A conceptual difference between the decision strategy proposed here and classical sequential decision strategies is that the focus is on the choice of preconditions which are logical expressions involving predicates whose truth depends on situations arising from previously executed actions. The precondition-action paradigm has been used in blackboard architectures studies in Artificial Intelligence (AI) and applied to ASR [4].

An action a_i is executed only if a precondition pc_i, describing a situation is evaluated to true based on the recognition results of a set of decoders. This is represented by the following rule:

$pc_i \rightarrow a_i$

For each action a_i, a set of uncertain events $\{E_{ij}\}$ is considered for describing action outcomes.

A set of consequences $\{c_{ij}\}$ is associated to the set of events.

If action a_i is taken and the event E_{ij} occurs, then a utility function $u(c_{ij})$ is associated to it while the *belief* of event E_{ij} is defined as $P(E_{ij}|a_i)$.

The principle of quantitative coherence states that, among the actions that can be executed at a certain point in time, preference should be given to the action a^* with maximum utility function defined as follows:

$$a^* = arg \max_i \mu(a_i) \tag{1}$$

$$\mu(a_i) = \sum_{j=1}^{J} u(c_{ij}) P(E_{ij}|a_i) \tag{2}$$

This principle can be applied for selecting a preferred action when more than one preconditions turn out to be true.

3.2. Application to multiple decoders

In our case, an action is the execution of one or more decoders, each one of which produces an N-best list or a lattice of hypotheses placed into a blackboard.

Preconditions are logical expressions of predicates describing the blackboard content. Primary focus on preconditions should depend on aspects of the blackboard content that are considered important based on the knowledge of the behavior of processes executed by actions. If the processes are decoders producing N-best list, it is reasonable to reason about the ranking of hypotheses generated by different decoders.

For example, if two decoders D_1 and D_2 agree on the choice of the top hypothesis of their N-best lists, this precondition can be expressed by: $pc = [W_{11} = W_{21}]$ where W_{ij} correspond to the hypothesis ranked j in the N-best list produced by decoder i.

The uncertain events E_{ij} attached to each action a_i correspond to the four following situations:

$$E_{ij} = \begin{cases} \text{accept correct} & cost = 0 \\ \text{accept error} & cost = \alpha \\ \text{reject correct} & cost = \gamma \\ \text{reject error} & cost = 0 \end{cases}$$

The consequence of each decision on each utterance has a cost which can be adapted following the specifications of the Directory Assistance system. The utility of each consequence is defines as $1 - cost(c_{ij})$ while the event probabilities can be estimated with a development set.

3.3. Risk function

A number of new decoders has to be considered and preconditions for their executions can be such that the corresponding action leads to the maximum reduction of the value of a risk function computed on the set of data to which the action is applicable. Enough rules have to be introduced in such a way that a recognition result can be produced for every input. Only in this case a strategy is complete. Once a complete strategy (S) is available, the total risk for the development set ($\mu(S_{dev})$) can be expressed as follows (according to equation 2 and the costs previously introduced):

$$\mu(S_{dev}) = \gamma \times N_{fa} + \alpha \times N_{fr} \quad (3)$$

where N_{fa} is the number of recognition errors (*false acceptance*) and N_{fr} is the number of cases in which the correct hypothesis is in the initial N-best lists and it is incorrectly rejected (*false rejection*). For the sake of simplicity, the following risk density will be used assuming all the costs per unit to be equal to one:

$$e = \frac{N_{fa} + N_{fr}}{N_{total}} \times 100 \quad (4)$$

N_{total} is the total number of items in the development set.

During the decoding process, the blackboard contains three types of information, namely the strategy rules, the hypotheses generated by the decoding processes and the scoring algorithms for making decisions when more than one rule can be applied. The application of a process generates a result which is described by a situation. For each situation one of the following actions can be taken: output a result; reject the input; execute a process.
Among the possible action, the one which results in the minimum risk should be taken.

4. EXPERIMENTS

4.1. Experimental setup

These experiments are carried out on a corpus of sequences *first name-family name* collected on the internal France-Telecom R&D Directory Assistance system. We use a $4K$ utterance development corpus and a $2K$ utterance test corpus. The lexicon used for obtaining the N-best lists contains about $100K$ names.

As presented in section 2 we use, in a first step, two decoders: one based on word-models and one using a phoneme string obtained from a lattice of phoneme hypothesis. The best decoder, according to the performance obtained on the development corpus is called D_1.

The strategy we implemented, following the formal description of section 3, consists in accepting or rejecting the best hypothesis of the N-best produced by D_1 and called W_{11}. The criteria used are based on the ranking of W_{11} in the N-best lists produced by our multiple decoders. This strategy is obtained on the development corpus and consists of two kind of processes: firstly, a set of rules precondition/action as presented in section 3.1 schedules the application of the different decoders. Secondly, an algorithm, specific to each situation, makes decision when more than one rule can be applied after the execution of a given action.

In addition to this method, and in order to compare our rejection strategy with a more standard one, we implemented a baseline strategy which simply estimates the difference of scores between W_{11} and W_{12} (the first and second hypotheses output by D_1). If this difference is bigger than a threshold δ, then W_{11} is kept otherwise we reject the whole N-best list.

4.2. Rejection strategy

The first step in our rejection strategy consists in defining the precondition/action rules which schedule the application of the different decoders. As we already said, we only use, for the moment, decoders sharing the same feature extraction process and the same phone models. That's why the design of the scheduling rules is very basic. Once different decoders specific to particular problems will we introduced, this process will be more complex.

The first rule is: $pc_0 \rightarrow D_1$ with pc_0 corresponding to an empty precondition starting the process. Once D_1 is performed, two situations are possible:

1. $score(W_{11}) - score(W_{12}) > \delta_0$: if the difference of score is bigger than a threshold δ_0 (corresponding to a high value), then the process is stopped and W_{11} is output as the solution without any further process;
2. otherwise this first hypothesis is not considered reliable enough, and the equation $score(W_{11}) - score(W_{12}) \leq \delta_0$ corresponds to the precondition pc_1.

The second rule is: $pc_1 \rightarrow D_2 D_I D_D D_S$ with D_2 corresponding to the phone-based decoder and D_I, D_D and D_S being the decoders that allow a distortion in the canonical forms of the hypotheses output by D_2 as presented in section 2.2.

The concept explored in this strategy is based on the rescoring of a limited set of previously generated hypotheses. If the same hypothesis gets the highest likelihood when the canonical forms is modified by limited perturbations, then it is likely that this hypothesis is the correct transcription of what has been uttered.

Once the N-best lists of all these decoders are obtained and placed in the blackboard, the strategy consists in trying to apply a list of precondition $pc_2, \ldots, pc_n$. Each of them corresponds to a logical expression about the ranking of the hypothesis W_{11} by the different decoders. When a precondition pc_i is applied to the N-best lists corresponding to an utterance, the following situations are possible:

1. pc_i is satisfied and the hypothesis W_{11} is accepted;
2. otherwise we try to apply pc_{i+1}

When no more preconditions can be applied, then the utterance is rejected.

This sorted list of preconditions is obtained by the following method:

- Firstly a set of logical expressions is empirically selected. These expressions indicate the position of the hypothesis W_{11} in the different N-best lists.
- Secondly an iterative process, using our development corpus, selects among all the possible expressions the optimal sequence that leads to the biggest decrease in the total error rate (e) as expressed in equation 4. At each step, all the possible logical expressions are applied to the development corpus. The expression pc_i that produces the smallest value of e on the sub-corpus containing all the utterances satisfying pc_i is selected. Then, the development corpus is reduced to the samples that don't satisfy pc_i and this process goes on until there is no more samples in the development corpus or no logical expression that can be satisfied on the remaining samples.

4.3. Results

The results, obtained on the $2K$ utterance test corpus (with no overlap with the development corpus), are given according to the ROC curve of figure 2. This curve presents the precision according to the false rejection rate. The precision is the percentage of correct answers on the total amount of answers (an answer is an utterance kept by the rejection strategy). The false rejection rate is the percentage of rejected utterances whose 1-best hypotheses produced by decoder D_1 were correct.

The baseline curve corresponds to a strategy based only on a rejection threshold. When the difference of score between W_{11} and W_{12} (in the N-best list produced by D_1) is below a given threshold, the utterance is rejected. The global WER of decoder D_1 is indicated at the extreme left of the curve (rejection threshold set to 0) and is about 30%. This curve highlights the lack of robustness of a rejection strategy based only on acoustic scores: at an operating point of 30% false rejection, the WER is 27%, which means that we reject 30% of the correctly recognised utterances with only a 3% improvement in the WER.

Strategy S_1 correspond to the method presented in section 4.2. As we can see, this strategy outperforms significantly the baseline method by improving the precision by more than 10% (absolute) at the same operating points of false rejection rate. For example, at 30% false rejection, the WER is 13.5%, which is a 50% improvement compared to the baseline. The lowest false rejection rate obtained with this method is 14% with a WER of 17%.

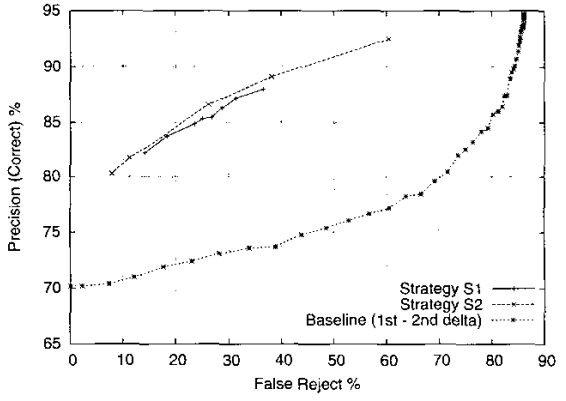

Fig. 2. ROC curves on the test corpus

Finally, the curve corresponding to strategy S_2 in figure 2 is a first attempt for merging the previous strategies based on acoustic scores and logical ranking by various decoders. In this experiment we add to the logical expressions presented in section 4.2 a criteria based on the proximity of the acoustic scores of the hypotheses produced by D_1. For each value of the threshold δ used in the baseline strategy, we calculate the number of items W_{1i} satisfying the following constraint: $score(W_{11}) - score(W_{1i}) < \delta$. This information is added to the logical expressions and the optimal rejection strategy estimated on the development corpus is performed in the same way as presented in section 4.2. The results given by the curve S_2 clearly indicate that this is a promising way, as on one hand the results are equal or even better on the portions of the curve covered by S_1 and on the other hand this strategy covers more operating points than strategy S_1 alone.

5. CONCLUSION

The rejection method presented in this paper is a first implementation of the blackboard based decision strategy presented in section 3. The results obtained clearly show that using the output of multiple decoders in order to accept or reject an utterance outperforms the results obtained with a standard rejection method based on acoustic confidence scores. This decision process may suggest new research directions. In fact, assume that the best HMM based decoders with certain features have been already used, the output errors and the false rejections can be collected and analyzed. The analysis may suggests that, for certain types of errors, certain processes may reduce the decision risk. For example, if there is a competition between two words, processes can be scheduled that have a high discrimination power for the phonemes which are different in the two words. This solution may be applicable in a limited number of cases and may not show a dramatic WER reduction, but it can be reused for other applications and many solutions of this type may show tangible advantages. A new perspective is thus open for going beyond the limits of actual ASR systems.

6. ACKNOWLEDGEMENT

The work described in this paper is part of research effort carried out in the SMADA project on Telephone Directory Assistance. This project is partially funded by the European Commission, under the Action Line Human Technology in the 5th Framework IST Program. The authors wish to express their thanks to Alexandre Ferrieux and Denis Jouvet from France Telecom R&D for the use of data.

7. REFERENCES

[1] P. Natarajan, R. Prasad, R. M. Schwartz, and J. Makhoul, "A scalable architecture for directory assistance automation," in *Proc. ICASSP'02, Orlando-Florida, USA*, 2002.

[2] Bechet F., Den Os E., Boves L., and Sienel J., "Introduction to the ist-hlt project speech-driven multimodal automatic directory assistance," in *in Proc. ICSLP 2000, Beijing, China*, 2000.

[3] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Rover," in *in Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, 1997.

[4] De Mori R., Lam L., and M. Gilloux, "Learning and plan refinement in a knowledge-based system for automatic speech recognition," in *IEEE Trans. On Pattern Analysis and Machine Intelligence*, 1987, vol. vol. 9, no.2.

[5] L. ten Bosch and N. Cremelie, "Pronunciation modeling and lexical adaptation using small training sets," in *Pronunciation Modeling and Lexicon Adaptation for Spoken Language, ISCA Workshop, Estes Park, Colorado, USA*, 2002.

[6] Strik H. and Cucchiarini C., "Modeling pronunciation variation for asr: A survey of the literature," in *Speech Communication, 29(2-4):225-246*, 1999.

[7] Bechet F., De Mori R., and Subsol G., "Dynamic generation of proper name pronunciations for directory assistance," in *in Proc. ICASSP'02, Orlando, USA*, 2002.