ON COMBINING RECOGNIZERS FOR IMPROVED RECOGNITION OF SPELLED NAMES

Denis Jouvet & Sébastien Droguet

France Télécom R&D, DIH/IPS,
Technopôle Anticipa, 2, Avenue Pierre Marzin,
22300 Lannion, France
denis.jouvet@francetelecom.fr

ABSTRACT

This paper deals with the recognition of spelled names over the telephone. Two recognition approaches are recalled. One is based on a forward-backward algorithm in which the spelling lexicon is handled by the A* algorithm in the backward pass. The other is a 2-step approach, which relies on a discrete HMM-based retrieval procedure. Both approaches integrate a rejection test. Combinations of the two approaches are investigated in this paper. First, a sequential combination is presented. The 2-step approach is used only when the forward-backward approach do not yield an answer because of memory limitations. This sequential combination, evaluated on field data collected from a vocal directory service, takes the best of both approaches. Results are presented for the recognition of valid spelled names as well as for the rejection of incorrect data. Finally, a detailed analysis of the recognition results of the 2 approaches shows that a comparison of the 2 recognition results leads to an efficient reliability criterion.

1. INTRODUCTION

Automatic recognition of names from their spelling is an important task with many obvious applications such as directory assistance or identification of city names. Natural spelling implies the recognition of connected letters, which is a difficult task, especially over the telephone. However, for many applications, the spelled names belong to a finite known list. Such a list provides a very useful information, and many recognition approaches differ in the way they handle it.

A basic approach relies on an unconstrained decoding of the sequence of letters, followed by a retrieval procedure. This 2-step approach is efficient but not optimal, and better performance is achieved by using the lexicon knowledge directly at the decoding level. This knowledge can be handled, for example, through a direct decoding in a finite state network [4, 2] or through dynamic grammars in a multi-pass procedure [6]. Trade-offs have to be found in order to obtain fast and accurate procedures. A forward-backward-based approach was proposed in [5] and is recalled later in this paper (section 2.1).

This work was partially supported by the SMADA European project. The SMADA project is partially funded by the European Commission, under the Action Line Human Language Technology in the 5th Framework IST Programme.

Almost all studies related to the recognition of spelled names deal exclusively with the recognition of correct spellings. Only a few of them have explored field data, and the main aspect they have studied was the handling of noise and extraneous speech before the spelling itself [1]. One of the main problems with spelling recognition is the detection of the end of the spelling utterance. A long hesitation often leads to a truncated spelling, which must be rejected, as well as the spelling of names which do not belong to the lexicon. Moreover, extraneous speech as well as noise tokens must also be rejected. The rejection of these various incorrect data was first investigated in [5], and will further be considered in this paper.

The organization of the paper is as follows. Section 2 recalls the 2 spelling recognition approaches: the forward-backward procedure and the 2-step approach, which relies on a discrete HMM-based retrieval procedure. The rejection of incorrect data is also recalled. Section 3 briefly presents the experimental setup. Section 4 details the sequential combination of the 2 approaches and its evaluation on field data. Section 5 presents a detailed analysis of the recognition results, provided by the 2 spelling recognition approaches, in view of defining a useful reliability criterion. Finally, a conclusion ends the paper.

2. RECOGNIZERS OVERVIEW

Two approaches are recalled here: a forward-backward procedure and a 2-step approach which relies on a discrete HMM-based retrieval procedure.

2.1 Forward-Backward Based Procedure

This approach is based on a forward-backward procedure as used in many N-best decoding algorithms. The two knowledge sources are the set of letter acoustic HMM models, and a finite state network which represents the lexicon. The forward pass uses an unconstrained letter model, and computes, for each state of the acoustic HMM and for each frame of the utterance, the likelihood along the best partial path ending in this state at this time. The backward pass is based on the A* algorithm and handles the spelling grammar (lexicon of allowed names). In the A* formalism, paths are ordered according to a global score. This global score is obtained by summing the score of the partial hypothesis already explored and an estimation of the score of the part that remains to be explored. In this type of forward-backward approach, this estimation results from the scores stored in the forward pass. To make the A* algorithm as efficient as possible, we check for every treated letter whether it constitutes, or not, a valid extension of a partial path according to the

grammar. If it does not, the hypothesis is discarded. This early discarding of non-valid paths avoids wasting CPU for determining the full hypothesis before checking its validity.

2.2 Discrete HMM-Based Retrieval Procedure

In this approach, the speech signal is first decoded using an unconstrained letter model. This decoding delivers a sequence of letters, which contains substitution, insertion and deletion errors. Then, knowing the recognized sequence of letters, the retrieval procedure searches for the most probable name in the lexicon. In [3] various retrieval procedures were compared. The best retrieval performance is obtained using a discrete Markov modeling approach, in which the recognized letters are the output symbols. An elementary Markov model, with 2 states and 3 transitions, is defined for each letter: a loop with an associated pdf to model insertions errors, a transition with an associated pdf to model substitution errors, and a null transition to model deletion errors. The pdfs (for insertion and substitution errors) define the emission probability of the recognized letters by the given (reference) letter model. This formalism allows to use HMM training procedures to estimate the optimal values of the model parameters. After the training process, these models represent recognition errors observed at the letter level.

2.3 Rejection Criteria

For recognition of spelled names, as for any other task, a rejection mechanism must be available to handle incorrect data. The mechanism presented here is based on the comparison of a likelihood ratio to a predefined threshold. As detailed in [5], the likelihood ratio which is computed is the ratio of the likelihood of the best lexicon-constrained decoding over the likelihood of the best unconstrained decoding. If this ratio is above a predefined threshold the recognized answer is accepted, if not it is rejected. In other words, the constrained answer is accepted if its score is not too different from the one of the unconstrained decoding.

In the forward-backward approach, the likelihood ratio is computed directly from the acoustic observation. The likelihood of the unconstrained decoding is available through the forward pass. The likelihood of the lexicon-constrained answer results from the backward pass.

In the discrete HMM-based retrieval procedure, a similar likelihood is computed from the recognized sequence of letters (letter symbols). The likelihood of the lexicon-constrained answer results from the retrieval procedure. A loop model is used to compute the likelihood of the unconstrained solution.

3. EXPERIMENTAL SETUP

The algorithms are evaluated on field data collected from a vocal directory task.

Because of possible variants in spelling double letters (“NN” for example may be spelled “N”.”N” or “2”.”N”) the word "2" is added to the letter vocabulary, and the spelling lexicon (or grammar) is enriched in order to take into account these spelling variants.

For the recognition experiments we have used the FTR&D (formerly CNET) lexicon which leads to 3600 different names yielding 4500 spelling variants (due to double letters).

The speech recognition system is based on mixtures of Gaussian function densities. Channel adaptation through blind equalization is included in the signal analysis. First and second order derivatives of the Mel cepstral coefficients are used in the modeling.

A double modeling of the letters is carried out. One is based on whole-word models. The size of the models was determined from the average duration of the letters estimated on a training set. The second modeling relies on a contextual modeling of the phonemes.

Field data was collected from the CNET Lannion vocal directory in operation since 1995. The database is divided into 4 subsets. One corresponds to the spelling utterances present in the lexicon. This set is referred to as valid-spelled names. The 3 other sets correspond to incorrect data. Non-valid spellings refer to all spelling utterances that do not belong to the lexicon, either because the user spelled a name not present in the lexicon, or because it is a truncated spelling. The 2 other subsets are made up of non-spelling speech data (such as command words, comments, …), and of noise tokens (noises detected by the endpoint procedure).

In Figures 1 to 4, the results are displayed by means of curves obtained by varying the rejection thresholds. The horizontal axis corresponds to the false rejection rate measured on valid spelled names, whereas the vertical axis corresponds to the most harmful error, that is substitution errors or false alarms depending on the type of data. Logarithmic scales are used on both axes.

4. SEQUENTIAL COMBINATION

The forward-backward approach suffers from memory limitations both in the forward pass where memory is required to store partial path likelihoods, and in the backward pass where memory is necessary for the A* heap. Because of the limited size of the heap used in the A* algorithm, and of the lexicon constraint, the search can stop before finding an answer. This occurs when no path has been found and no more cell is left in the A* heap. This phenomena usually occurs when the best lexicon-constrained answer has a score much lower than the unconstrained answer. This provides a way to reject incorrect data. Increasing the size of the heap reduces the number of cases where no answer is found, but at the expense of extra memory and CPU time. On the opposite, the basic retrieval procedure will always find an answer, namely the lexicon entry having the best retrieval score. The rejection test is then applied to accept or reject the answer. Hence the idea of combining the 2 approaches.

4.1 Sequential Combination

As the forward-backward based approach is the most efficient one (see results in Figures 1 to 4), this approach is favored, and the 2-step approach is only used as a repair procedure.

Hence, the forward-backward procedure is applied first. If an answer is found (either correct, error or rejection) this answer is kept.

But if the forward-backward procedure is not able to deliver an answer because of memory limitations, the 2-step approach is applied. This occur when there is not enough space to store all the partial path likelihood data (for any active state at each frame) in the forward pass, or when the A* heap is exhausted before a constrained answer is found. In such cases discrete HMM-based retrieval procedure is applied on the best-unconstrained letter decoding.

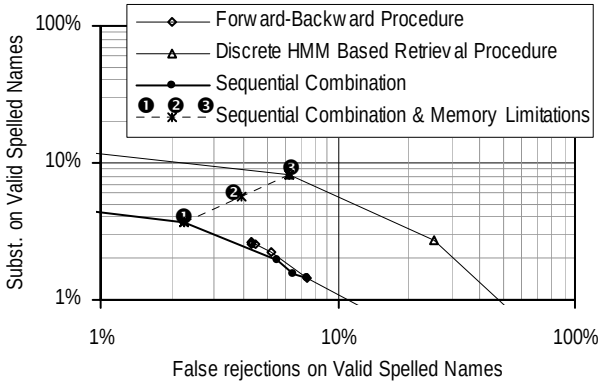

Figure 1 – Error Rates on Valid Spelled Names.

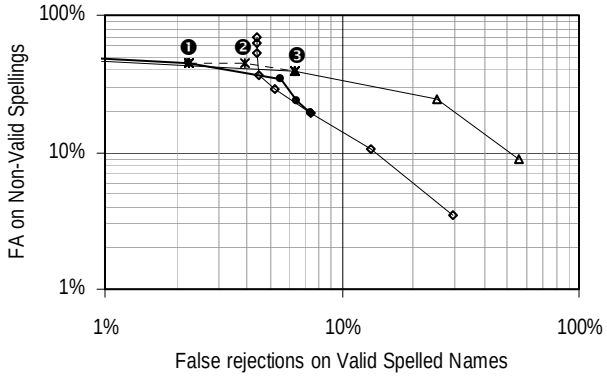

Figure 2 – Error rates on Non-Valid Spellings.

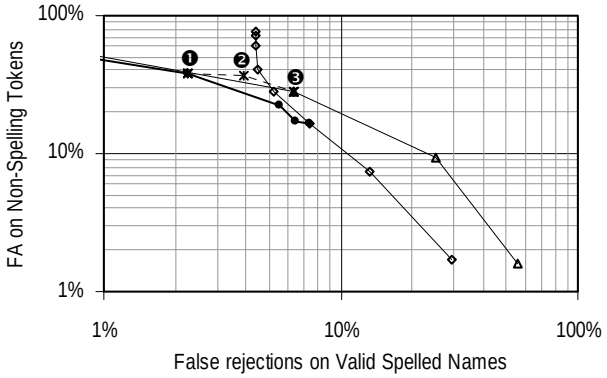

Figure 3 – Error rates on Non-Spelling Tokens.

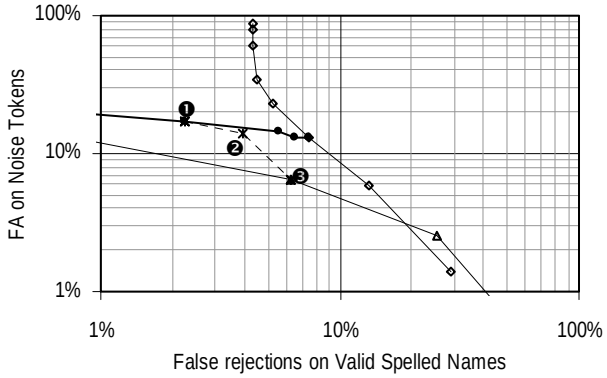

Figure 4 – Error rates on Noise Tokens.

4.2 Experiments and Discussion

Figures 1 to 4 show the ROC curves on the 4 subsets of the field database. It clearly appears that on valid spelled names the forward-backward based approach (diamonds) leads to a much smaller substitution rate than the 2-step approach (triangles). This is mainly due to a tight integration of all available knowledge sources (acoustic models and spelling grammar) in the decoding process, as opposed to the 2-step approach which successively uses these two knowledge sources. On non-valid spellings and non-spelling tokens, the forward-backward approach leads to a smaller false alarm rate when false rejection rate is greater than 4-5%. However on noise tokens the 2-step approach yields better results.

When both approaches are sequentially combined (circles), good results are obtained. As the 2-step approach is applied only when the forward-backward approach does not find an answer because of memory limitations, this combined approach takes the benefits of both procedures. It provides substitution and false alarms rates comparable to those of the forward-backward-based approach, and it also allows to achieve a much smaller false rejection rate, without increasing too much the substitutions and false alarms errors.

Another interesting feature of the combined approach is it graceful behavior when memory limitations becomes stronger. The dotted lines with stars (and numbers) are obtained by reducing more and more the memory devoted to the storage of the likelihood scores in the forward pass, and to the heap in the A*-based backward pass. The rejections thresholds were not modified. As can be seen from the curves, the error rates slowly move from the combined approach (❶) to those of the 2-step approach (❸). This corresponds to an increased usage of the 2-step approach when memory limitations becomes stronger.

5. TOWARDS A RELIABILITY CRITERION

A detailed analysis of the recognition results was also conducted. As the 2 recognition approaches do not behave exactly the same way, it is interesting to study the cases when they lead to the same result, and when they do not.

Table 1 – Amount of same & different (not rejected) answers, and amount of answers provided only by the 2-step approach.

	Valid Spellings	Non-Valid Spellings	Non-Spellings	Noise Tokens
Same Answer for 2 Approaches	70.4% (1065)	4.3% (120)	1.8% (68)	0.4% (13)
Different Answer for 2 Approaches	22.1% (335)	14.8% (416)	14.6% (537)	12.6% (416)
2-Step Only (and not rejected)	1.9% (29)	15.0% (422)	5.8% (215)	1.5% (49)
2-Step Only (and rejected)	5.6% (84)	65.9% (1852)	77.8% (2868)	85.5% (2830)

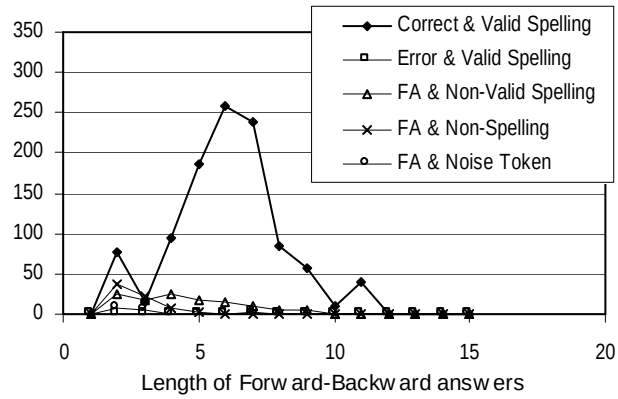

Figure 5 – Number of correct and incorrect answers when they are the same.

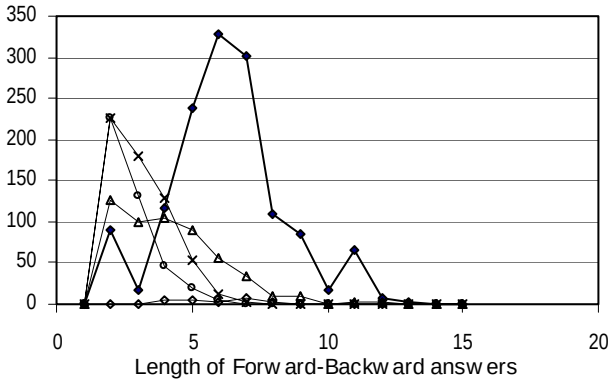

Figure 6 – Number of correct and incorrect forward-backward answers.

5.1 Comparison of the 2 Recognizer Answers

Table 1 summarizes the number of same and different answers (between the 2 approaches) for a given set of rejection thresholds. It clearly appears from the last 2 lines of the table, that when only the 2-step procedure provides an answer (i.e. no answer provided by the forward-backward approach), most of these answers are rejected.

Moreover same answers are mainly observed on valid spelling data, and different answers on the incorrect data subsets. Table 2 reports the number of correct and incorrect answers when they are the same for the 2 recognition approaches. Thus comparing the answers of both recognizers provides a strong hint on whether the answer is correct or not, however other information must also to be taken into account.

Table 2 – Number of correct and incorrect same answers.

	Valid Spellings	Non-Valid Spellings	Non-Spellings	Noise Tokens
Same Answer & Correct	1062	0	0	0
Same Answer & Incorrect	3	120	68	13

5.2 Taking into Account the Answer Length

One feature to consider is the length of the answer (number of letters). It seems quite obvious that if the recognizer recognizes a long sequence of letters with a lexicon constraint, it is more likely to be correct than if it is a short one.

Figure 5 shows the number of correct answers on valid spellings (thick line) and the number of incorrect answers in the various subsets for each answer length, but only when both approaches deliver the same answer. For comparison purposes, Figure 6 displays the same curves for all the answers provided by the forward-backward approach. From these figures, it clearly appears that if both approaches lead to the same answer, and if the answer has 5 or more letters, the answer can reliably be considered as correct. This criterion validates about 60% of the valid spellings. This means that almost 60% of the valid spellings can be considered as reliably recognized (i.e. for example no confirmation required in a dialogue) with only very few errors accepted. Figure 6 shows that the length criterion alone is not applicable, as it would accept too many false alarms.

6. CONCLUSION

Combinations of 2 recognition approaches was investigated for recognizing spelled names. One of the approaches is based on a forward-backward procedure as used in N-best approaches. An optimized handling of the spelling lexicon in the backward A*-based pass makes the algorithm efficient. The other approach runs in 2 steps and relies on a discrete HMM-based retrieval procedure.

A sequential combination of the 2 approaches (recognizers) allows taking the best of each procedure. This yields the same substitution error rate as the forward-backward procedure, and this allows to obtain small false rejection rates as with the 2-step procedure. Also, the combination of both approaches makes the global approach less sensitive than the forward-backward approach, to memory limitations.

Finally, a reliability criterion was proposed based on the comparison of the answers provided by the 2 recognition approaches, and on the length of the recognized answers. Such a reliability criterion successfully applies to almost 60% of the valid spelled names with very few errors.

REFERENCES

[1] Galler M. & Junqua J.-C. (1997), Robustness Improvements in Continuously Spelled Names over the Telephone. *Proceedings ICASSP'97*, Munich, Germany, pp. 1539-1542.

[2] Hild H. & Waibel A. (1996), Recognition of Spelled Names over the Telephone. *Proceedings ICSLP'96*, Philadelphia, PA, USA, pp. 346-349.

[3] Jouvet D., Lainé A., Monné J. & Gagnoulet C. (1993), Speaker-Independent Spelling Recognition over the Telephone. *Proceedings ICASSP'93*, Minneapolis, Minnesota, USA, vol. II, p. 235.

[4] Jouvet D., Lokbani M. N. & Monné J. (1993), Application of the N-best Solutions Algorithm to Speaker-Independent Spelling Recognition over the Telephone. *Proceedings EUROSPEECH'93*, Berlin, Germany, pp. 2081-2084.

[5] Jouvet D. & Monné J. (1999), Recognition of Spelled Names over the Telephone and Rejection of Data out of the Spelling Lexicon. *Proceedings EUROSPEECH'99*, Budapest, Hungary, pp. 283-286.

[6] Junqua J.-C., Valente S., Fohr D. & Mari J.-F. (1995), An N-best Strategy, Dynamic Grammars and Selectively Trained Neural Networks for Real-Time Recognition of Continuously Spelled Names over the Telephone. *Proceedings ICASSP'95*, Detroit, Michigan, USA, pp. 852-855.