Towards an oral interface for data entry: The MAUD System

Dominique Fohr, J.-P Haton, Jean-François Mari, Kamel Smaïli, Imed Zitouni

▶ To cite this version:

HAL Id: hal-01112900
https://hal.science/hal-01112900

Submitted on 4 Feb 2015

Towards an oral interface for data entry: The MAUD System

Fohr, J.-P. Haton, J.-F. Mari, K. Smaïli and I. Zitouni

CRIN-CNRS/ INRIA Lorraine

BP239 54506 Vandoeuvre Lès-Nancy France

e-mail: {fohr, jph, jfmari,smaili, zitouni}@loria.fr

Abstract

This paper deals with use of *MAUD* as an oral interface for data entry and the description of the speech component of this system. The interface of *MAUD* has to combine voice-driven and keyboard dialogue in order to allow the user to use both keyboard and voice. The speech recognition system participated at the AUPELF-UREF evaluation of dictation machines. *MAUD* uses a vocabulary of 20 000 words. The mode of speech is continuous, and the language model was built with corpora of over than 48 million of words extracted from French newspapers.

I/ Introduction

The oral entry of texts with a dictation machine is an important field of application of automatic speech recognition, and the first existing prototypes are encouraging the development of such applications. However, some research effort has still to be carried out in several areas: lexicon, linguistics and in the design of a user friendly interface. When one uses a keyboard to enter texts, one needs tools to remove inadequate words, to skip a line, to save a file and, etc. In fact, we need a real editor for processing words. For oral entry of texts, there are several fundamental problems in speech processing, and the one concerned with human factors have not yet been sufficient addressed. In the near future, the oral entry of texts has to be used, as any other basic software, by a large population. In order to make the use of the dictation machine friendly, we have to take into account the characteristics of the potential users. The system has to make the difference between the data and the commands entry without changing oral entry mode and without any restriction for the user. The oral communication is more natural for human being than any other channel of communication. That is why the oral entry of data can be considered under some considerations as a friendly man-machine process of communication. The user dictates a text, and during this process, he/she communicates with the system by a command language. A command is recognised and executed by the system, and the resulting effect is displayed on the screen. Our main aim is to design a multimodal dictation machine. The multimodal interface of our system has to combine voice-driven and keyboard dialogue which allows the user to use both keyboard and voice for processing words and for issuing commands. MAUD is currently used in voice activated typewriter system capable of editing cardiology reports and in as an oral interface for navigating on the WEB.

In this paper, we describe the *MAUD* dictation machine of the RFIA group. The new version of our prototype is based on a second-order Hidden Markov Model (HMM 2) and on an hybrid language model which uses both a statistical and a formal grammar. At present, the system uses a large vocabulary of 20 000 most frequent words extracted from the French newspaper Le Monde. It is trained acoustically on the Bref corpus and linguistically on more than 48 million of words extracted from the newspapers : « Le Monde » and « Le Monde Diplomatique ». This system is based fundamentally on a stochastic approach. At present, the system is not able to run on real time, that is why the ergonomic interface has not yet been included.

II/ Description of the MAUD System

The system *MAUD*, is a 20k words continuous dictation system using a stochastic model in both acoustic and linguistic level. Before going further in the description of the system, we give details about the language model component used by MAUD.

II.1/ MAUD Language Model

The language model of *MAUD* is made up of a combination of a stochastic model and a formal grammar. During the recognition process, the stochastic language model suggests partial hypothetical sentences, and the unification grammar checks their correctness. If a rule can be applied, the acceptance of the sentence is decided by the formal component, otherwise the decision is left to the stochastic component.

II.1.1/ Stochastic component

Even though an explicit formal grammar for natural language is more expressive, stochastic n-gram language models are still preferred for building operational large vocabulary speech recognition systems, since they can be trained automatically on large corpora. In stochastic language models, the basic idea is to find the most likely words that matches the acoustic signal. For that purpose, it is necessary to collect the probabilities of a word in all possible contexts. In our model, we use a combination of n-grams and n-class in accordance to the following formula :

$$P(w_1 w_2 \cdots w_n) = \prod_{i=1}^{n} P(c_i / c_{i-2} c_{i-1}) P(w_i / c_i)$$

Where $w_1 w_2 ... w_n$ is the sequence of words to recogniser. C_i determine the syntactic category of W_i. In order to estimate the probability $P(c_i / c_{i-2} c_{i-1})$, we need to tag each word of the training corpus. Consequently the dictionary of the application need a syntactic field for each entry. This involves that some words have to be duplicate if they appear in more than one class. From the eighth elementary grammatical classes of French, we build up about 230 classes including punctuation [1]. To learn the language model, we use this high number of classes because in order to build a model which is sufficiently predictive and highly selective. We used, in a first step, a corpus of 0,5 million words which has been accurately labelled. The model learned during this process has been used in a second step to automatically label each word of « Le Monde » and « Le Monde Diplomatique» corpora. From these corpora, we obtained a model made up of about 17500 bi-class and 245000 tri-

classes. In spite of the amount data of the corpora, we need to use a smoothing method to take into account the unseen and the uncommon events [2].

II.1.2/ Formal component

It is well known that such a stochastic model does not take into account complex linguistic phenomena, since it operates only on a short « history » of a word (generally 2 or 1 adjacent words). As a result, serious problems are encountered when n-gram models are used in a real continuous speech recognition system. As a matter of fact, even if *n* is high and the total amount of data is sufficient to estimate probabilities, this kind of language model cannot take into account all the phenomena of a natural language. This is especially the case when the agreement restrictions are between two words which are separated by more than *n* words. To handle such phenomena, we have enlarged the *MAUD* language model by a unification grammar. This grammar is implemented as an *ATN*, accurately captures phenomena of the French language, through the use of features like the following : Gender, Number, Person, Verb-Form, ...

II.2/ Sentences recognition by stochastic processes

The answer of *MAUD* is deducted from many treatments on the signal and its successive representations (fig. 1) [3]. This system proceed in 4 steps:

- Gender identification;
- Building a world lattice with the use of a second order Hidden Markov Model (HMM2) and a bigram language model;
- Building the N-best sentences with by using a trigram language model and the word lattice;
- Filtering this sentences by using a 2-class, a 3-class language modeling and a unification grammar.
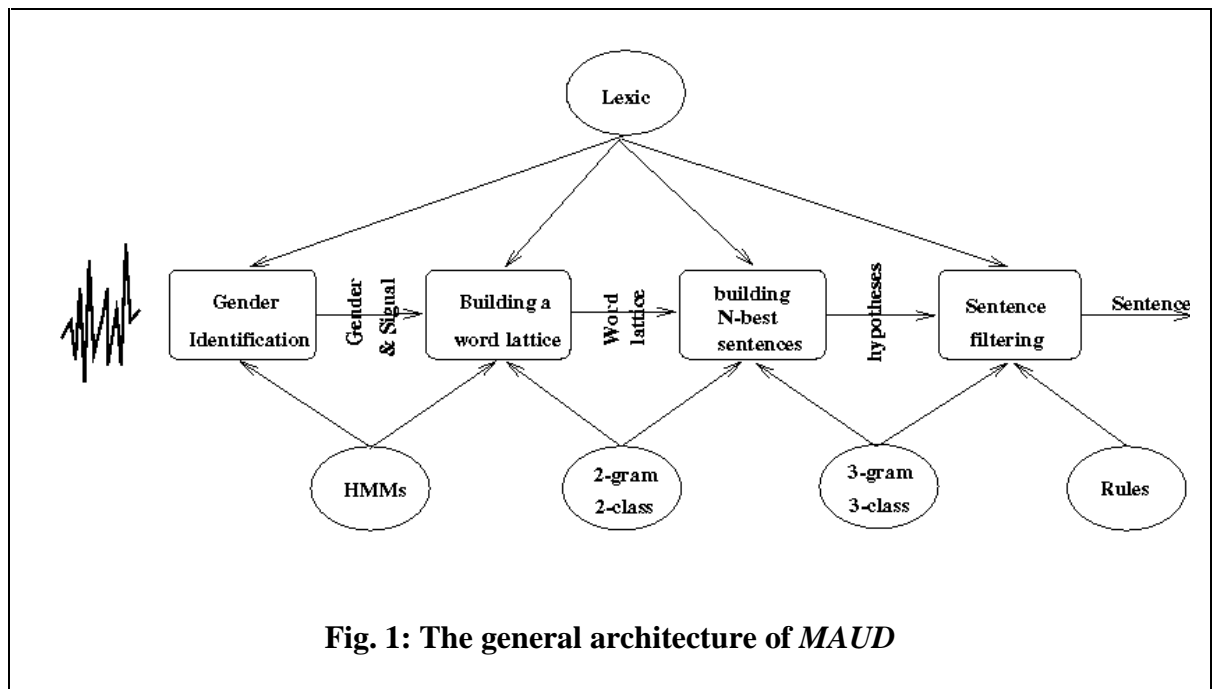

Fig. 1: The general architecture of *MAUD*

II.2.1/ Gender Identification

The signal is parametrised with 12 MFCC coefficients, with their first and second derivatives. Each frame is computed every 8 ms. Two recognition systems are used in parallel: the first one uses 35 context independent phonetic models constructed for male speakers, while the second one uses a similar model dedicated for female speakers. The best likelihood algorithm determines the speaker gender. In this step, in order to accelerate the identification process, we use a very narrow beam search for recognising.

II.2.2/ Building a word lattice

The goal of this step is to build a word lattice from the speech signal. For that, a context dependent acoustic models are used according to the results of the first step. Each phoneme in context (242 male diphones and 261 female diphones) is modelled by a second order Markov model with 3 states (HMM2). This HMM2 has been trained on a wide corpus of labelled speech. Therefore all the contextual variations have been captured and modelled by means of probability density functions. The second order Markov chain governs the segmentation in stationnary and transitional areas. Each word in the lexicon is represented by the concatenation of the HMM2 diphones which compose it. To obtain the word lattice, our system uses a slightly modified block-Viterbi algorithm [4] which takes into account the usual phonological alterations (deletion, liaisons,...) of spoken French language and a bigram language model. The goal of the bigram language model is to predict the following word in accordance with the previous one.

II.2.3/ N-best sentences building

This step builds the N-best sentences using the word lattice and a rigram language model. The goal of the trigram language model is to predict the following word in accordance with the two previous words. The trigrams appear with different probabilities that can be estimated by their frequencies in a training corpus of written words. Many trigrams have never been seen; in this case, the trigram probability is approximated by various back-off methods based on the probabilities of the bigrams. A beam search is used accounting for both acoustic scores and alignment calculated during the previous step (no acoustic recalculation is required). The result is a list of ordered sentences. Each sentence's score is obtained by a combination of the acoustic and the language model notes.

II.2.4/ Sentence filtering

Sentence filtering is carried out using a probabilistic model based on 2-class and 3-class improved by grammatical rules which reduce the ambiguities inherent to a positional classic model. These grammatical rules are based on the unification grammar formalism. The aim of this grammar is to take into account phenomena such as agreement in gender and number [2]. The sentences outputted in the previous step are syntactically labelled and are scored again in order to keep the *N* best sentences according to this formula:

$$\alpha P(c_i/c_{i-2}\ c_{i-1}) + \beta P(c_i/c_{i-1}) + \gamma P(c_i) + \theta$$

where $\alpha + \beta + \gamma + \theta = 1$, $P(c_i/c_{i-n}...c_{i-1})$ is the n-class probability and c_i is the class assigned at the word w_i. This formula guarantees a nonzero estimate for the 3-class

language model. The N best sentences kept are examined by the unification grammar in order to eliminate the sentences which do not respect the agreement grammatical rules.

The answer of *MAUD* system is the best sentence accepted by the unification grammar.

III/ Results

MAUD, as a stand alone system has been assessed in the first AUPELF-UREF campaign. All the laboratories which participated to the test used the same corpora for testing their systems. Three hundred sentences drawn from the newspaper « Le Monde » have defined the test corpus on which 5 systems have been assessed. MAUD got the second place with 67% word recognition rate. The main issue is to handle the out of vocabulary words.

IV/ Conclusion

We have described a speaker independent word recognition system. The stochastic principles have been used at the three levels of MAUD : phonetic, lexicon and grammar. For achieving more robust estimation of probabilities, the words are clustered on 230 classes that specify similar grammatical behaviours. We are currently integrating this system into a multimodal station in order to access internet for navigating on the WEB.

References

[1] K. Smaili, F. Charpillet and J. -P. Haton « A new Algorithm for Word Classification based on an Improved Simulated Annealing Technique » 5th International Conference on the Cognitive Science of Natural Language Processing », Dublin, 1996.

[2] K. Smaïli, I. Zitouni, F. Charpillet and J.-P. Haton « An Hybrid Language Model For a Continuous Dictation Prototype », 5th European Conference on Speech Communication and Technology, Vol 5, PP 2727, Rhodes-Greece, 1997

[3]   D.Fohr, J.P.Haton, J.F.Mari, K.Smaïli and I.Zitouni   «*MAUD*: Un prototype de machine à dicter vocale», 1ères JST Francil 1997, PP25-30, Avignon-France, 1997.

[4] A. Krioule, J.-F. Mari and J.-P. Haton « Some Improvements in Speech Recognition Algorithms Based on HMM. Proceedings IEEE ICASSP 90, PP 545-548, Albuquerque, 1990