INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team cordial

Man-machine oral and multimodal communication

Rennes

THEME COG

Activity Report

2006

Table of contents

1. Team

Head of project (Responsable scientifique)

Laurent Miclet [Faculty member (Professor), Enssat]

Administrative staff

Joëlle Thépault [Administrative assistant, Enssat]

Faculty members (University of Rennes 1)

Nelly Barbot [Associate Professor, Enssat]
Vincent Barreaud [Associate Professor, Enssat, from September the 1st, 2006]
Olivier Boëffard [Professor, Enssat]
Arnaud Delhay [Associate Professor, IUT]
Marc Guyomard [Professor, Enssat]
Jacques Siroux [Professor, IUT]

Ph. D. Students

Pierre Alain [bourse INRIA, until September the 31th, 2006]
Sabri Bayoudh [bourse INRIA, from October the 1st, 2004]
Ali Choumane [bourse Région, from October the 1st, 2005]
Karl DeVooght [FTR&D, from January the 1st, 2005]
Josselin Huaulmé [CIFRE Télisma, from October the 1st, 2003]
Damien Lolive [bourse MENRT, from october the 1st, 2005]
Larbi Mesbahi [bourse Région, from november the 1st, 2006]
Salma Mouline [FTR&D, from January the 1st, 2003]
Sylvie Saget [bourse Région, from October the 15th, 2003]

Post Doctoral Students

Sébastien Carbini [Post doc INRIA from November the 1st, 2006]
Laure Charonnat [PostDoc ANR VIVOS, from October the 1st, 2006]

Technical staff

Laurent Blin [Research engineer, FNADT, from December the 1st, 2005]
Gaëlle Vidal [Assistant engineer, ANR VIVOS, from July the 10th, 2006]

Associate members

Jean-Christophe Pettier [Associate Professor, Enssat]

2. Overall Objectives

2.1. Overall Objectives

The Cordial project explores several aspects of multimodal man-machine interfaces, with speech components. Its objectives are both theoretical and practical : on the one hand, no natural dialogue system can be designed without an understanding and a theory of the dialogic activity. On the other hand, the development and the test of real systems allow the evaluation of the models and the constitution of corpora.

The conception of a man-machine interface has to take into account the communication habits of the users, which have been developed within interpersonnal communication. This is particularly true for interfaces using speech, which is a medium quite performant and spontaneous. Users have great difficulties to communicate through an oral dialogue with a machine having a speech interface of mediocre quality. The dialogue phenomena are complex [28], involving spontaneous speech understanding, strong use of pragmatics in the dialogue process, prosodic effects, etc.

Dialogue modeling

When multimodal dialogue is involved, the interference between speech phenomena and tactile actions or mouse clicks brings up problems of interpreting the coordination of the different actions of the user.

When a user makes a communication action towards the dialogue system, he certainly has an intention; but often, this intention is not explicitely present in the communication. A major problem for the system is to extract it, in order to be able to give a satisfactory answer. This requires a theory coping with the notions of intention, background knowledge, communication between agents, etc. We modelize the dialogue phenomena by using the concepts of speech acts and dialogue acts, and we consider that a sequence of exchanges can be analyzed as the result of a planning. This model gives a satisfactory modeling of many phenomena in real dialogues, such as the coordination between different negotiation phases or the management of the user's knowledge base.

However, several points are not straightforwardly modeled in such a theory: parts of the dialogue do not carry any obvious intention or errors in understanding may mistake the planner, etc. Moreover, the extraction of the dialogue acts from the speech of the user is a complex problem, as is also the restitution of the dialogue acts of the system into synthetic speech.

Machine learning

In addition to the modeling of the core of dialogue phenomena, the Cordial project has also a particular interest in machine learning from corpora at different stages of a dialogue system. It covers the extraction of semantics from the outputs of a speech recognizer. It also tackles the problems of constructing the prosody of the machine synthetic speech or helping the dialogue engine to compute an answer. Machine learning [2] is a field with many different facets, spanning from the inference of finite automata from symbolic sequences data to the optimization of parameters in stochastic processes. Our research in this field makes use of quite different techniques, reflecting the variety of the data and of the models met at the different stages of a dialogue system.

Speech synthesis.

The front-end part of an oral dialogue system consists in a text generator producing a sequence of words, corresponding to the message to be emitted [1]. This part of text is then converted into an oral message through a speech synthesis system [35], [84], [87]. The text to speech technology has still to increase its quality, especially in a dialogue environment, in order to produce a speech as natural as possible. This can be made partly in producing a good prosody, but also in working on the quality of the acoustic signal. In a dialogue system, the speech synthesizer can be given extra information on the semantics and the pragmatics of the situation and therefore produce a speech with special effects: delivering information, stressing on a detail, insisting on a misunderstanding, repeating an information, etc. This can influence the way the message has to be delivered, especially concerning its prosody. In the same way, a text-to-speech system built from the target application can lead to significative improvements [40]. Lastly, an interesting problem is to diversify the synthetic voices, without having to record and index a new corpus. Acoustic voice transformation techniques can be used, but changing a voice into another requires also a modification of the segmental and prosodic characteristics.

3. Scientific Foundations

3.1. Introduction

Our activities are distributed into four complementary domains. The first one is concerned with both *the coding and the structure of interaction*. It also deals with the applications. The second one deals with *multimodality and system prototyping* (architecture and evaluation). The third one is concerned with *machine learning* techniques and their application to dialogue phenomena and speech technologies. The last area deals with *speech synthesis* adapted to dialogue.

3.2. Dialogue and modeling

Keywords: *Speech Acts*, *plan recognition*, *planning*.

We use a family of dialogue models based on speech acts plans. This modeling takes into account the general framework of communication and makes easier the implementation on computer. But it does not solve some problems like extracting speech acts from utterances or the integration of different information sources and miscommunication between participants.

Man-Machine interaction can be seen as a sequence of particular actions: speech acts [33], [76] called in our context *dialogue acts* which support both the function of the act in the dialogue (for example: requesting, querying, ...) and a propositional content (for example: the theme of the query). These acts can also be characterized by their conditions of use which are concerned with the mental states of the participants (intention, knowledge, belief). The most accurate computerized model is the planning operator [31], [61] in which preconditions and constraints as well as effects of an act can be represented. For example, the act to ask for somebody to perform one action can be modeled as follows:

Request(Speaker, Hearer, Action(A))

precondition-intention: *Want(Speaker, Request(Speaker, Hearer, Action(A)))*

precondition preparatory: *Want(Speaker, Action(A))*

Body: *Mutual Belief(Hearer, Speaker, Want(Speaker, Action(A)))*

effect: *Want(Hearer, Action(A))*

This can be interpreted as: when an agent wants that its listener performs an action A, it can use the action labeled *Request* whose goal is to build up a consensus between participants in order to perform A. Realizing this consensus is the task of another action which is not described here. The set of actions which are necessary for reaching a goal is named a plan. This approach makes the hypothesis that each dialogue partner participates in the realization of the other's plan. This dialogue act modeling allows to consider several types of automatic reasoning in order to manage the dialogue. The first one is concerned with the contextual understanding of user's utterances by means of a mechanism so-called *plan recognition*. It aims at rebuilding a part of the other participant's plan; if this part is correctly identified, it allows to give an account of the explicit motivations and believes of the other participant. A second process aims at computing a relevant response by means of a planning mechanism which is able, because of the nature of the modeling itself, to take into account the known information and the possible misunderstandings. This type of modeling makes easier the implementation in some simple situations but does not deal with some important problems in various fields.

3.2.1. Dialog act extraction

The first problem is to translate the sentence uttered by the user into a dialogue act. This process is not a simple transcoding problem. It is necessary to take into account altogether a large collection of knowledge (mental states, presuppositions, prosody, ...) as well as some indices present in the sentence (syntactic structure, lexical items, ...). In addition, the surface form of speech sentences contents a lot of irregularities (problems of performance) which complicates the speech recognition task as well as the understanding and interpretation tasks.

3.2.2. System modeling

The second problem takes place in the use of the planning formalism [8] in order to associate three points of view [69]: the one of the application, the one of the main dialogue (which is concerned with user's intentions towards the application) and the one of the dialogue management (meta dialogue and phatic dialogue). Some partial solutions have been found [61] but they are not well adapted to data management applications (querying data base) or applications which allow several parallel tasks and the processing of certain functions for communication management. A possible approach to deal with this problem could be a multi-agent modeling. Indeed, this conceptual framework allows to combine *a priori* exclusive models and dialogue contexts in order to increase the number of dialogue problems dealt with. Therefore, the problem is partly moved from dialogue modeling towards integration modeling.

3.2.3. Communication errors

The third problem arises frequently in interaction: it concerns bad communication. Each of the two participants (*i.e.* the human and the system) can indeed have some erroneous knowledge about the application, about the other participant's abilities and about current references used to point out objects during the interaction. One error which concerns this information, may (in the long or short run) leads to a failure, i.e. to an impossibility for the system to satisfy the user. Detecting and dealing with these errors basically requires a characterization process and a plan based modeling.

3.2.4. Application modeling

In an interactive system, the application has to behave as an active component. In current systems, the application modeling affords two types of main defaults. The task model may be too rigid (for example: plans in the systems for transmitting information) constraining too heavily the user's initiative. The task model may also be based on constraints (as in CAD application), allowing in this way a user's activity more free but causing a lack of co-operation for helping the user to reach its goal. We believe that the task model has to include the following elements: data and their ontology, knowledge about the use of data (operating modes) and the interface with the rest of the system. Lastly, the modeling has to be designed in order to make easier the changing of the task.

3.3. System and multimodality

Keywords: *educational software*, *multimodality*, *reference*, *teaching and learning languages*.

We are studying an additional modality, a tactile screen, in order to avoid some of the problems coming from using only speech. The problems to deal with due to this new modality are concerned with integrating messages coming from the different channels, processing of references as well as evaluating systems. The aim of the Ordictée study is to design and to develop educational software for helping to teach and to learn languages.

The use of speech technologies in interactive systems raises problems and difficulties spanning from the design of complete softwares (including the research of the task) to the architecture design, including a particularly good quality speech synthesis and the introduction of a new modality.

3.3.1. Multimodal interactive system

Human communication is seldom monomodal: gesture and speech are often used jointly because of functional motivations (designing elements, communication reliability). In a speech environment, introducing an additional modality -in our case, gesture by means of a tactile screen- allows to overcome some speech recognition errors.

But it raises also new difficulties. The first one is that the informations come from various communication channels: at which level (syntactic, semantic, pragmatic) has the integration to be done? What kind of modeling has to be used? In the literature, few satisfactory responses can be found. We chose to lean on Maybury's works [62], performed in a different context (the generation of communicative acts for the system ouput). Maybury proposes several levels of communicative acts which allow to integrate at each level information coming from different modalities. We adopt this principle (which is fully coherent with our dialogue modeling) but we use it for recognizing the act: the tactile and speech modalities are processed separately as communicative acts which are merged in speech acts.

The second difficulty is the processing of references, particularly in the framework of the chosen application (querying a geographical and tourist database). Indicating the interesting objects during the dialogue is done both by means of speech sentence and gesture (pointing out, drawing a zone) and takes into account the application context (the user can follow the outline of a cartographic object with her finger).

Studies in this domain are in the linguistic field and in the artificial intelligence field. Some linguists [89] propose very precise studies about the condition of use of prepositions (functional approach) in the designation of objects. We think that these results are interesting and we have adapted them for our parsing of sentences. In the artificial intelligence field, several modeling of spatial relations have been proposed. We use the one proposed by IRIT (Toulouse) [93] in order to check the semantic coherency of referential expressions in the framework of our application. This modeling is based on certain characteristics (dimension, morphology, ...) of elements which govern the use of linguistic items in the expressions.

The ambition to put dialogue systems on the market needs to comply with requirement about the quality of interaction. It is necessary to be able to evaluate and compare different systems using different points of view (speech recognition rate, dialogue efficiency, language and dialogue abilities,...) in the framework of equivalent applications, and eventually for the same system, to evaluate different approaches. Various metrics have been yet proposed [83], [47] (for example: length of dialogue, number of speech turns for recovering speech recognition errors), but they do not take into account all the dimensions of an interactive system. Some new solutions are currently under consideration (for example in the CLIPS labs in Grenoble): they are based on pragmatics issues such relevance, or based on the concept of system self evaluation which consists in doing process by the system, or by one part of it, pieces of dialogue which present some difficulties, giving it all necessary contextual information.

3.3.2. Language teaching

The use of ORDICTÉE is concerned with the primary class exercise called dictation. In this application, a speech synthesiser reads French text while the pupil writes the orthographic transcription on his keyboard. The reading speed is continuously tailored to the speed of the typing. The pupil can correct the text whenever he wants. This application is based on the design and the development of specific tools such as the alignment of the text provided by the teacher and the pupil text.

3.4. Machine learning in dialogue systems

Keywords: *Kalman filter*, *grammatical inference*, *hidden Markov model*, *machine learning*, *speech data bases*.

This research theme focuses on the elaboration of machine learning methodologies in all the stages of a dialogue system.

Machine Learning can be seen as the branch of Artificial Intelligence concerned with the development of programs able to increase their performances with their experience[2]. It is basically concerned with the problem of *induction* or *generalization*, which is to extract a concept or a process from examples of its output. From an engineering point of view, a Machine Learning algorithm is often the search for the best element h^* in a family $\mathcal{H}$ of functions, of statistical parameters or of algorithms. Such a choice is done in optimizing a continuous or a discrete function on a set of learning examples. The element h^* must capture the properties of this learning set and generalize its properties.

Machine Learning is a very active field, gathering a variety of different techniques. Grossly speaking, two families of techniques can be distinguished. On the one hand, some Machine Learning algorithms use learning sets of symbolic data and discover a concept h^* which is also symbolic. For example, Grammatical Inference learns finite automata from set of sentences. On the other hand, other Machine Learning algorithms extract numerical concepts from numerical data. Neural networks, Support Vectors Machines, Hidden Markov Models are methods of the second kind. Some methods can work in examples with both numerical and symbolic features, as Decision Trees do. Some concepts that are learned may have both a structure and a set of real values to optimize, as Bayesian Networks or stochastic automata, for example.

The Cordial project is concerned with the introduction of Machine Learning techniques at every stage of a dialogue process. This implies that we want to learn concepts which basically produce time ordered sequences. That is why we are interested in learning from sequences, either in a symbolic background or in a statistical one.

3.4.1. Grammatical inference.

In the frontal part on an oral dialogue system, the incoming speech is processed by a recognition device, generally producing a *lattice* of word hypotheses, i.e. the lexical possibilities between two instants in the sentence. Then a syntax has to be used, to help producing a sequence of words with the best conjoint lexical and syntactic likelihood.

The syntactic analysis can be realized either through a formal model, given *a priori* by the designer of the system, or through a statistical model, the simplest being based on the counting of how grammatical classes follow each other in a learning corpus (*bigram* model).

Both types of models are of interest in Machine Learning : grammatical inference is basically the theory and the algorithmics of extracting formal grammars from samples of sentences; the discovery of a statistical model from a corpus is an important problem in natural language processing. It is interesting to combine both approaches in extracting from the learning corpus a stochastic finite automaton as the language model. It has the advantages of a probabilistic model, but can also exhibit long distances dependencies reflecting a real structure in the sentences.

We have worked on grammatical inference in the recent years, especially within a contract with FTR&D between 1998 and 2001. The field is always very active in the Machine Learning community. Many progresses in grammatical inference have recently be done in the framework of Language and Speech processing [48], [29], [97], [98], [99], [100].

We are now interested in the learning of a special class of finite automata called *transducers*. They read a sentence to produce another one, on a different alphabet. The machine learning of transducers from sets of couples of sentences is a well mastered problem (some real size experiments in language translation have been already made, [91], [70]). We want to experiment and improve these techniques in the framework of the transformation of the outputs of a speech recognizer into a sequence of dialogue acts (see 7.4). In particular, we will consider the introduction of domain knowledge in the learning algorithm.

3.4.2. Nearest Neighbors learning of tree structures

Any sentence is both a sequence of words and a hierarchical organization of this sequence. The second aspect is particularly important to analyze if one wants to understand syntactic and prosodic aspects in oral speech. Producing synthetic speech in oral dialogue requires a good quality prosody generator, since much information is carried through that channel. Usually, the prosody in synthetic speech is made by rules which use syllabic, lexical, syntactic and pragmatic information to compute the pitch and the duration of every syllabe of the synthetic sentence.

An alternative issue is to consider a corpus of natural sentences and to use some machine learning algorithms. More precisely, any sentence in this learning set must be described both in terms of relevant information with regards to its prosody (syllabic, lexical, etc.) and in terms of its prosody. The machine learning task is to produce explicit or hidden rules to associate the description with the prosody. At the end of the learning procedure, a prosody can be associated to any sentence described in the same representation.

The learning methods used in the bibliography make use of neural networks or decision trees, ignoring the hierarchical nature of the organization of the syntax and the prosody, which are also known to have strong links. This is why we have represented a sentence by a tree and made use of a corpus-based learning method. In a first step, we have used the nearest-neighbour rule.

Given a learning sample of couples of trees (sentences) and labels (prosody), $\mathcal{S} = \{(t_i, p_i)\}$ and a tree x, the nearest-neighbour rule finds in $\mathcal{S}$ the tree $t^{\star}$ which is the closest to x and adapts to x a prosody p_x directly deduced from $p^{\star}$.

This raises two problems: firstly to find a good description of a sentence as a tree, secondly to define a distance between trees. We have worked on these questions during the last years [37], [36].

3.4.3. Learning by analogy in sequences and trees structures

In the context of speech synthesis, we would like to use now a more sophisticated lazy learning method: *learning by analogy*. Its principle is as follows: *knowing a sentence x to synthesize, look for a triplet of sentences (b, c , d) in $\mathcal{S}$ such that x is to b as c is to d.*

Actually, we do not yet study learning by analogy directly on trees, but on sequences. The reason is that we use a distance between the trees and the sequences (the edit distance) which is much easier to manage on the universe of sequences.

We have firstly worked on defining what is *solving an analogical equation on sequences* when the edit distance is introduced. In general, an analogical equation can be described as follows: *find x from a triple a, b and c such that a is to b as c is to x* and is often written by

$$a : b :: c : x$$

3.4.3.1. Solving analogical equations

The idea is to generalize the studies of Lepage [59] and of Yvon [96] for whom the edit distance is a trivial case. The classical definition of $a : b :: c : d$ as an analogical equation requires the satisfaction of two axioms, expressed as equivalences of this primitive equation with two others equations [58]:

$$\begin{array}{rl} \textit{Symmetry of the 'as' relation}: & c : d :: a : b \\ \textit{Exchange of the means}: & a : c :: b : d \end{array}$$

As a consequence of these two primitive axioms, five other equations are easy to prove equivalent to $a : b :: c : d$.

Another possible axiom (*determinism*) requires that one of the following trivial equations has a unique solution (the other being a consequence):

$$\begin{array}{c} a : a :: b : x \Rightarrow x = b \\ a : b :: a : x \Rightarrow x = b \end{array}$$

We can give now a definition of a solution to an analogical equation which takes into account the axioms of analogy : x is a *correct solution* to the analogical equation $a : b :: c : x$ if x is a solution to this equation and is also a solution to the two others equations: $c : x :: a : b$ and $a : c :: b : x$.

Solving analogical equations between sequences has only drawn little attention in the past. Most relevant to our discussion are the works of Yves Lepage, presented in full details in [59] and the very recent work of Yvon and Stroppa [81], [96].

Our approach to solving equations on sequences is based on classical edit distance and uses deletion, insertion and substitution. We did not assume that inclusion property is true for analogy. That is where our approach generalizes the studies of Yvon and Lepage.

We consider that the relation "is to" is defined with the alignment between two sequences, and that the relation "as" requires to compare two alignments, which are themselves sequences (or more simply, "as" can be the equality).

3.4.3.2. Aims of this study

We aim at giving a sound definition of analogy in sequences as a first step, then in prosodic tree structures in a second step. With this definition of analogy, we will implement an algorithm for solving analogical equation. Then, in the learning by analogy problem, the adaptation of fast NN-algorithms, such as AESA [64], is necessary. AESA is interesting as it gives a nearest neighbour in constant time on average, with the cost of a pre-computation that is linear in time and space.

3.4.4. Learning to improve the dialogue management

We modelize the dialogue phenomena by using the concepts of speech acts and dialogue acts, and we consider that a sequence of exchanges can be analyzed as the result of planning. Machine learning can also be used to increase the efficiency of the planner. A well-known topic in Artificial Intelligence is the use of experience to increase the efficiency of inference engines, planners, generally speaking every kind of reasoning system. Often used is the framework of Case-Based Reasoning, which uses corpus of previous experience to discover "shortcuts" or memorize often used pieces of elaborated information. Another possibility is to use statistics on the sequencing of actions for making decisions informed by experience.

This work is part of the CRC[1] "Machine learning in man-machine interaction" between the Cordial project and France Télécom Recherche et Développement, DIH/DII. This contract is described in section 7.4.

3.5. Speech processing

Keywords: *Kalman filter*, *grammatical inference*, *hidden Markov model*, *machine learning*, *speech data bases*.

The research activities in speech processing accomplished by the Cordial project lie in a general scientific framework which is the *automatic speech transformation*. This framework, in particular, makes it possible to have some studies in unit selection for speech synthesis, voice transformation, speech segmentation, etc.

What makes a voice specific, the fact that we can recognize a familiar voice at phone for example, is a relatively complex concept to encircle and define. The main first concept considering voice quality is certainly the perceived timbre of speech but it masks other factors like suprasegmental one (melody, duration of phonemes, energy, focus, etc). In this context, a system of voice transformation tries to modify the acoustic characteristics of a *source* speaker voice so that this voice is perceived like that of a *target* speaker.

This research subject coherency can be divided along three different technological axis: text-to-speech synthesis (TTS), biometry, and finally pathological voices.

From a TTS point of view, the source voice corresponds to a standard TTS voice for which a very strong manual expertise was necessary. The target voice corresponds to a *voice footprint* easy to record and prepare. Transforming a reference TTS voice, making it as close the target voice as possible, avoids the discouraging amount of time and cost necessary for the construction of a new reference voice. Under this methodological assumption we can thus consider new applications, unrealistic for the moment, which will consider a voice profile. This profile would enable a user to listen to his emails using the voice timbre of a person who is dear to him. In this case, the constraint of the target voice is relaxed, and instead we try to answer to the following question: does this transformed voice sounds like a human voice? , meaning, are the characteristics of the transformed voice corresponding to a human voice ? even if a corresponding natural voice does not exist.

The repercussions of the proposed studies are immediate in the field of speaker identity. Indeed, two essential elements can explain the interest of a voice transformation study for the development of a speaker identification system. First element is applicative and aims to increase the robustness of an automatic speaker authentication system against impostures. The second corresponds more clearly to a prospective research: it is considered whereas this transformation can be used as a solution to accept his identity considering that a speaker has a voice transformation which allows a synthesis of a high quality vocal signal. The rates of false acceptance and false rejection are the two criteria most frequently used to evaluate the intrinsic performances of an automatic authentication system. However other factors such as acceptability by the user and especially complexity and cost of an imposture are crucial for a real application. The state of the art about controlling imposture techniques must ameliorate the robustness of such systems, increasing the cost of an imposture and thus the attractivity of such technologies.

[1] *Contrat de Recherche Coopérative*, Cooperative Research Contract

It is also conceivable to bring innovative technological answers in the field of handicap for adult as well voices as for children voices. In particular, we think that voice transformation techniques can correct some articulation glitches [20]. Articulation disorders concerns the incapacity of correctly pronouncing one or several sounds. These disorders can be due to delayed development, a lake of muscles control, a cleft lip or a cleft palate, an auditive deficiency or even learning difficulties.

Considering these technological challenges, only some research axes are developed here. We put a stress on the process of acoustic unit selection, an optimal building of linguistic corpus, the automatic annotation and the segmentation of the speech signal, the language models, and, finally, on speaker transformation systems.

3.5.1. Optimal speech unit selection for text-to-speech systems

The TTS issue is interpreted here as a voice transformation task located at a strictly phonological level. To build a *target* voice, we search speech units from a continuous speech database.

For this kind of TTS system, exploiting a continuous speech database, the crux of the problem, is no more the database itself. But the algorithm which selects the best sequence of speech units and finding a best sequence is a combinatorial task. The majority of systems based on that approach avoid the difficulty by using an *a priori* heuristic that permits an acceptable resolution of the sequencing problem. The treatment is generally applied from the beginning to the end of the sequence searching for the longest phonological sub-sequences. This graph search is undertaken while forgetting to specify clearly any assumption on the optimality of the treatment [2]. This assumption leads to dynamic programming algorithms like Viterbi or Disjkstra. The unit selection system then offers acceptable solutions in terms of time complexity. To our knowledge, few works integrate an experimental checking of this optimality principle.

We think that the compromise between a speech inventory with strong linguistic expertise like a diphone database and a continuous speech database is for the moment badly formulated. There are two plausible assumptions to reformulate it:

- Preserving a minimalist algorithm of selection. It is then necessary to reconsider a definition of the speech inventory with more linguistic constraints.
- Having an algorithm of selection with sufficient phonetic and phonological knowledge to find an acceptable sequence; indeed, brute force cannot suffice. Notably, proposing relevant pruning heuristics (taking into account the acoustic criteria while searching for the optimal unit sequence).

3.5.2. Optimal corpus design

The definition of an acoustic unit inventory is a crucial step for the database construction. The final corpus has to satisfy the following properties:

- Covering as well as possible the most of the acoustic transitions in accordance with a language. A prosodic description of units can be combined with a phonemic description (segment units recorded in different prosodic contexts).
- Containing explicit descriptive information at phonologic and linguistic levels. These information permit to characterize the sound elements that will be incoporated in the continuous speech database.
- Guaranteeing a constant vocal quality during the whole inventory. The vocal quality can be deteriorated by a change of the recording procedure[3], or by a modification of some extra-linguistic factors proper to the speaker [4]. Therefore, it is necessary to minimize the global recording duration and so the size of the continuous speech database.

[2] In the sense that any optimal solution carried out on a sub-sequence belongs to the optimal solution of the sequence.

[3] For example, the change of the microphone can introduce enough heterogeneousness, like phase problems, to alter a base.

[4] For example, the speaker catches a cold between two recording sessions.

The linguistic definition of a continuous speech database can be formulated as a set covering problem. Indeed, we have the most complete possible linguistic corpus, composed of millions textual sentences, and we want to condense it by reducing the redundant elements in order to avoid their recording. Each sentence is described by a attribute vector which exhaustively characterizes the considered task. To construct a continuous speech database for the speech synthesis task, each textual sentence is represented by its phonemes, di-phonemes, phonetic and syllabic classes, etc. This set covering problem is NP-complete, and there is no exact algorithms applicable in a reasonable time. That the reason why, for large corpora, methods based on simplifying heuristics are used.

In the general framework of speech area, numerous methods have been proposed in the literature. We find notably the greedy algorithm which consists of the iterative construction of a sentence subset by adding step by step a sentence chosen according to a performance criterion. This performance score aims to reveal the sentence which should contribute at best to the covering construction. Considering the goal to reach, the score can be calculated by different ways, according to sentence units, their frequency [49], their context [88], [39], or a unit distribution to reach in the reduced database [54]. Some variants of the greedy algorithm have been also proposed, like inversed greedy method and pair exchange [7].

These works anchor in the problematic of a continuous speech database for text-to-speech synthesis but present some interest in other linguistic tasks like, for instance, speech recognition or speaker identification [14].

3.5.3. Automatic Speech Labeling and Segmentation

In speech processing, as in many other fields, automatic machine learning methodologies require databases of consequent size [5]. These linguistic samples collected through experiments have the complexity of the various factors that one seeks to model. Thus for speech processing, usually one wishes to establish an explicit relationship between an acoustic level and the phonological level of the language. In this context, a segmentation task consists in labeling the acoustic speech signal by phonological or linguistic events.

By considering an acoustic signal associated with a phonetic transcription, a task of speech segmentation into phones consists in finding the precise time instants of beginning and end of the phonetic segments. This task can be more or less difficult according to the phonological assumptions. In the most favorable case, one has the exact phonetic transcription from the speaker. This case is not too realistic because it requires a human expertise made on the recordings at a phonetic level [6]. An acceptable solution consists in supposing known the textual transcription of the recorded message and to apply an automatic phonetic transcription system [7]. However the automatic segmentation task is more complex because the grapheme/phoneme transcription has no chance to correspond to true elocution of a speaker. One can also think about another solution more acceptable from a practical point of view but more complex to implement if we suppose now that the exact word transcription is unknown.

Concerning speech segmentation under the assumption of a perfectly known phonetic sequence, the most powerful systems consider Markovian models [43]. A sequence of Hidden Markov Models, HMM, is built starting from the phonetic description. Since the main task of a speech segmentation system concerns the precise time location of the phone transitions, monophone models are mainly used. In a training stage, the parameters of each phone model are learned from a corpus of examples using an EM methodology [72]. In a second stage, known as the decoding process, the segmentation system seeks the most probable alignment between the sequence of models and the sequence of the acoustic observations. The temporal stamps delimiting the phonetic segments are easily found considering the transitions between HMM on the optimal path of alignment.

[5] Even if the size of a database remains correlated with the task, the distribution of linguistic or phonological events follows a power law and for this reason one needs very large corpus size.

[6] Skills requested are those of an expert in acoustics and phonetics.

[7] A human expertise is always necessary but the competence required is less accurated because it concerns only a checking of a textual transcription by listening.

Work which we undertake in speech segmentation takes place under the assumption of a relaxed phonological and linguistic sequence. We take for working hypothesis the observation of the speech signal associated with a partially known textual form. Various problems rise from this assumption:

- How to translate automatically a word sequence to a phonemic description ? in particular, integrating all the variants of pronunciation. The theoretical modeling support is the graph of the phonemic sequences.
- Starting from the graph of the phonemic sequences, how to find, by using an adequate acoustic modeling - typically an HMM, a mapping between the speech signal and the phonetic labels ?
- Which confidence measures make it possible to locate dissimilarities between the real pronunciation made by the speaker and and the phonetic hypothesis found in the graph of transcription ?
- These confidence measures are then used to control a speech segmentation process by manual expertise. The expert will concentrate its work only on the incorrectly segmented speech sounds.
- Finally, to propose solutions to soften the constraint of a perfectly known text. Here we think about using traditional speech recognition techniques only on small portions of the word sequence indicated by confidence measures.

3.5.4. Language modeling

Language models are involved in numerous information processing systems given that the data processed are word sequences. Domains such as speech processing, automatic translation, data mining or natural language interfaces may be concerned by this methodology. Using more or less restrictive assumptions, a language model describes the sequence of words in a sentence.

From a methodological point of view, there are two ways to test the effectiveness of a language model: either by evaluating the model itself with no regard for any applicative context or assuming that it is restricted to a task (as for example a speech recognition system). This latter case means that the performance of the task qualifies the effectiveness of the language model.

The cross entropy between a language model and a test set is the most widely used criterion to evaluate a language model. The less the entropy is, the better is the model. However, the perplexity criterion, based on the cross entropy, suffers a few drawbacks. The probabilistic assumptions is one of them. It's the main reason why experimental procedures require protocols promoting the evaluation of the entire task. But this methodology is as contestable as the former. It is difficult indeed to distinguish, on an entire performance, which is the part of the language model from the rest of the system. Besides, transposing the results acquired from different domains is quite uneasy.

In order to alleviate this drawback, we propose to adopt an evaluation framework for language models, independent on the task, and to abandon the perplexity criterion. We privilege a framework based upon a statistic on the ranks of prediction, which does not require any assumptions on the models. A predictive framework is not a new way for evaluating language models; Shannon worked on a prediction framework in the beginning of 1950 [78].

3.5.5. Voice transformation

A Voice transformation system modifies a *source* speaker's speech utterance to sound as if a *target* speaker had spoken it. This technology offers a number of useful applications in computer interfaces and biometrics. For instance, human/computer voiced interaction would be enhanced if a large variety of high quality synthesized voices were used. On the biometrics level, a transformed voice could pose as a real voice in order to test a voice-based authentification system.

The specificity of a voice resides in two acoustic notions. The first one, the timbre, qualifies the speech signal on the segmental level. The second one aggregates supra-segmental characteristics such as melody, speech rate, phone length and energy. Most of voice transformation researches are conducted at a segmental level and supra-segmental variations are not addressed. We think that the quality of a transformed voice could benefit from joint processing of the signal on the segmental and supra-segmental level.

At a segmental level, a voice transformation technique implements a transformation on an acoustical representation of the *source* speaker's speech signal. The modified speech signal should be *perceptually* close to the one that would have been uttered by the *target* speaker. As a consequence, two sub-problems should be addressed to perform this voice transformation. First, the speech signal parametrization should take into account the voice characteristics to be transformed. Second, the transformation should be found (computed).

The first voice transformation techniques were based on methods used in speech recognition [30]. A speech recognition system gives improved performances if it is designed to recognize a specific voice. That is why speaker adaptation methods seek to tune the recognition system to a single speaker (the user), and speaker normalization methods transform voices from a multitude of speakers into a single voice.

The segmental acoustic space of a speaker is hard to model and is strongly linked to the phonetic characteristics of the language. A segmental voice transformation technique, should rather take into account the perceptive differences [8] between two utterances of a same sound by two speakers than the phonetic characteristic of the sound itself. As a result, the acoustic space of the speakers should be quantified in order to minimize the representation of socio-linguistic characteristics and thus to reveal the speakers acoustical characteristic. Consequently, most of speech transformation methods use the same course of action. First the acoustic space of a speaker is segmented. Then, a specific transformation is separately implemented on each segment of this acoustic space. We propose to follow this general approach of voice transformation. Our goal is to include original algorithms for automatic classification and automatic transformation learning. These transformation could be linear or non-linear.

For what concerns the voice transformation in a supra-segmental point of view, few publications focus on this subject to our knowledge. We want to apply a similar approach to the previous one at the segmental level. In order to derive a transformation between the source and target prosodic spaces, we need to define and determine the prosodic space proper to a locutor. Although the intonation is a combination of numerous factors, we first focus on the automatic modeling of fundamental frequency F_0. F_0 contours, extracted from the speech signal, ensue from the evolution of the vibration of vocal folds over time and are mono-dimensional signals with real values. In addition to the modeling problematic, the model has to be appropriate to a classification providing a melodic space characterization of the locutor by a class set. Numerous models of such contours are introduced in the literature, like symbolic models based on tags [63], quantitative generative models using phrase and accent commands [50], dynamical state space model which consider that the observation of a portion of the melody is explained by a stochastic state variable [73]. Especially, a wide range of publications deals with stylization of F_0 contours using polynomial functions. We can cite models like MoMel [51], [66], Tilt [86], as well as Sakai and Glass's model [75] based on regular spline functions.

In the scientific framework that we have previously covered, our undertaken works in voice transformation answer the followings problematics:

- To determine a segmental transformation of high quality by the proposal of relevant models of speech signal analysis and synthesis (models of type FD-PSOLA [67], harmonic and noise model [56] which include the transient analysis [90]). It is necessary to take account a model of glottal wave signal [94] ?
- To propose unsupervised classification solutions of segmental spaces.
- To derive by unsupervised learning the prosodic space of a speaker (joint treatment of the melody and the duration).
- To transform jointly functions of segmental and prosodic spaces.

All these theoretical works are validated in two experimental domains: the one of speech synthesis for what concerns the problematic to diversify the synthetic voices and the one of falsification of voice authentification systems.

[8] These differences are due to variation in the speakers' physiology and their sociological background.

4. Application Domains

4.1. Application Domains

The application domains for our researches are all the situations where man-machine communication requires speech or where the use of speech brings more comfort. These applications are in general complex enough to require a real dialogue situation, and would be tedious if used through a simple sequence of guided short answers.

Examples for these applications are : information services on a personal computer or on a public, booking services by telephone, computer assisted language learning.

5. Software

5.1. Introduction

We develop our applications on the CNRT platform DORIS, to promote joints projects with industrial research. The GEORAL system is a demonstrator of touristic information services, with oral dialogue and tactile screen. We also have a "dictation" software called ORDICTÉE, which has been experimented in primary schools.

5.2. DORIS platform

Participants: Laurent Miclet [*correspondant*], Jacques Siroux, Olivier Boëffard.

The Cordial project aims to promote its research activities by means of technological demonstrations. To achieve this point, hardware and software resources have been defined to build a R&D platform named DORIS and dedicated to man-machine interaction, in particular with the use vocal and dialogue technologies. The main funding comes from IRISA/INRIA, the Regional Council of Brittany and Cordial public contract funding. DORIS, in the context of the CNRT-TIM Bretagne, has vocation to promote joint projects between institutional and industrial research.

DORIS is concerned by the different research projets like GEORAL (see sections 5.3 and 6.2) and ORDICTÉE (see section 5.4). In November 2005, an research engineer (founded by FNADT) will be full time on the DORIS platform to manage the technical aspects and to develop new softwares for the previously quoted projets.

5.2.1. Hardware architecture

On the powerhouse systems side, a Compaq AlphaServer system has been chosen to support the calculation power needs, especially for speech processing. In addition, the platform includes a Network Appliance file server with a storage capacity up to 350 Go.

In order to facilitate technical access for industrial partnership, the platform includes fast secure network access. DORIS inherits from the ENSSAT-Université de Rennes 1 network. We propose high internet connection with VPN access.

On the client side, PCs with an up-to-date sound configuration are used. These computers are meant for software development within DORIS. They are nowadays used by engineers, PhD and postgraduate students involved in the CORDIAL project. Touch screens have been purchased in order to facilitate the development of multimodal man-machine interfaces.

This client-server configuration is fully functional inside the ENSSAT campus. Further improvements will be focused on lightweight clients and resources sharing with external partners (see section 5.2.3).

5.2.2. Software architecture

The DORIS platform main goal is to group research projects that deal with the man-machine interaction field. In this entity, they shall take advantage of other teams works and tools.

We first direct our efforts towards the installation of a multi-agent[9] architecture. It satisfies the needs for modularity, quick and clean development and interoperability. To fulfill this role, we chose and installed JADE[10], a software framework fully implemented in Java language. It allows the implementation of multi-agent systems through a middle-ware that complies with the FIPA[11] specifications. The agent platform can be distributed across machines, which do not need to share the same OS.

We made this choice to simplify the development while ensuring standard compliance. Furthermore, the Java technology allows us to use already developed libraries that are not necessarily in our sphere of competences (e.g. sound or speech coding, framing, streaming) and therefore to concentrate on the scientific interests of the team.

Today, one main project has taken place inside the DORIS platform: GEORAL (see sections 5.3 and 6.2).

5.2.3. New steps with DORIS

Thanks to the CNRT TIM Bretagne, Télisma and France Télécom R&D are actively involved within the DORIS project. Télisma proposes their software suite for speech recognition and France Télécom R&D for text-to-speech synthesis.

Several publications have reported on efforts in building such a platform and several issues need to be addressed. Among those, we focus in this work on the distributivity of the solution based on an Agent architecture, and on the use of Voice over IP solutions, and we illustrate such issues through a demonstration application built upon such an architecture. Additionally, this platform helps us to integrate different third-party solutions – speech bundles, VoIP protocols, applications, etc. – and test them in an acceptable technological environment.

A salient feature of the proposed solution is to mask the third-party API specificities behind the MRCP protocol, Media Resource Control Protocol. MRCP controls media service resources like speech synthesizers, recognizers, signal generators, signal detectors, fax servers etc. over a network. This protocol is designed to work with streaming protocols like RTSP (Real Time Streaming Protocol) which help establish control connections to external media streaming devices, and media delivery mechanisms like RTP (Real Time Protocol). RTSP protocol is a standard protocol for controlling the delivery of data with real-time properties. The main contribution of this protocol to our platform concerns the negotiation of the RTP, setup parameters (client and server port numbers, session id) and the transport of MRCP messages between client and proxy-Agents dedicated to speech ressources. We have defined half-duplex streaming. A client can initiate a session on the DORIS platform from one source, for example a PDA, and get a speech feedback from another source, for example with a cellular phone. We developed a complete stack following the MRCP specifications and other necessary protocols like RTSP and RTP. An API for MRCP clients has been developed in Java (the MRCP stack and the client API is about 12000 lines of code).

Several communications have been done during the year 2004, including presentations and demonstration with industrials, local organisations and journalists. The INRIA associate engineer managing the platform has taken part in a congress (Synerg'Etic, Nantes).

The multiagent/IP possibilities of the DORIS platform has been presented in Interspeech 2004 congress in Corea, [55].

5.3. Georal

Participants: Jacques Siroux [*correspondant*], Marc Guyomard, Sébastien Carbini, Ali Choumane.

[9] An agent is an independent and autonomous process that has an identity, possibly persistent, and that requires communication with other agents in order to fulfill its tasks.

[10] Java Agent Development Framework, a free software distributed by Telecom Italia Lab (TILAB).

[11] Foundation for Intelligent Physical Agents, which purpose is the promotion of emerging agent-based applications, services and equipment. This goal is pursued by making available internationally agreed specifications that maximize interoperability across agent-based applications, services and equipment.

GEORAL TACTILE is a multimodal system which is able to provide information of a touristic nature to naive users. Users can ask for information about the location of places of interest (city, beach, chateau, church,...) within a region or a subregion, or distance and itinerary between two localities. Users interact with the system using three different modalities: in a visual mode by looking to the map displayed on the screen, in an oral and natural language mode (thanks to a speech recognition system) and in a gesture mode pointing to or drawing on a touch screen. The system itself uses both the oral channel (text-to-speech synthesis) and graphics such as the flashing of sites, routes and zooming in on subsections of the map, so as to best inform the user.

The GEORAL project started in 1989 and is the origin of various works since then. It was fully developed in Visual Prolog 4.0. We decided to re-implement GEORAL making the most of the capabilities of the DORIS platform.

The foundation stone of this re-implementation was to split the initial Prolog modules (syntactic and semantic analysis, dialogue management and tactile screen management) taking into account the multi-agent paradigm (one module for one functionality). We assigned one agent for each specific role, agents that are written in Java. However, we kept core functions written in Prolog, in order to take advantage of the fact that this language is really convenient for tasks like natural language processing. But all peripheral functions from screen management to client-server communication have been rewritten in Java and C/C++ languages.

The call of Prolog predicates from agents written in Java was not straightforward. After a benchmarking phase, we decided to use a Java package that allows such calls (*tuProlog*). This implied studying the existing Prolog files to extract useful predicates and to correct them to bring the code closer to the ISO Prolog. Furthermore, the work on the Prolog code allowed an improvement of the GEORAL engine capacities. A larger range of queries are now accepted by the system and some bugs have been fixed. Improvement have also been made on the gesture management side. The touch screens prove to be useful to process new kind of drawings like windings follow-up.

A text-to-speech server has been installed and a dedicated agent communicate with it. The processing time and the sound quality are very good, but we are using a local network for the moment. We have in mind to insert the Internet between the clients (possibly wireless devices) and the server. An important work on data coding and communication protocol has to be made beforehand.

Several analyses and use tests have been made on the different normalized communication protocols between agents (FIPA normalization). The dialog engine has been modified to make the dialog be the most natural (taking ellipses, anaphores and interruptions into account). Student projects have been integrated to model new types of tactile acts. Agents for speech recognition and speech synthesis have been developed for communication with the server. The grammar of speech recognition of GEORAL has been written.

Finally, several improvements have been added, including the improvement of screen display, the speed up and debugging of the code, and internal facilities for software development (abstract agents, agents managing functions, simplified communication interface).

To learn more about GEORAL system, see section 6.2.

5.4. Ordictée

Participants: Marc Guyomard [*correspondant*], Olivier Boëffard.

As explained in section 3.3.2, ORDICTÉE is a software that allows a pupil to perform a dictation exercise on his one. It is made up of three modules: The pupil module, which, together with the pupil itself, carries out the dictation exercise, the teacher module, which allows the teacher to design his own dictation texts, and the administrator module which is devoted to set the application parameters. One of the main function of ORDICTÉE is to follow the typing, i.e. to adapt the reading rhythm to the typing speed. This function is based on the one hand on the hypotheses that mistakes do not affect the pronunciation, and on the other hand on the phonetic closeness of the two texts (the pupil text and the teacher one).

5.5. Semantic Parrot

>From a usability point of view, the semantic *parrot* that we propose consists in taking a speech message from a standard audio input (Personal computer, PDA, Cell phone), to understand the underlying concepts and finally to generate a paraphrase using a speech ouput.

>From a technological point of view, the semantic parrot implements techniques of speech recognition, automatic speech understanding, and finally of concept to speech synthesis.

Currently, a first demonstrator is built on the DORIS platform (see section 5.2) implementing a technology of speech recognition provided by TELISMA and a technology of speech synthesis provided by FTR&D.

5.6. Collaborative software

A set of collaborative softwares is available through the DORIS platform as to make the Cordial project members' work easier.

These tools have also been made available to the different research partners of the Cordial projet, in particular to the members of the VIVOS project (see section 7.2).

These softwares are:

- A wiki engine, used to build a type of web site that allows the visitors themselves to easily add, remove, and otherwise edit and change some available content. This ease of interaction and operation makes a wiki an effective tool for collaborative authoring. Inside the Doris platform, this tool is used to present the research activities of the different projects, their achievements, to share scientific analyzes, to write working memos or reports, to describe software formats used for some task, etc. All the reference documents of the projet are available through this web site;
- Subversion, a version control file system. With such a repository system, registered users can share files, read or write them simultaneously under a powerful conflict management policy, while a complete modification history is logged, allowing to retrieve any previously store file revision. The Cordial project members use this tool to write and manage its scientific publications, its software developments and its official reports. The Vivos partner members use it to share some of their speech data, their common specifications, to write their ongoing research reports, etc.;
- a bibliography server, to unify and share all the publication references used by the project members;
- E-GroupWare, a collaborative projet management software, dedicated to the Vivos project. It is used to share contacts, to setup meetings, to plan and follow tasks, etc.;
- a bulletin board system, also dedicated to the Vivos project. It allows the members informal discussions about any topics of the ongoing project.

Available though simple web interfaces, the two last tools are especially useful for the Vivos project and its geographically distant members.

6. New Results

6.1. Dialogue and modeling

6.1.1. Logical modeling for dialogue processing

Participant: Jean-Christophe Pettier.

This research topic has no new results in 2006.

6.1.2. Dialogue systems evaluation

Participant: Jacques Siroux.

There is no new results in this theme as no manpower could have been devoted to this task since the end of the contract of our engineer.

6.1.3. Modelling the Communicative Common Ground (CCG) in term of Collective Acceptance

Participants: Sylvie Saget, Marc Guyomard.

This study is covered within the framework of a PhD thesis funded by the grant A3CB22 / 2004 96 70 of the regional council of Brittany. Work began on October 15, 2003.

The problem, underlying this study, is to enhance the interactivity of spoken dialog systems through the modelling of negotiation sub-dialogs at the dialog level (meta-dialog). Modelling reference negotiation sub-dialogues is a way of handling "communicative errors", by giving a dialog system and its users, the capacity to interactively refine their understanding until a point of intelligibility is reached. The approach chosen within the framework of this thesis is based on the explicit modelling of the collaborative aspects of dialogue in order to obtain an explicative as well as generic model. Besides, such a modelling is also interesting in regards of unsolved questions concerning the design of team members ([52]).

This study focus on defining the proper mental attitude in order to formalize the Conversational/Communicative Common Ground (CCG), that is the set of coordinations involved in dialog. Actually, dialog partners have to face several problems while trying to understand each other: high ambiguity of language, the vocabulary problem, indirect accessibility of other agents' mental state, the distance of their mental states (due to different perceptions, different judgements, different levels of knowledge,...). Recent works ([41], [71]) in psychology show that human dialog partners address these problems notably by aligning with linguistic objects they choose in order to achieve a particular communicative intention: lexical choice, conceptual pact, grammatical structure and so on, and by aligning with coordinate on the way of producing utterance: clarity of articulation, accent, speech rate and so on.

The first part of the thesis is the improperness of considering utterance generation and interpretation as being done in regards of dialog partners' subjective mutual beliefs *in the general case* ([27], [26]).

Generally speaking, a spoken dialog system is commonly considered as being rational. The system's rationality is notably transcribed by its sincerity (following Grice's Maxim of Quality) and by the coherence of its mental state, [57]. Moreover, utterance treatment (generation and interpretation) is notably based on the (Subjective) Common Ground, among dialog partners, [80]. Accommodation, [60], is then a way to ensure the coherence of their epistemic state while solving coordination problems. Respecting this fundamental hypothesis constraints spoken dialog systems to support rich epistemic states (containing mutual beliefs and nested beliefs) and the associated reasoning process. On the whole, concrete spoken dialog systems are not able to deal with these constraints ([85]).

Is the sincerity hypothesis necessary in the general case? We claim that it is not the case ([27], [26]): utterance (in particular its semantics) has to be consider as a *tool* in order to achieve communicative goals. In the particular case of reference, basing on A. Kronfeld's work ([53]), the goals underlying referential acts should be defined as:

1. Literal goal: that as a result of the hearer's recognition of the noun phrase as a referring expression, the hearer will generate a local individuating set that will determine this very same object;
2. Discourse purpose: that the hearer will apply various operations to the newly created individuating set so that it will meet the appropriate identification constraints.

Besides, according to H.H. Clark et al. in [46], the literal goal has to be refined in considering this two interrelated goals:

1. Identification: Speaker is trying to get hearer to identify a particular referent under a particular description.
2. Grounding: Speaker and hearer are trying to establish that the addressee has identified the referent as well enough for current purpose.

The choice of a particular referring expression is then based at the first place on its *usefulness* to reach these communicative goals. Thus the mutual evidence on the true value of the referential expression may be useful but it is not necessary. In order to formalize such a choice and the result of mutual understanding, the intentional mental attitude of *acceptance* is used and its multi-agent counterpart *collective acceptance*. (Collective) Acceptances have the following properties, in contrast with beliefs [95]:

- They are voluntary (or intentional);
- They hold on utility or success (thus we can accept something we believe false);
- They do not required justifications;
- All or nothing: we decide to accept or not to accept.

The philosophical fundament of this approach has been presented in [27]. Moreover, a dialog model as well as a model of referential act based on acceptance has been proposed in [26] and a first step in linking such models to reference treatment has been explored and described in a paper accepted at IJCAI 2007.

6.1.4. Hability modeling

Participants: Karl DeVooght, Marc Guyomard.

In the Russel and Norvig's book [74], an agent is defined by three main features:

- An agent perceives his environment, adapting himself and acting in consequence.
- An agent persists in the time, and can consequently perceive its own dynamic and those of the environment.
- An agent evolves in an autonomous way: he can learn in order to refine his initial partial and incomplete beliefs.

In our work, we relate at least the two first features to the study of agent cognitive skills. A cognitive skill is an agent capability to realize a cognitive process i.e. a process based on the knowledge of an agent. Subsequently, our approach is based on formal models on which we aim at describing the agent mental state with mental attitudes (e.g., close to a BDI-like approach).

For the moment, such existing models do not fulfil completely expectations mentioned above in the case of agent cognitive skills consideration. First, few models get an agent self-aware on what he can do whenever it is in question long-term processes. This lack does not enable an agent to perceive fully its own dynamic. Secondly, the description of cognitive skills rest generally upon a notion of action as a change of state. This is pretty weak whenever one wants to model cognitive skills requiring different agent's behaviour (e,g, helping some user vs negotiating some contract). Finally, we argue a cognitive agent cannot accurately adapt himself by evolving his mind in the time since he is endowed with a monolithic reasoning capability which does not favour an agent to behave suitably in a set of specific and various situations.

For dealing with these problematic points, we proposed a cognitive agent model characterizing by three main concepts: capability, activity and context. Such a model was conceived for matching with some theoretical intuitions as well as for being a pattern in order to develop more flexible and complex cognitive agents. We plan to illustrate concretly our approach on an agent framework, called JADE (Java Agent Developpment Framework).

6.2. System and multimodality

A study about referring phenomena in an enlarged version of GEORAL had been led. We also continued activities to improve the ORDICTEE software (dealing with faults coming from phonetic, following typing).

6.2.1. Georal Tactile and reference

Participants: Jacques Siroux, Ali Choumane.

Recent progresses in speech recognition allow to plan new important developments inside the dialogue system GEORAL TACTILE [79]. Increasing the vocabulary size gives the users the possibility to utter more complex linguistic sentences. We use this fact to enrich the application world with new elements on the map which is the support for querying. In this new framework, several issues are studied: modeling the cartographic context, linguistic and gestural of users referencing elements on the map, and at last the architecture of the system.

In a first time we have made an experiment in order to determine the linguistic behaviour of the users when they reference elements on the map. A large number of linguistic forms and of tactile built up elements (for example referencing a triangle using particular points) have been observed. A new type of gesture (following a line) has also been observed [42].

We have proposed a syntactic model in order to parse and filter referential expressions in the user utterances. This model is based on Vandeloise and Borillo's works [89], [38] which take into consideration the spatial characteristics of the handled elements. Next we have developed a semantic model which allows to filter more precisely the output of the syntactic parser. The model is derived from the Aurnague's one [32] which uses specific attributes of the elements (for example size, consistency, position, ...). We only use three attributes (dimension, consistency and form) but we combine them in order to take into account the possible syntactic forms.

As far as the cartography is concerned, we developed a new data model and search algorithms that are better adapted to handled elements.

Finally, we have redesigned the architecture of the system and the processing flow in order to deal with various facts: more complex gestures, references on objects which are not stored in the database and a two stages processing. By contrast with the current version, we have given priority to gesture activity over speech activity; this principle allows to progressively check and possibly correct the referential linguistic expressions, to determine referents on the map and to build up, if necessary, new elements in the database. Some of these algorithms have been implemented and we are integrating them in the system.

We began studies, firstly in order to model in uniform way the different semantic points of view (natural language, graphics) from the Pineda and Garza's work [Pin00], secondly to bring together the processing on references in GEORAL and the plan-based modeling of dialogue. We began to studying the use of the concept of salience taking into account the results from LORIA project-team Langue et Dialogue. We especially studied the processing of some tactile designations: those that appear when user touches the screen following the cartographic representation of roads, rivers, ... Some referring ambiguities may arise if two cartographic elements are very close or if the user's performance is fuzzy. We propose to solve these ambiguities using a salience score to choose the best candidate. Some preliminary results are encouraging but we have to experiment the algorithm with naive users in real conditions and with more complex geographic maps and elements.

We have started another study in order to design the best way for representing linguistic knowledge (from lexical level to contextual level). The best way means that the design and implementation would be on the one hand, less expensive as possible, and on the other hand, reusable and easily integrable within the system.

We complemented the above studies on referential problems by studying two complementary ways. The first one is concerned with works on written natural language understanding for applications as data mining, question answering, message understanding, etc. Some of these works [92], [65] are interesting for our purpose because they are using poor knowledge and light parsing in order to solve anaphora. But, they need using corpora in order to tune up the values of the different parameters used. The second one is concerned with text generation studies [Man03]. In this thesis work, the author shows that it is necessary to use linguistic knowledge in order to generate referential relationships and that this knowledge could be deduced from experiments and corpora. It could be interesting to merge this knowledge with the Vandeloise's results.

During the first year of the REPAIMTA project (partially funded by the regional council of Britany), we produced a state of the art on automatic processing of referential expressions (pronominal anaphora, definite description anaphora). We defined a first version of a model for dealing with referential expressions within the Georal context. The model includes four representation languages. The first one is concerned with the

natural language modality; it allows to parse the oral input, to determine what kind of referential expression is present in the utterance (taking into account lexical and semantic information and some results from Vieira and Poesio) and if possible to solve the referential expression. The deictic expressions as well as expressions which are referring to entities on the displayed map can't be solved during this step. The second language represents information displayed on the map. We pay special attention to the visual saliance of the entities on the map. This visual saliance can be one of the parameters needed to choose the better referent during the solving of referential expressions. The third one is concerned with the tactile activities: kind of drawing, coordinates, ... The goals of the last language are to merge the information coming from the modalities, to solve the pending referential expressions and to check the coherency between modalities. Some of the results coming from Vandeloise and Aurnague works are used to lead the needed inferences. This proposal has been presented in the InScit2006 conference [21]. The first steps for the implementation are under process.
We have intention to start a new study which will be concerned with a new mode of interaction with Georal prototype. S. Carbini studied during his PhD thesis algorithms and programs to detect (by mean of camera) hand moves used to interact with a system. We want to study the use of these new mode and modality for Georal.

6.2.2. Ordictée

Participants: Marc Guyomard, Olivier Boëffard.

A new algorithm for the identification of the pupil spelling mistakes is implemented. It is expected to overcome some of the major drawbacks of the usual alignment algorithms. As far as the following of the typing is concerned, new features are under investigation. They aim at a better synchronisation between the pupil text and the teacher module utterances.

6.3. Machine learning in dialogue systems

6.3.1. Grammatical inference

Participants: Erwan Livolant, Laurent Miclet.

A thesis has started in 2004 with the following topics : the adaptation of the actions of an agent in a communication situation. The main issue is to give the agent a capacity of analysis on the ongoing dialogue, in order to adapt dynamically its strategy if necessary.

This research topic has no new results in 2006 as the thesis has been stopped.

6.3.2. Nearest Neighbors learning of tree structures

Participants: Sabri Bayoudh, Arnaud Delhay, Laurent Miclet.

This research topic has no new results in 2006.

6.3.3. Learning by analogy in sequences and trees structures

Participants: Sabri Bayoudh, Dorian Le Direac'h [Master Student], Arnaud Delhay, Laurent Miclet.

After having given in the last years a definition of the analogical dissimilarity between objects (including sequences) and an algorithm for computing it, we have focused this year our attention on learning a classification rule based on analogical proportion and analogical dissimilarity (recall that four objects have a null analogical dissimilarity iff they are in analogical proportion). This has been experimented on different data, especially those described by binary and nominal attributes.

We have devised an original non-parametric classifier, which allowed us to obtain promising classification rates. This technique uses the notion of analogical dissimilarity between four objects at the learning step and the resolution of analogical equations on the class labels at the decision step. Basically, it examines all the triples in the learning set and computes for each one its analogical dissimilarity with the object to classify. Then it retains the best triples (those with low analogical dissimilarity and a non ambiguous decision on the classes) and makes a vote on their class labels.

We have strongly improved the performance of this basic method by weighting the attributes. This is an idea used by many classifiers, but we develop it by weighting an attribute according to the departure and the arrival class in the analogical equations involved, which makes this method quite specific to classifying by analogy. The departure class is the class of the two first objects in the analogical equation "A is to B as C is to X" and the arrival class is that of C (and of X when solving this equation). Consequently, each attribute has C^2 different weights, where C is the number of classes, according to the type of analogical equation in which it is involved. The weights are learned from the learning sample.

Using this weighting matrix has dramatically increased the quality of the classification rate. We have obtained the best classification rate on some classical data set when comparing with standard classifiers, such as the K-nn technique, the multi layer perceptron, a decision table, PART, ...The weighting technique and the results are described in an paper presented in CAp06 [19] and in a paper accepted at IJCAI 2007.

Still the weak point of our method, what could be called an "analogical proportion classifier" or "analogical dissimilarity classifier" is the computation time, since it uses all possible triples from the learning set ; the basic complexity is $O(m^3)$ (m is the number of instances in the learning set). In order to fasten the algorithm, some methods derived from the LAESA algorithm have been already developed in the previous year [19].

In September and October 2006, Sabri Bayoud has been invited by Jose Oncina (author of the LAESA algorithm) from the laboratory of "Lenguajes y Sistemas Informaticos" in the university of Alicante in Spain. During the stay, he studied in detail different fast research algorithms for the K nearest neighbors classifier and how to adapt them to our problem. Then, he implemented the kd-tree algorithm for the "analogical proportion classifier" and tested it on several data sets. The results were that, on the one hand it was faster than the LAESA algorithm on binary and nominal data, on the other hand it computes more analogical dissimilarities than the LAESA algorithm.

6.3.3.1. Solving analogical equations

The "analogical proportion classifier" has also been applied to sequences. In this case, computing the analogical dissimilarity is costly, because of the alignment of the four sequences using a four dimensional dynamic programming algorithm. On this type of data the LAESA algorithm is at the contrary faster than the kd-tree method. Comparisons between the Fukunaga algorithm and the kd-tree method have also been made on binary and nominal data as well as on numerical data.

We also have worked this year on the recognition of written letters. The data have been proposed to us by the IMADOC project. We have encoded in a naive manner these caracters as sequences of 8 elementary directions. Interesting results have been obtained, showing that analogy is able to capture both the nature of the character and the style of the scriptor. More experiments are to be devised, in collaboration with IMADOC, especially the generation of letters "in the style" of a given scriptor. This work was the subject of the research period in the Master 2R of D. Le Direach.

6.3.4. Learning to improve the dialogue management

As indicated in section 7.4, some collaborative work has started in 2004 in the framework of the CRC with FTR&D. No manpower is explicitly devoted to this task, since no PhD student within the CRC has been oriented towards this activity.

6.4. Speech Processing

6.4.1. Optimal speech unit selection for text-to-speech systems

Participants: Hung Cao [Master Student], Olivier Boëffard.

Searching for a sequence of units is generally based on a graph of candidate sequences associated with a metric qualifying the global cost of a sequence. For the continuous speech corpora, the support of representation is usually the phone. The problem is hard to solve and corresponds to a blind sequencing task: at the same time, one needs to find unknown relevant units and to form optimal sequences with these units. Once mapping functions between candidate and target units are given as well as a metric about the overall acoustic quality, then the system tries to find N-best paths in a valuated graph. It is in practice impossible, for constraints of space and temporal complexity, to enumerate exhaustively all the candidate sequences. To limit the search space for the best acoustic sequence, current TTS systems impose an search heuristic usually based on a dynamic principle (additive cost functions, etc).

During 2006, with the help of Hung CAO, Master of Research in Informatics, we propose to explore a new strategy allowing to find the N-best acoustic unit sequences. These N-best sequences will be then analyzed from a perceptive point of view so as to validate or not our assumptions. We propose to explore heuristics with less constraints than the ones proposed by the community taking into account a descriptive model of the acoustic coarticulation. We propose to connect the acoustic quality of a sequence of units candidates to a HMM probabilistic model.

Given a continuous speech corpus annotated with phones, the task consists in studying a N-best search algorithm in a DAG using a A^* algorithm and a pruning heuristic formulated with the help of a HMM model sequence.

6.4.2. Optimal corpus design

Participants: Nelly Barbot, Jonathan Chevelu [Master Student], Arnaud Delhay, Olivier Boëffard.

The study of a continuous speech database is a part of H. François PhD works, attended in December 2002. The main difficulty consists in determining in a automatic way a minimal set of sentences taken from a huge set of written sentences. This selected subset is then recorded and is used as examples for learning methods of acoustic prediction models. This selection has to be optimal according to a criterion of precision of acoustic models. The problem can be formalized like a minimal set covering one. The conducted works led to two significant results:

- The construction of a textual database from various sources (dialogue transcriptions, literature, TV series scripts, medicine courses). This database contains approximately 310,000 sentences, and which its elements are annotated using phonologic information.
- Obviously, it is impossible to record the speech equivalent of this first database. Using a greedy algorithm, a subset of 4,000 sentences has been extracted, ensuring 95% of the allophone covering.

A score function links each weighted sentence with several phonological criteria. This scoring takes into account the number of sentence units and the unit frequencies in the initial corpus. Optimizing a unit coverage is a NP-complete problem and one needs to handle matrices of size around $310,000 \times 30,000$. Some variants of the optimization algorithm have been implemented: greedy algorithm, spitting method and the pair exchange method [5].

During 2006, we continue the work initiated by H. François PhD in the framework of the Master training period of Jonathan Chevelu. We keep all the different score functions introduced in [7], and we study the implementation of an integer programming approach. Indeed, the set covering problem (SCP) is widely studied in the literature and the list of applications of Lagrangian relaxation includes some difficult combinatorial optimization problems of huge size. Our corpus condensation problem being a large scale SCP, we consider Lagrangian-based heuristic approaches in order to prune the search space. Our study is based on works for solving crew scheduling in an italian railway company [44], [45] and aims to compare the performances of these approaches with the previous results obtained by a greedy methodology.

6.4.3. Automatic speech labeling and segmentation

Participants: Laure Charonnat, Gaëlle Vidal, Vincent Barreaud, Josselin Huaulmé, Olivier Boëffard.

This study is covered within the framework of a ANR/RIAM project funded by OSEO/Anvar 7.2. Work began on April 1, 2006.

An expertise on speech recordings provided by a dubbing company has been carried out by Gaelle Vidal who has developed recommendations regarding data preparation for automatic segmentation (file format, accommodation and use of the software Transcriber).

6.4.4. Language modeling

Participants: Pierre Alain, Nelly Barbot, Vincent Barreaud, Olivier Boëffard.

This study is covered within the framework of a PhD thesis funded by INRIA within a scientific collaboration with FTR&D Lannion (FTR&D/DIH/D2I) 7.4. Work began on October 1, 2003.

During 2006, we focused on language model evaluation tasks. Considering the main language models evaluation methodologies, we can distinguish those integrating a model in a complete system (Recognition Score, Precision and Recall Score, BLUE Score, etc), from those which are task independent.

In [16] and [17], we proposed to evaluate language models by their *language prediction capacity*. This prediction capacity is characterized by, given a starting sequence and requiring the model to predict the following words, the number of proposals necessary for the model to guess the correct words. The more the number of proposals is important, the worse the model is as a language predictor. This evaluation methodology is an extention of the Shannon's game [77]. Instead of taking a prediction from a known history of 0 (the letters are distributed according to their appearance frequency in the language, the most frequent letter *E* is predicted in first, ...), to 100 (the subject knows 100 letters in a text, and must predict the following one), we can now ask the evaluated models to predict several joint words in a single row, and in open vocabularies (30 000 English words). The experiments are performed in English, we obtain ranks distributions for 12 000 two words frames for different language models (a 3-gram, and several 2-multigram models). The mean ranks, and the related confidence intervals allow us to compare these models.

As the representation of the prediction frame, we used a multivaluated graph. The search for the prediction rank can be then viewed as a graph traversal. The weak path in the graph is the best prediction for the language model, the second weakest is the second prediction, ...We used an A^* algorithm to carry out this N best path search. To reduce space and time combinatorial issues, we proposed in [15] several pruning heuristics (the path with higher cost than the frame to be predicted can be safely pruned).

Shannon established a bond between a rank distribution and a crossed entropy of the language model and a test corpus. Shannon proposed two bounds for the entropy, which are only valid for monotonous decreasing distributions (which is the case for power laws distributions). We carry out some experiments (Wall Street Journal, 30 000 words vocabulary, 12 000 predictions frames, several language models), and we obtain, for multiple words prediction frames, a *second* mode in the ranks distributions. This second mode prohibit us from the use of the Shannon's bounds. In [18], we proposed a short grammatical analysis of the frames involved in this second mode. With this analysis, we conclued that these frames are generally extremely difficult to predict (use of proper names, numbers, etc), or placed on a gap in the sentence (presence of a comma between the history and the window to be predicted, etc). We argue that an analysis of the rank distributions, based only on the first mode of these distributions remains valid.

6.4.5. Voice transformation

Participants: Nelly Barbot, Vincent Barreaud, Damien Lolive, Larbi Mesbahi, Olivier Boëffard.

This research field is addressed on two levels: segmental and supra-segmental. The study of segmental voice transformation is conducted in the framework of Larbi Mesbahi's PhD thesis which is funded by a project initiated by la Région Bretagne, 7.3. This thesis started on November 2006. In this thesis, the relevant segmental criteria can be linked to the signal's source (strength of the air flow, breathyness, harmonic fullness, etc) as well as to the speaker articulators (timbre, nasal flow, etc). Our methodology is based on the observation of large corpora. Relevant speech phenomena will be modeled as random variables. Some of them will be unobserved or latent. Our main goal is to perform unsupervised learning of both the segmentation and the transformation functions. The thesis will adopt the following framework:

- Acoustic signal uttered by the *source* and *target* speakers will be parametrized by LSF or MFCC vectors. LSF vectors are derived from LPC vectors which are computed with a covariance method and model for speech signal's source. The transformed voice's source signal will thus be obtained by reverse filtering. This source, combined with (a yet to defined) glottal model, will give the full speech signal of transformed voice.
- The speaker acoustic spaces will be modeled by Gaussian mixtures. Two Gaussian mixtures can be used to model the signal's source and the vocal tract separately, for each speaker. The coupling of those two models will be studied. The models parameter will be estimated by EM algorithm [82]. The number of classes is yet to be determined. It will be computed thanks to the minimum description length criterion (MDL).
- Finally, we will set the transformation functions between the acoustic spaces of the *source* speaker and the *target* speaker. The quality of this transformation will be evaluated by objective measurements as well as perceptive tests.

Concerning the transformation of the supra-segmental level, this work is covered within the framework of the PhD thesis of Damien Lolive, funded by the French Ministry of National Education and Research, and started on October 2005. In this study, we are particularly interested in the melodic contour transformation from a source speaker into a target speaker.

The first step of this work concerns F_0 contours modeling assuming a parametric model based on a B-spline model. This model has smoothing properties and local irregularities which capture the global shape of the F_0 curve and the breaks of curvature and discontinuities. Moreover, few parameters are needed to characterize a B-spline curve, they are the degree of the B-spline, the number of knots, the location of knots and control points. A B-spline curve of degree m is the sum of the control points weighted by B-spline functions of degree m. Between two successive knots, these B-spline functions are non-negative polynomial functions of degree m and their degrees of continuity at a knot depend on the knot multiplicity. For a given degree (generally $m = 3$) and sequence of knots, the control points are estimated using the least-squares error criterion. For what concerns the knot placement and multiplicity, we propose a global optimization algorithm using a simulated annealing procedure [34].

The comparison between a spline model, providing a regular curve, and a B-spline model is realized in [22]. On a methodological point of view, we compare these two models with equal degrees of freedom. Performances of both models are evaluated thanks to confidence intervals on the mean estimation error. The results show that the confidence intervals enable to dissociate the two models and underline the higher accuracy of the B-spline model which intrinsically take into account the irregularities of the melodic contours.

The next step of this work [25] introduces a criterion that enables selecting automatically the model number of knots. This selection is based on the parsimony principle and is obtained by applying an MDL (Minimum Description Length) methodology. This criterion gives a compromise between the model's complexity and its accuracy in estimating the F0 contour. Besides, this criterion makes possible to have parameters with variable precision in function of the curve to estimate. Consequently, the criterion automatically adapts to each curve to obtain the best compromise. Experimentally speaking, we use a 7000 syllables corpus. For each syllable, the optimal parameter set is estimated with a simulated-annealing algorithm. We store, for each syllable, the model that minimizes the MDL criterion, that is to say the best compromise between precision and complexity. The results shows a quite low mean RMS (Root Mean Square) error on the corpus $(0.55Hz)$ and a limited number of freedom degrees (68% of the full model).

In addition, in [25], we made the general hypothesis according to which the knots of the B-spline model are not necessarily placed at observation places. Consequently, searching the optimal model is complex and expensive in terms of computing time, and the estimation error is finally worse. Thus, we can restrict the position of the knots on observation places. Thereby, in [24], [23], the mean RMS error on the corpus is lower $(0.42Hz)$ with a number of degrees of freedom smaller (63% of the full model).

7. Contracts and Grants with Industry

7.1. Néologos

The main topic of this project relates to the creation of new telephone vocal data bases for the French language. This project has two main objectives : a multi-speaker speech database with children voices (1000 speakers) and a multi-speaker speech database with adults voices.

Cordial is mainly concerned with the second task. We aim to define, for French, a speech database of reference speakers, i.e. a speech corpus where each speaker will have pronounced sufficient statements so that one can exploit them to characterize his voice. To achieve this goal, we need more than only 50 statements to record for each speaker. We plan to record a database where 200 reference speakers have recorded over the fixed phone network 500 well defined statements to cover the main coarticulation features of the language.

In addition to speech recognition systems, such a corpus is also useful for the research and the development of the techniques of speaker identification and authentification, voice transformation, voice characteristics for Test-To-Speech systems.

The partners of the project are of three types :

- Academic laboratories undertaking an active research on vocal technologies (IRISA, LORIA, and FTR&D), whose main contribution will be done on the supply of research tools and on the realization of validation tests.
- Industrial partners (TELISMA and DIALOCA) marketing products of speech recognition, whose contribution will be done by the organization itself of the collection and the realization of "industrial" tests more intended to show the contribution of the corpus for the improvement of the products.
- The ELDA (European Language Resources Distribution Agency) whose vocation is to distribute linguistic resources, and who leads an activity of creation of corpus.

7.2. Vivos

In July, 2006, Cordial entered a RIAM[12] (Research and Innovation Audiovisual and Multimedia Innovation Network) project focused on synthesis of expressive voices for multimedia applications, such as movie dubbing or games for mobile phones.

Three research themes are addressed by this project: expressive speech synthesis, voice transformation, and speech synthesis from recordings initially non-dedicated to this task.

Cordial contribution mainly takes its place in the third theme. To this day, speech synthesis is based on speech units from large amounts of natural speech recorded by professional speakers. One goal of the VIVOS project is to perform speech synthesis from any spontaneous speech records. The task of the Cordial projet is to realize phoneme segmentations of speech corpora from dubbing or documentary spontaneous speech records.

To achieve this, Cordial hired Gaëlle Vidal in July 2006, for manual segmentation, and Laure Charonnat in October 2006, for automatic segmentation, extending the study realised by Samir Nefti [68].

Vivos partners are:

- Academic laboratories undertaking an active research on vocal technologies (IRISA, IRCAM and FTR&D),
- Industrial partners: Chinkel, a dubbing company, and Betomorrow, developer of mobile applications.

7.3. Transpar

A contract between la Région Bretagne and the l'Université de Rennes 1 - ENSSAT allows a PhD student, Larbi Mesbahi, working on the field of voice transformation for a three year period.

[12] Recherche et Innovation en Audiovisuel et Multimedia

7.4. Dialogue and Semantics

In 2003 has been finalized the CRC[13] "Machine learning in man-machine interaction" between the Cordial project and France Télécom Recherche et Développement, DIH/DII, Lannion.

The subject is of common interest between our two research units. The CRC federates all the manpower in both teams involved on the topic. It covers the thesis of P. Alain, described at section 3.5.4, another thesis at FTRD DIH/DII, started Feb 2002 and the thesis of E. Livolant started in January 2004 and stopped at the end of 2005. The total manpower in permanent researchers is of 0.125 man-year at FTRD and at Cordial (scientific management of the CRC and direction of the thesis).

8. Other Grants and Activities

8.1. International networks and workgroups

The Cordial team is a member of the European Network of Excellence in Human Language Technologies Elsnet, and of the French-speaking network FRANCIL (Réseau FRANCophne d'Ingénierie de la Langue).

9. Dissemination

9.1. Leadership within scientific community

Laurent Miclet has been the President of the scientific committee of the French Machine Learning Congress [12], Conférence d'Apprentissage *CAp 2006*.

Arnaud Delhay has been president of the organisation comittee of *CAp 2006* and Olivier Boëffard has been a member of the same comittee.

Olivier Boëffard has been member of the organisation comittee of *JEP 2006* (Journées d'Étude sur la Parole).

9.2. Teaching at University

Olivier Boëffard teaches the course *Speech Synthesis* in the Master STIR, Rennes 1 (option Signal, orientation 2) and takes part in the module Data Mining (*Fouille de données*) in the Master Informatique de Rennes 1.

Marc Guyomard and Jacques Siroux teach the module *human-machine communication* at Enssat, Lannion (Lannion part of the Master Informatique de Rennes 1).

Laurent Miclet teaches a course in Pattern Recognition *Reconnaissance des Formes* in the Master STIR and a part of the module *Apprentissage et Classification* (AC) in the Master Informatique de Rennes 1. In the Lannion part of the Master Informatique de Rennes 1, for which he is the coordinator, he teaches a module of Machine Learning *Apprentissage Artificiel* and takes part in the module Data Mining (*Fouille de données*).

Laurent Miclet has been one of the two coordinators of the french translation ([13]) of the book *Artificial Intelligence*, by S. Russel and P. Norvig.

9.3. Conferences, workshops and meetings, invitations

Laurent Miclet has been a reporter for the Ph.D. thesis: G. Valétudie. *Nouvelles méthodes en data-mining et extraction de connaissances à partir de données : application au complexe Mycobacterium Tuberculosis.* Thèse de l'Université des Antilles et de la Guyane, 15th of September 2006.

Olivier Boëffard has been a reader for the Ph.D. thesis: G. Valétudie. *Prise en compte de critères acoustiques pour la synthèse de la parole* Thèse ENST / Université de Rennes 1, 27th of September 2006.

[13] *Contrat de Recherche Coopérative*, Cooperative Research Contract

9.4. Graduate Student and Student intern

We have this year two Master students in a research period.

10. Bibliography

Major publications by the team in recent years

[1] O. BOËFFARD, C. D'ALESSANDRO. *Synthèse de la parole*, Hermès Science, New-York, 2002.

[2] A. CORNUÉJOLS, L. MICLET. *Apprentissage artificiel : méthodes et algorithmes*, Eyrolles, 2002.

[3] A. DELHAY, L. MICLET. *Analogie entre séquences : Définitions, calcul et utilisation en apprentissage supervisé.*, in "Revue d'Intelligence Artificielle.", vol. 19, 2005, p. 683–712.

[4] P. DUPONT, L. MICLET, E. VIDAL. *What is the search space of the regular inference ?*, in "Lecture Notes in Artificial Intelligence, Grammatical Inference and Applications, Berlin, Heidelberg", vol. 862, Springer Verlag, September 1994.

[5] H. FRANÇOIS, O. BOËFFARD. *The greedy algorithm and its application to the construction of a continuous speech database*, in "Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC)", vol. 5, 2002.

[6] H. FRANÇOIS, O. BOËFFARD. *Evaluation if units selection criteria in corpus-based speech synthesis*, in "Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech), Geneva, Switzerland", 2003, p. 1325–1328.

[7] H. FRANÇOIS. *Synthèse de la parole par concaténation d'unités acoustiques : construction et exploitation d'une base de parole continue*, Ph. D. Thesis, Université de Rennes 1, 2002.

[8] M. GUYOMARD, P. NERZIC, J. SIROUX. *Plans, métaplans et dialogue*, Technical report, n^{o} 1169, Irisa, September 1998.

[9] S. NEFTI, O. BOËFFARD, T. MOUDENC. *Confidence measure for phonetic segmmentation of continuous speech*, in "Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech), Geneva, Switzerland", 2003, p. 897–900.

[10] J. SIROUX, M. GUYOMARD, F. MULTON, C. RÉMONDEAU. *Oral and gestural activities of the users in the géoral system*, in "Intelligence and Multimodality in Multimedia, Research and Applications, AAAI Press", John Lee (ed), 1998.

[11] F. VIOLARO, O. BOËFFARD. *A hybrid model for text-to-speech synthesis*, in "IEEE Transactions on Speech and Audio Processing", vol. 6, n^{o} 5, 1998, p. 426–434.

Year Publications

Books and Monographs

[12] L. MICLET (editor). *Actes de la 8ème Conférence Francophone sur l'Apprentissage Automatique (CAP 2006)-Trégastel, France*, Presses Universitaires de Grenoble, 2006.

[13] S. RUSSELL, P. NORVIG. *Intelligence artificielle*, Coordination de la traduction : L. Miclet et F. Popineau, Pearson, 2006.

Articles in refereed journals and book chapters

[14] S. KRSTULOVIC, F. BIMBOT, O. BOËFFARD, D. CHARLET, D. FOHR, O. MELLA. *Optimizing the coverage of a speech database through a selection of representative speaker recordings*, in "Speech Communication", vol. 48, n^{o} 10, 2006, p. 1319-1348.

Publications in Conferences and Workshops

[15] P. ALAIN, O. BOËFFARD. *Algorithme de recherche d'un rang de prédiction. Application à l'évaluation de modèles de langage*, in "Actes des XXVIèmes Journées d'Etudes sur la Parole, Dinard, France", 2006, p. 321-324.

[16] P. ALAIN, O. BOËFFARD. *Evaluation de modèles de langage indépendamment d'une tâche. Extension du jeu de Shannon à la prédiction de séquences de mots conjoints*, in "Actes de la 8ème Conférence Francophone sur l'Apprentissage Automatique - Trégastel, France", L. MICLET (editor). , Presses Universitaires de Grenoble, 2006, p. 389-390.

[17] P. ALAIN, O. BOËFFARD. *Using a general rank-based statistics framework to evaluate language models*, in "Proceedings of the 11th International Conference on Speech and Computer (SPECOM), Saint Petersburg, Russia", 2006, p. 457-462.

[18] P. ALAIN, O. BOËFFARD, N. BARBOT. *Evaluating language models within a predictive framework: an analysis of ranking distributions*, in "Lecture Notes in Artificial Intelligence - Proceedings of the 9th International Conference on Text, Speech and Dialogue - Brno, Czech Republic, Berlin, Heidelberg", P. SOJKA, I. KOPEČEK, K. PALA (editors). , vol. 4188, Springer Verlag, 2006, p. 319-326.

[19] S. BAYOUDH. *Apprentissage par analogie : classification de données binaires et nominales*, in "Actes de la 8ème Conférence Francophone sur l'Apprentissage Automatique - Trégastel, France", L. MICLET (editor). , Presses Universitaires de Grenoble, 2006, p. 299–314.

[20] D. CADIC, A. L. FORESTIE, E. GOUGIS, T. MOUDENC, A. FURBY, O. BOËFFARD. *Etude préliminaire d'une nouvelle synthèse vocale destinée aux patients atteints de sclérose latérale amyotrophique*, in "Journées de Neurologie de Langue Française, Toulouse, France", 2006.

[21] A. CHOUMANE, J. SIROUX. *Toward a Generic Model Including Knowledge and Treatments for Multimodal Reference Resolution*, in "Proceedings Inscit2006, Mérida - Spain", V. P. GUERRERO-BOTE (editor). , vol. 2, n^{o} 2, 2006, p. 298-302.

[22] D. LOLIVE, N. BARBOT, O. BOËFFARD. *Comparing B-spline and spline models for F0 modelling*, in "Lecture Notes in Artificial Intelligence - Proceedings of the 9th International Conference on Text, Speech and Dialogue - Brno, Czech Republic, Berlin, Heidelberg", P. SOJKA, I. KOPEČEK, K. PALA (editors). , vol. 4188, Springer Verlag, 2006, p. 423-430.

[23] D. LOLIVE, N. BARBOT, O. BOËFFARD. *Melodic contour estimation with B-spline models using a MDL criterion*, in "Proceedings of the 11th International Conference on Speech and Computer (SPECOM), Saint Petersburg, Russia", 2006, p. 333-338.

[24] D. LOLIVE, N. BARBOT, O. BOËFFARD. *Modélisation B-spline de contours mélodiques avec estimation du nombre de paramètres libres par un critère MDL*, in "Actes des XXVIèmes Journées d'Etudes sur la Parole, Dinard, France", 2006, p. 499-502.

[25] D. LOLIVE, N. BARBOT, O. BOËFFARD. *Proposition d'un critère MDL pour l'estimation de courbes ouvertes modélisées par des B-splines*, in "Actes de la 8ème Conférence Francophone sur l'Apprentissage Automatique - Trégastel, France", L. MICLET (editor). , Presses Universitaires de Grenoble, 2006, p. 219-234.

[26] S. SAGET, M. GUYOMARD. *Goal-oriented Dialog as a Collaborative Subordinated Activity involving Collaborative Acceptance*, in "Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue (Brandial 2006), University of Potsdam, Germany", 2006, p. 131-138.

[27] S. SAGET. *In favour of collective acceptance: Studies on goal-oriented dialogues*, in "Proceedings of Collective Intentionality V, Helsinki, Finland", 2006.

References in notes

[28] R. DE MORI (editor). *Spoken Dialogue with Computers*, ISBN 0122090551, Academic Press, 1998.

[29] V. HONAVAR, C. DE LA HIGUERA (editors). *Machine Learning Journal - Special Issue on grammatical inference*, 1-2, vol. 44, 2001.

[30] M. ABE, S. NAKAMURA, K. SHIKANO, H. KUWABARA. *Voice conversion through vector quantization*, in "Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'88)", 1988, p. 655–658.

[31] J. ALLEN. *Natural language understanding*, Benjamin/Cummings Menlo Park, 1987.

[32] M. AURNAGUE. *A unified processing of orientation for internal and external localization*, Groupe Langue, Raisonnement, Calcul, Toulouse, France, 1993.

[33] J. AUSTIN. *Quand dire c'est faire*, Editions du seuil, Paris, France, 1970.

[34] N. BARBOT, O. BOËFFARD, D. LOLIVE. F_0 *stylisation with a free-knot B-spline model and simulated annealing optimization*, in "Proceedings of the 9th European Conference on Speech Communication and Technology (Eurospeech), Lisbon, Portugal", 2005, p. 325–328.

[35] D. BIGORGNE, O. BOËFFARD, B. CHERBONNEL, F. EMERARD, D. LARREUR, J. L. LE SAINT-MILON, I. MÉTAYER, C. SORIN, S. WHITE. *Multilingual PSOLA text-to-speech system*, in "IEEE International Conference on Acoustics, Speech, and Signal Processing", vol. 2, 1993, p. 187–190.

[36] L. BLIN. *Apprentissage de structures d'arbres à partir d'exemples : application à la prosodie pour la synthèse de la parole*, Ph. D. Thesis, IRISA – Université de Rennes 1, 2002.

[37] L. BLIN. *Génération de prosodie par apprentissage de structures arborescentes*, in "Actes de la Conférence d'Apprentissage, Laval, France", 2003.

[38] A. BORILLO. *Le lexique de l'espace : les noms et les adjectifs de localisation interne*, in "Cahiers de grammaire", vol. 13, 1988, p. 1–22.

[39] B. BOZKURT, O. OZTURK, T. DUTOIT. *Text design for TTS speech corpus building using a modified greedy selection*, in "Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech), Geneva, Switzerland", 2003, p. 277-280.

[40] O. BOËFFARD, F. EMERARD. *Application-dependent prosodic models for text-to-speech synthesis and automatic design of learning database corpus using genetic algorithm*, in "Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech), Rhodes, Greece", vol. 5, 1997, p. 2507–2510.

[41] S. E. BRENNAN, H. H. CLARK. *Conceptual pacts and lexical choice in conversation*, in "Journal of Experimental Psychology: Learning, Memory and Cognition", vol. 22, 1996, p. 482-1493.

[42] G. BRETON. *Modélisation d'un contexte cartographique et dialogique*, ENSSAT, Technical report, DEA Informatique de Rennes 1, 1998.

[43] F. BRUGNARA, D. FALAVIGNA, M. OMOLOGO. *Automatic Segmentation and Labeling of Speech based on Hidden Markov Models*, in "Speech Communication", vol. 12, 1999, p. 357–370.

[44] A. CAPRARA, M. FISCHETTI, P. TOTH. *Algorithms for set covering problem*, Technical report, n^{o} 406, IASI-CNR, Rome, 1995.

[45] S. CERIA, P. NOBILI, A. SASSANO. *A Lagrangian-based heuristic for large-scale set covering problems*, in "Mathematical Programming", vol. 81, 1998, p. 215-228.

[46] H. CLARK, A. BANGERTER. *Changing conceptions of reference*, in "Experimental pragmatics, Basingstoke, England", I. NOVECK, D. SPERBER (editors). , Palgrave Macmillan, 2004, p. 25-49.

[47] A. COZANNET, J. SIROUX. *Strategies for oral dialogue control*, in "Proceedings of International Conference on Spoken Language Processing (ICSLP), Yokohama, Japan", vol. 2, 1994, p. 963–966.

[48] P. DUPONT, L. MICLET. *L'inférence grammaticale régulière : fondements théoriques et principaux algorithmes*, Technical report, n^{o} 3449, INRIA, 1998, http://hal.inria.fr/inria-00073241.

[49] H. FRANÇOIS, O. BOËFFARD. *Design of an optimal continuous speech database for text-to-speech synthesis considered as a set covering problem*, in "Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech), Aalborg, Denmark", 2001, p. 829-833.

[50] H. FUJISAKI. *Information, prosody, and modeling - with emphasis on tonal features of speech*, in "Proceedings of Speech Prosody Conference, Nara, Japan", 2004, p. 1–10.

[51] D. HIRST, A. D. CRISTO, R. ESPESSER. *Levels of representation and levels of analysis for the description of intonation systems*, in "Prosody : Theory and Experiment", vol. 14, 2000, p. 51–87.

[52] G. KLEIN, D. WOODS, J. BRADSHAW, R. HOFFMAN, P. FELTOVICH. *Ten Challenges for Making Automation a 'Team Player' in Joint Human-Agent Activity*, in "IEEE Intelligent Systems", 2004, p. 91-95.

[53] A. KRONFELD. *Goals of referring acts*, in "TINLAP-3", 1987, p. 143–149.

[54] A. KRUL, G. DAMNATI, F. YVON, T. MOUDENC. *Corpus design based on the Kullback-Leibler divergence for text-to-speech synthesis application*, in "Proceedings of the International Conference on Spoken Language Processing (ICSLP), Pittsburg, USA", 2006, p. 2030–2033.

[55] J. L'HOUR, O. BOËFFARD, J. SIROUX, L. MICLET, F. CHARPENTIER, T. MOUDENC. *DORIS, a multiagent/IP platform for multimodal dialogue applications*, in "Proceedings of the International Conference on Spoken Language Processing (ICSLP), Jeju Island, Korea", 2004, p. 736-739.

[56] J. LAROCHE, Y. STYLIANOU, E. MOULINES. *HNM: a simple, efficient harmonic+noise model for speech*, in "IEEE Workshop on Applications for Signal Processing to audio and Acoustics", 1993.

[57] M. LEE. *Rationality, Cooperation and Conversational Implicature*, in "Proceedings of the Ninth Irish Conference on Artificial Intelligence", 1997.

[58] Y. LEPAGE, S.-I. ANDO. *Saussurian analogy: a theoretical account and its application*, in "Proceedings of COLING-96, København", 1996, p. 717–722, http://www.slt.atr.co.jp/~lepage/ps/coling96.ps.gz.

[59] Y. LEPAGE. *De l'analogie rendant compte de la commutation en linguistique*, Habilitation à diriger les recherches, Université Joseph Fourier, Grenoble, 2003.

[60] D. LEWIS. *Scorekeeping in a language game*, in "Journal of Philosophical Logic", vol. 8, 1979, p. 339-359.

[61] D. J. LITMAN. *Plan Recognition and Discourse Analysis : An Integrated Approach for Understanding Dialogues*, Ph. D. Thesis, University of Rochester, TR 170, 1985.

[62] M. MAYBURY. *Communicative Acts for Explanation Generation*, in "International Journal of Man-machine studies", vol. 37(2), 1990, p. 135–172.

[63] P. MERTENS. *Synthesizing elaborate intonation contours in text-to-speech for french*, in "Proceedings of the Speech Prosody Conference, Aix-en-Provence, France", 2002, p. 499-502.

[64] M. L. MICÓ, J. ONCINA, E. VIDAL. *A new version of the nearest-neighbour approximating and eliminating search algorithm (AESA) with linear preprocessing time and memory requirements*, in "Pattern Recognition Letters", vol. 15, n^{o} 1, 1992, p. 9-17.

[65] R. MITKOV. *Anaphora resolution*, Lonman, 2002.

[66] S. MOULINE, 0. BOËFFARD, P. BAGSHAW. *Automatic adaptation of the Momel F_0 stylisation algorithm to new corpora*, in "Proceedings of the International Conference on Spoken Language Processing (ICSLP), Jeju Island, Korea", 2004, p. 961–964.

[67] E. MOULINES, F. CHARPENTIER. *Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones*, in "Speech Communication", vol. 9, 1990, p. 453–467.

[68] S. NEFTI, O. BOËFFARD. *Acoustical and topological experiments for an HMM-based speech segmentation system*, in "Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech), Aalborg, Denmark", 2001.

[69] P. NERZIC, M. GUYOMARD, J. SIROUX. *Reprise des échecs et erreurs dans le dialogue homme-machine*, in "Cahiers de linguistique sociale", vol. 21, 1992, p. 35–46.

[70] J. ONCINA, P. GARCÍA, E. VIDAL. *Learning subsequential transducers for pattern recognition and interpretation tasks*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", vol. 15, 1993, p. 448-458.

[71] M. PICKERING, S. GARROD. *Toward a mechanistic psychology of dialogue*, in "Behavioral and Brain Sciences", vol. 27, n^{o} 169-225, 2004.

[72] L. RABINER. *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*, in "Proceedings of the IEEE", vol. 77, n^{o} 2, 1989, p. 257–286.

[73] K. N. ROSS, M. OSTENDORF. *A dynamical system model for generating fundamental frequency for speech synthesis*, in "IEEE Transactions on Speech and Audio Processing", vol. 7, n^{o} 3, 1999, p. 295–309.

[74] S. RUSSELL, P. NORVIG. *Artificial Intelligence: A Modern Approach*, 2nd edition, Prentice-Hall, Englewood Cliffs, NJ, 2003.

[75] S. SAKAI, J. GLASS. *Fundamental frequency modeling for corpus-based speech synthesis based on statistical learning techniques*, in "Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop, St. Thomas, U.S. Virgin Islands", 2003, p. 712–717.

[76] J. SEARLE. *Sens et expression*, Les éditions de minuit, 1982.

[77] C. E. SHANNON. *A mathematical theory of communication*, in "Bell System Technical Journal", vol. 27, 1948, p. 379–423.

[78] C. E. SHANNON. *Prediction and entropy of printed English*, in "Bell System Technical Journal", vol. 30, n^{o} 1, 1951, p. 50-64.

[79] J. SIROUX, ET AL.. *Multimodal References in Georal Tactile*, in "Proceedings of the workshop Referring Phenomena in a multimedia Context and their Computational Treatment, SIGMEDIA and ACL/EACL, Madrid", 1997, p. 39–44.

[80] R. STALNAKER. *Pragmatic presuppositions*, in "Context and Content", 1974, p. 47-62.

[81] N. STROPPA. *Définitions et caractérisations de modèles à base d'analogies pour l'apprentissage automatique des langues naturelles*, Ph. D. Thesis, Ecole Nationale Supérieure des Télécommunications, 2005.

[82] Y. STYLIANOU, O. CAPPE, E. MOULINES. *Continuous probabilistic transform for voice conversion*, in "IEEE Transactions on Speech and Audio Processing", vol. 6, n^{o} 2, March 1998, p. 131 – 142.

[83] SUNDIAL. *SUNDIAL, Prototype performance evaluation report*, Deliverable, n^{o} D3WP8, projet Sundial P2218, September 1993.

[84] P. TAYLOR, A. BLACK. *The architecture of the Festival speech synthesis system*, in "Proceedings of the 3rd ESCA Workshop on Speech Synthesis", 1998, p. 323–327.

[85] J. TAYLOR, J. CARLETTA, C. MELLISH. *Requirements for belief models in cooperative dialogue*, in "User Modeling and User-Adapted Interaction", vol. 6, n^{o} 1, 1996, p. 23–68, http://citeseer.ist.psu.edu/232252.html.

[86] P. TAYLOR. *Analysis and synthesis of intonation using the Tilt model*, in "Journal of the Acoustical Society of America", vol. 107, 2000, p. 1697-1714.

[87] P. TAYLOR. *Concept-to-speech by phonological structure matches*, in "Philosophical Transactions of the Royal Society, Series A", vol. 358(1769), 2000, p. 1403–1416.

[88] J. VAN SANTEN, A. BUCHSBAUM. *Methods for optimal text selection*, in "Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech), Rhodes, Greece", 1997, p. 553–556.

[89] C. VANDELOISE. *L'espace en français*, Éditions du seuil, Paris, 1986.

[90] T. VERMA, S. LEVINE, T. MENG. *A Flexible Transient Analysis/Synthesis Tool for Transient Signals*, in "International Conference of Computer Music", 1997.

[91] E. VIDAL, F. CASUBERTA. *Learning Finaite-State Models for Machine Translation*, in "Proceedings of the 7th International Colloquium on Grammatical Inference", 2004.

[92] R. VIEIRA, M. POESIO. *An empirically based system for processing definite descriptions*, in "Computational Linguistics", vol. 26, n^{o} 4, 2000, p. 539-545.

[93] L. VIEU. *Sémantique des relations spatiales et inférences spatio-temporelles : une contribution à l'étude des structures formelles de l'espace en langage naturel*, Ph. D. Thesis, Université Paul Sabatier, Toulouse, 1991.

[94] D. VINCENT, O. ROSEC, T. CHONAVEL. *Estimation of LF glottal source parameters based on ARX model*, in "Proceedings of the Interspeech Conference", 2005.

[95] K. WRAY. *Collective Belief and Acceptance*, in "Synthese", vol. 129, 2001, p. 319-333.

[96] F. YVON. *Des apprentis pour le Traitement Automatique des Langues*, Habilitation à diriger des recherches, Université Paris 6, 2006.

[97] C. DE LA HIGUERA. *A Bibliographical Study of Grammatical Inference*, in "Pattern Recognition", 2005.

[98] C. DE LA HIGUERA, J. ONCINA. *Learning Context-Free Languages*, in "To appear in Artificial Intelligence Review", 2006.

[99] C. DE LA HIGUERA, F. THOLLARD, E. VIDAL, F. CASACUBERTA, R. CARRASCO. *Probabilistic finite-state machines - Part I*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", 2005.

[100] C. DE LA HIGUERA, F. THOLLARD, E. VIDAL, F. CASACUBERTA, R. CARRASCO. *Probabilistic finite-state machines - Part II*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", 2005.