A continuous speech recognition approach for the design of a dictation machine

K.Smaïli, F.Charpillet, JM.Pierrel, JP.Haton

CRIN / INRIA Lorraine BP 239 54506 Vandœuvre
email: smaili@loria.crin.fr

Abstract

The oral entry of texts (dictation machine) remains an important potential field of application for automatic speech recognition. The RFIA group of CRIN / INRIA has been investigating this research area for the French language during the past ten years. We propose in this paper a general presentation of the present state of our MAUD system which is based upon four major interacting components: an acoustic phonetic decoder, a lexical component, a linguistic model and a user interface.

Keywords: dictation machine, linguistic model, blackboard

I Introduction

The role of the MAUD (Machine AUtomatique à Dicter) project is to give any French speaker the possibility to dictate his letters or reports to his workstation. MAUD uses a vocabulary made up of the most common words of French and contains 37 000 words (including inflections). A problem may be mentionned when a large vocabulary is used: a great number of possible sentences which match perfectly with phonemic lattice can be recognized. Therefore, it is necessary to design a linguistic model which efficiently discards the wrong interpretations. In this paper, we discuss some ideas concerning the design of MAUD and present the main characteristics and original features of this system. Part of this work has been carried out in the framework of the Multiworks ESPRIT project [Smaïli 90b].

II MAUD architecture

MAUD is based upon three major components :

- an acoustic-phonetic decoder which provides a phonemic lattice enriched with phoneme likelihoods. This decoder is based on our APHODEX expert system [Carbonell 89];
- a lexical component which is intended to extract the most likely words. This component represents the kernel of the recognition and understanding process since it is called upon at three different processing levels, i.e. semantic, syntactic and phonological;
- a linguistic model which is in charge of predicting the next words to verify in the sentence and of selecting syntactically correct interpretations. The linguistic model of MAUD is a probabilistic one based on bi- and tri-class frequencies but it also uses linguistic concepts.

II.1 The lexical component

Any man-machine communication system in natural language oral or written, uses a more or less complex lexicon. However, it is only when the vocabularies reach some few thousand words, that it becomes necessary to develop real lexical components such as those planned for the fifth generation computers [Simons 83]. This module constitutes one of the most important components of our system. As a mere matter of convenience, it can be considered that the lexicon intervenes at four different fundamental levels: semantic, syntactic, morphological and phonological. We will now describe the various aspects of this lexical component.

II.1.1 Lexical classification

From the 8 elementary grammatical classes of French, we got 201 refined syntactic-semantic classes. Increasing the number of classes allows us to accurately define the possible syntactic-semantic contexts. The rule applied to classify the dictionary is the following: one word belongs to a class if any other word of the same class can be substituted to it without any change in the syntactic structure of the sentence.

II.1.2 The vocabulary

The system works presently with a 7 000 entries dictionary and the final version of the system will run with a 37 000 entries dictionary. Both dictionnaries are extracted from the most common words of the BDLEX lexical data base [Pérennou 87], all conjugations and all types of variations in gender or number being included. The problem of words which have several syntactic functions is solved by the duplication of the entry in each class to which it belongs. Each entry of our dictionary comprises the following fields :

- an accentuated graph representing the orthographic form of the entry;
- the phonetic form associated with the orthographic representation;
- a phonological end point: boundaries of words are often subject to particular alterations such as latent consonant, denasalization, etc. This field contains phonological knowledge for taking into account these phenomena;
- the syntactic-semantic class;
- syntactic features: in this field we find the number and gender of the adjective or noun entry, person and conjugation time for the verbs.

Recently, we have added two important types of information which are not used in the present version of the system. This information consist of the relative and absolute word frequencies. The relative frequency is obtained from the labelled corpus. As for the absolute frequency, we have extracted it from an orthographical list of French [Catach 84].

II.1.3 Lexical access

The lexical search is extremely difficult because the phonetic transcription may contain substitution, omissions and insertion errors and because a speaker uses a system of phonological rules to modify and more specially to simplify the pronounciation of individual words in certain contexts. Since the dictionary is accessed very frequently, it is necessary to find a strategy which allows for a fast access to all the words in the dictionary. In MAUD, this process is controlled by the syntactic component. In other words, when the system tries to identify a word, it looks for it only in those syntactic-semantic classes which can follow the current syntactic-semantic class. For that purpose all the words which belong to the same syntactic class have been grouped in the same zone. As for the problem of identifying phonetic words, we use some phonetic rules, among them some are represented as a classical production rules and the others are compiled in the four following matrices :

1 / A confusion matrix in which the likelihood of possible confusions between phonemes is stored. It is an inter-phoneme confusion matrix built independently of the acoustic-phonetic decoder of the system. Values in this matrix are between 0 and 1, 1 means that there is no confusion between the two phonemes to compare.

2 / The confusion matrix of the system : this confusion matrix is dependent upon the acoustic-phonetic decoder.

3 / The omission matrix : in the same idea, the matrix gives the probability to elide a phoneme. This matrix is used when the lexicon component emits an omission hypothesis.

4 / The insertion matrix : it contains the probabilities that the acoustic-phonetic transcriber inserts a given phoneme.

Results contained in the three last matrices were obtained from the decoding a corpus of 14 500 phonemes.

III.2 The linguistic model

The goal of syntax in speech recognition is to participate in the choice of word candidates in a sentence and in the elimination of false interpretations which may be generated. Two possible approaches are possible to design such model. If the domain of the application is restricted, all the phenomena concerning the application could be modelized. Consequently, the solution in this case is to use a grammar [Pierrel 87]. On the other hand, if we are interested in natural language, the problem becomes much harder. As a matter of fact, linguists have worked for a long time in this area trying to develop a theory and / or linguistic models. The approach used in a system which processes natural language is the probabilistic model. At the present time, it is the only practical way which allows a speech recognition system to use efficiently a syntactic-semantic model. For any given language, when one takes a string of words which has a length equal to n (n>1), there are p manners ($p \leq n!$) for arranging a word string of length n ($n > 1$) in order to build up a correct sentence. Rigourously, the probability for a word to be produced, should depend conditionally upon the whole past string. But in practice, this probability would be difficult to estimate and would require too much storage. A possible solution as suggested by Jelinek is to take into account only the two last words. These three words constitute a trigram and when only the last word is taken into account, we will speak about bigram [Jelinek 82].

II.2.1 Presentation of the model

The linguistic model used in MAUD is strongly dependent on biclass and triclass frequencies, but it also uses some other concepts which are specific to French language.

This model is based upon seven syntactic-semantic modules :

- a syntactic-semantic preprocessor (SSP) which collaborates with the lexical component to filter the first flow of words. The PSS is exclusively based upon a biclass model. The PSS allows us to have rapidly a first interpretation of the APL;
- a syntactic-semantic processor which uses the list of words given by the PSS and provides a list of syntactic pattern (SPA) (these patterns represent the best syntactic path given by the stochastic module). A plausibility coefficient is assigned to each SPA. This coefficient is computed as follows:

$$CP = \prod_{i=2}^{n} F^{C_i}_{C_{i-2}C_{i-1}} + a * F^{C_i}_{C_{i-1}} + b$$

a and b being two constants which vary according to the number of biclasses and triclasses;

- a filter which cancels the SPA which have not a grammatical availability. This filter is based on frequency of metaclasses, where the metaclass is merely an existing triclass;
- a sentence generator which uses the SPA and the list of words and provides in the output the most likely sentences;
- a system based on grammatical rules, which allows to take into account the syntactic-semantic phenomena which cannot be processed only by the frequency model;
- a system based on phonological rules which is specific to the French language. This system allows the elimination of wrong phonological sentences;
- finally, a sentence filter made up of more complex grammatical rules which can eliminate more sentences than the one mentioned above. This filter could be extended to act as a syntactic corrector.

II.2.2 The design of the probabilistic model

To collect the frequencies of the grammatical biclasses and triclasses, an important corpus of text has to be labelled, i.e. the grammatical classes have to be assigned to each word of the text. The dictionary used for the labelling was the whole dictionary of BDLEX (230 000 entries). The labelling has first been done manually with a text of 3 500 words. Statistics have been kept and stored in the appropriate precedence matrices. From the computed probabilities, we have labelled automatically a scientific book [Pierrel 87] of 23 500 words and finally the labelling was done on a text of 350 000 words. The corpus has been extracted from a regional daily newspaper.

From the results obtained[Smaïli 90a], we have noticed that the triclass curve is practically in a linear progression, this means that much text remains to label. Two solutions are possible for solving the convergence problem. The first one consists in labelling more corpus from other sources. The second solution is to eliminate the noise from the model (the human errors of labelling) and to use a classification system which must be more constrained. Since, the first solution is very costly, we opted for the second solution. Another idea has been used to find a solution to the convergence problem of the syntactic model. It concerns the use of a triple classification. Since it is difficult to have a convergent model with 201 syntactic-semantic classes, we will combine three different models. The first model is based on 201 classes, the second model on 138 classes and the third one on 38 classes. These two last models use subsets of classes from the first model. In this case when a model fails, the system passes control to another model with a narrower classification. For that, we have calculated the biclass and the triclass statistics which are

presented in figure 1.

Number of words / Number of classes	58 614	126 703	196 675	269 775	338 302	373 761
38	821 / 4431	979 / 6212	1061 / 7536	1124 / 8516	1166 / 9217	1181 / 9508
108	2022 / 9115	2677 / 14 083	3188 / 17 956	3408 / 21 012	3628 / 23 360	3724 / 24 436
201	2814 / 10654	3959 / 17 313	4806 / 22 708	5437 / 27 266	5890 / 30 817	6103 / 32 504

Fig 1

Statistics concerning the triple classification (first value: biclass, second value: triclass)

II.3 The editor associated to the dictation machine (DIM)

II.3.1 The necessity of using an editor

When one dictates a letter, frequently one needs to change a word by another or more simply to eliminate it. A simple ASR cannot do that. It is absolutely necessary in this case, to add some tools to the ASR to have a system which allows modifications on the recognized sentences. Two approaches are possible: The first one consists of the use of a speech editor. The user will have a list of keywords that he can use while dictating a text. The second one is based on a graphical editor which is accessible by the user at every moment. Choosing a speech editor presents some problems in continuous speech recognition. Indeed, the recognition system must detect from the flow of speech the keywords of the editor language. The second approach is much more accessible and especially in multi-media workstation. It is sufficient to allow the interruption of the recognition system and to pass control to the user.Thus, it is sufficient to drag the mouse to the function he needs and choose it like in Dragon system.

II.3.2 The editor of MAUD

Because this system will be integrated in a multi-media workstation, we have thought that it is more interesting to use the second approach. The present editor is a multi-window editor. All its functions are selected by the mouse. Two kinds of functions are available :

- the user functions, such as the choice of a word among others, the addition of words, the backtrack to the previous words and so on,
- the designer functions such as displaying word scores and intermediary results, asking the system why a word does not belong to the list of recognized words and so on.

The user or the designer can interrupt the system at every moment to start the command editor.

III Some results of recognition

The transcription system made up of the linguistic model, the decoding modules, the graphic interface and the whole source knowledge has been developed on a UNIX Concurrent system and implemented recently on a SUN station. Because the system has been tested on only 7 phonetically balanced sentences [Combescure 81] uttered by 4 male speakers, we can not give a complete assessment of our system. These sentences have been decoded by Aphodex. Each sentence has been segmented in phonetic units. On average the decoder has proposed 4 phonemes for each segment and has made 8 errors of interpretation for each sentence: 2 omissions, 2 insertions and 4 substitutions. The sentences are made up of 21 phonemes on the average.

III.1 Examples

Figure 2 presents a copy of a working session screen. The tested sentence listed is made up of 7 words uttered continuously. In the present version of the system, a manual word segmentation has been done. Among the 49 words which compose our corpus 37 have been recognized without any user's intervention with an average of 15 candidates for each word (for more details see [Smaïli 90a].

III.2 Discussion

We can note that the SSP has recognized all the words of this sentence. No external intervention was necessary. The uttered sentence has been recognized but some others too. A lot of sentences could not be eliminated except by the introduction of very constrained semantic information. Other sentences could not be eliminated by the introduction of such information but necessitable improvement of the acoustic-phonetic decoder.

At present, we are working in two different ways. The first one consists in enlarging our system to the continuous speech. The system is under development and it is based on the following concepts:

- All the available linguistic knowledge is used at the same time.
- At each step of the recognition, only the n highest scored words are kept.
- The recognized words are stored in a framework tree. Each node of this tree contains all the words which have the same length. Inside the node structure, the recognized words are grouped in accordance to their syntactic classes.
- Each node of this tree is in relation with all its successors and not with nodes belonging to the same level.

The second research area is related to the use of a blackboard model for implementing MAUD. We use the ATOME blackboard tool developed in our group [Lâasri 88] to design and to implement the new version of MAUD. This new architecture is made up of two blackboards: a classical blackboard partitionned into four levels which correspond to the different levels of the interpretation process, and an historical blackboard which contains the interpretations which have not been kept in the first blackboard. Figure 3 shows the overall architecture of the system.

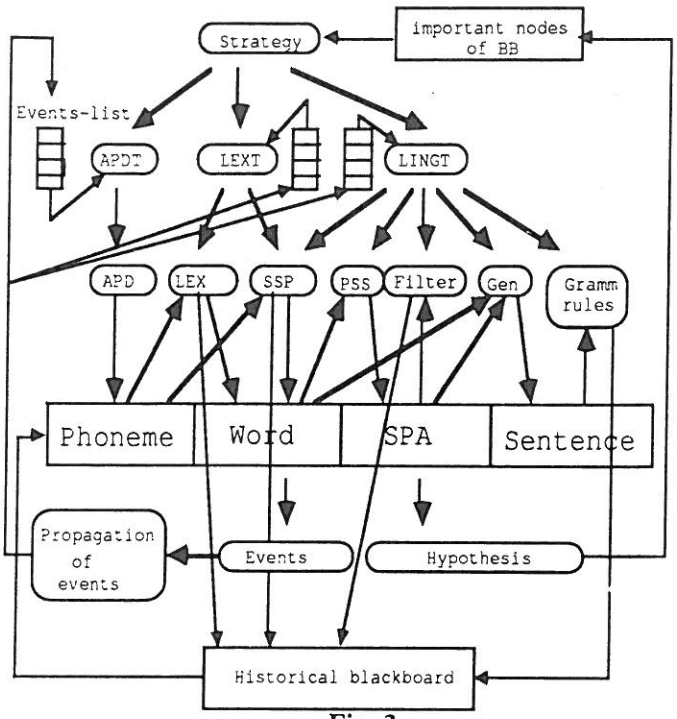

Fig 3

The new architecture of MAUD based on a blackboard approach

IV Conclusion

In this paper, we have presented the MAUD prototype system for the dictation of texts in natural language with a large vocabulary. Results are encouraging and will be improved in the next few months. In the next version of the system a triple lexical classification will be used. The two frequencies will be integrated in the computation of the global score sentence. A lot of work has to be done at the acoustic-phonetic level to make the phonetic lattice more reliable. Indeed, the addition of specific rules in the acoustic-phonetic base will improve the recognition of some phonemes such as /l/ /m/ /n/ /r/. The use of several complementary processors and knowledge sources for phonetic decoding will highly improve the overall performances. MAUD is based on an advanced blackboard architecture thus allowing for thus a combination of processors and for the easy test of different interpretation strategies.

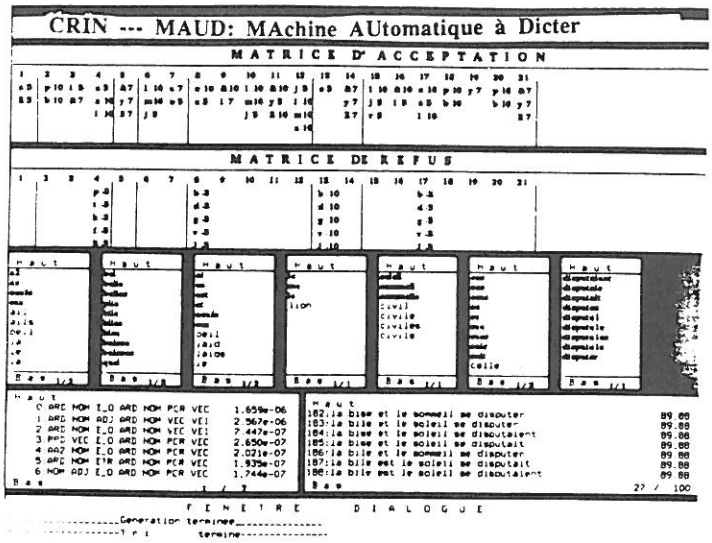

Fig 2

Decoding of the sentence *" la bise et le soleil se disputaient "(the north wind and the sun were quarrelling)*

References

[Catach 84]: N.Catach " Les listes orthographiques de base du français", *Nathan recherche, 1984.*

[Carbonell 89]: N.Carbonell, D.Fohr, JP.Haton " APHODEX an acoustic-phonetic decoding expert system", *International journal of pattern recognition and artificial intelligence (IJPRAI), vol 1, n°2,1987.*

[Combescure 81]: P.Combescure " Vingt listes de dix phrases phonétiquement équilibrées",*Revue d'acoustique, 14, N° 56,1981.*

[Jelinek 82]: F.Jelinek, R.L Mercer, LR.Bahl " Continuous speech recognition: statistical methods", *CSR group, IBM TJ Watson Research Center, 1982.*

[Lâasri 88]: H.Lâasri, B.Maître, T.Mondot, F.Charpillet, JP.Haton " ATOME: A blackboard architecture with temporal and hypothetical reasoning", *Proceedings of the 8th European conference on artificial intelligence (ECAI), 1988.*

[Pérennou 87]: G.Pérennou "Bdlex : a data and cognition base of spoken French" *Research and development in language processing, JP.Haton et G.Pérennou eds INRIA, 1987.*

[Pierrel 87]: JM.Pierrel " Dialogue oral homme machine", *HERMES, 1987.*

[Simons 83] : GL.Simons "Toward fifth generation computers", *NCC publication , 1983.*

[Smaïli 90a]: K.Smaïli, F.Charpillet, JM.Pierrel, JP.Haton " Idées et concepts de réalisation d'une machine à dicter destinée aux grands vocabulaires", *Actes des XVIIIèmes journées d'étude sur la parole, Montréal 1990.*

[Smaïli 90b]: K.Smaïli, Y.Laprie " Continuous speech recognition: application for dictation machine", *Multiworks report, ESP 2105, 1990.*