INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team cordial

Man-machine oral and multimodal communication

Rennes

THEME COG

Table of contents

1. Team

Head of project (Responsable scientifique)

Laurent Miclet [Faculty member (professeur), Enssat]

Faculty members (Personnel Université Rennes 1)

Nelly Barbot [maître de conférences, Enssat]
Olivier Boëffard [maître de conférences, Enssat]
Arnaud Delhay [maître de conférences, IUT]
Marc Guyomard [professeur, Enssat]
Jean-Christophe Pettier [maître de conférences, Enssat]
Jacques Siroux [professeur, IUT]

Ph. D. Students

Pierre Alain [bourse INRIA, from October the 1st, 2003]
Samir Nefti [bourse EGIDE, defense December the 16th, 2004]
Sylvie Saget [bourse Région, from October the 15th, 2003]
Salma Mouline [FTR&D, from January the 1st, 2003]
Sabri Bayoudh [bourse INRIA, from October the 1st, 2004]
Josselin Huaulmé [CIFRE Télisma, from October the 1st, 2003]
Erwan Livolant [FTR&D, from February the 1st, 2004]

Technical staff

Johann L'Hour [Junior technical staff (Ingénieur associé INRIA),until November the 1st, 2004]

2. 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 very 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 [19], 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 dealing 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 the models met at the different stages of a dialogue system with an oral component.

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. 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 [30]. A last 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 *coding and 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*.

speech act: In the Speech Act Theory initiated by Austin [24] and developed by Searle [56], the main axiom claims that the emission of a utterance can be assimilated to the performance of actions which modify the mental states of the participants.

plan : sequence of actions which aims at the realization of an intentional goal.

plan recognition: Recognizing a plan given a sequence of observed actions consists in identifying underlying relations between these actions in order to guess goals and possible continuations of the current plan.

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 [24][56] 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 [20][41] 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,)

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 [5] [49] in order to associate three points of view: 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 [41] 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. So, 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.* human and system) can indeed have some erroneous knowledge about the application,

about other's abilities and about current elements dialogue as 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 the 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: *multimodality*, *reference*.

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 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.

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 concerned with the ways by which information coming from the 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 [43], performed in a different context (the generation of communicative acts for the system ouput). Maybury proposes several levels of communicative acts (type of speech act) which allow to integrate at each level information coming from different modalities. We take up 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 [60] 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) [62] 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 [58][35] (for example: length of dialogue, number of speech turn 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.

3.4. 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 ([37]).

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, [61][52]). 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 7.2). 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_a which is the closest to x and adapts to x a prosody p_x directly deduced from p_a.

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 [27][26].

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 [40] and of Yvon [63] 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 [39]:

Symmetry of the 'as' relation:	$c : d :: a : b$
Exchange of the means:	$a : c :: b : d$

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):

$a : a :: b : x \Rightarrow x = b$
$a : b :: a : x \Rightarrow x = b$

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 [40] and the very recent work of Yvon.

Our approach to solving equations on sequences is based on classical edit distance and uses deletion, insertion and substitution. We did not assume that distribution 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 [46], 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 speech units for speech synthesis

Text-to-speech synthesis(TTS) can be carried out by concatenation of acoustic units obtained from a continuous speech database. The state-of-the-art TTS systems consist in juxtaposing pre-recorded acoustic units, typically phones, diphones or units of non-uniform length.

An alternative to the production of speech from a dictionnary of diphones consists in using a indexed corpus of continuous speech [59]. When one has to produce a sequence of phonemes, the idea is to get in the corpus the best acoustic sequence. It is selected according to several criteria: its phonetic correspondance, its length, position in the sentence, etc. The relative importance of these criteria can be tuned by learning.

The multiple representation of these configurations at the acoustic and phonological levels enables voice quality to be improved significantly [51][25]. Furthermore, one can consider that the acoustic segments used to build an acoustic utterance no longer have predefined linguistic definition. We consider here that the phonological units are not defined *a priori* over a finite set of phonemes. Therefore, we face a combinatorial issue where linguistic units [55], that we no longer know, are useful to parse a phoneme sequence in order to find the best acoustic segments.

Our methodological research framework try to answer the following points :

- From a linguistic point of view, how can we automatically build a set of phonological units [29] ?
- From an acoustic quality point of view, given a continuous speech database with phone labels, how can we characterize the best sequence of linguistic units?
- From an algorithmic point of view, in a graph search framework, what are the best heuristics to solve this combinatorial problem [38] ?
- From a pragmatic point of view, given a target application, what could be the best set of pre-recorded speech sentences yielding the best TTS quality?

3.4.5. Automatic Speech Labelling and Recognition

Machine learning methodologies based on statistical approaches require databases of relatively consequent size. The examples taken from these databases show the relationship between the numerous variables involved in the studied phenomenon. In voice synthesis as well as in voice recognition, one wishes to be able to have an explicit relation between the acoustic level and the phonological level. If an automatic labelling starting from phonetic sequences is a task which finds acceptable solutions, the process is more complex when only the text is known.

In such a context, given a speech utterance realized by a speaker and its particular phonetic transcription, the precise location of temporal marks delimiting phone boundaries on the speech is required. The state-of-the-art systems use a markovian description of the speech in an appropriate acoustic space [32].

Sequences of Hidden Markov Models (HMM) are built from the phonetic description of the acoustic observations. As one needs to discriminate phone boundaries, the majority of the phone segmentation systems postulates a monophone modeling hypothesis. During a learning phase, the parameters of each phone model are learned through a set of exemples using the well known EM iteration scheme [53]. During a decoding phase, the segmentation system finds the most probable alignment between the sequence of models and the observations. The temporal marks delimiting each phone are easily recovered using the model transitions on the optimal path alignment.

We propose to weaken our previous hypothesis by relaxing the exact phonetic transcription with the exact phonemic sequence. The phonemic sequence is built automatically from the text. Under the same phonemic symbol, various acoustic realizations can be found depending on the coarticulation context of the realized phone.

Firstly, we developed a baseline speech segmentation system based on HMM – Hidden Markov Models as briefly exposed section 3.4.5. The scores of segmentation which we obtained are equivalent to those of the state-of-the-art systems from literature. Moreover, to analyze the behavior of this baseline segmentation system, we carried out experiments on two axes : the topological definition of the HMM and the acoustic analysis of the speech [48].

Secondly, we focus our activity on the automatic phone labelling from the text, not from the true phonetic sequence. Given an automatic phonemic transcription from the text, various acoustic realizations can be found depending on the coarticulation context of the realized phone and depending on the speaker. Since the HMM framework is well adapted to introduce variants of pronunciation, all is needed is to extend the graph of model hypotheses and let the decoding phase find the best alignment. The main drawback of this scenario is that as the degrees of freedom increase, the system becomes instable and less accurate.

3.4.6. Learning semantics from speech

Currently, many automatic systems delivering information suppose that the user of such a service must be able to adapt himself to the implicit requirements of the automatic system.

We postulate that a man-machine interface being based on the natural language must facilitate the access of the greatest number of us with this type of services [44]. Under this assumption, the machine must make the maximal effort to adapt itself to the user.

There already exists many information systems which technology is based on a man-machine oral dialogue. In a first stage, the speech, the entry of such a system, is translated into a sequence of words. This sequence of words will be then treated by a pragmatic entity taking into account a dialogue model. In return, the machine response is stated by a speech synthesis system starting from the text or a concept modeling.

Within this experimental framework, a problem which still today does not find satisfactory scientific and technological answers is that of the semantic treatment.

A semantic function in a context of natural speech processing has a double objective. Firstly, a pragmatic treatment carried out on a sequence of concepts is more relevant than carried out directly on a sequence of words – words are sensitive to the errors of the recognition system. Secondly, a system which would be able to *understand* the message can propose different alternatives, what cannot do the current speech synthesis systems starting from the text.

The proposed research framework is settled between the output of an automatic speech recognition system and the input of a dialogue management system. From the description of a statement recognized by the ASR system and translated into a lattice of words, the goal consists in providing the sequence of the underlying concepts. We propose to explore the temporal dimension of the sequence of the concepts in an oral statement starting from its relationship to syntax and more particulary the discovery of the set of thematic roles [54].

We will adopt a methodology based on the observation of corpus of examples within a statistical theoretical framework. However, it is difficult to find corpora annotated by semantic elements (particularly in French). The phenomenon will be described by random variables partially observed [36], [47]. Then, an objective consists in determining the optimal quantity of annotated information required for training and mixing them with unlabelled corpora under an assumption of unsupervised training [50] [22].

3.4.7. Learning prosody from speech

On the one hand, speech processing state of the art can efficiently model acoustic voice characteristics starting from a voice print. On the other hand, few studies are interested in the suprasegmental aspect of the speech, more precisely on the automatic modeling of melody contours [45] [33]. One could find multiple objectives, setting up automatic voice transformation systems, merging prosodic and acoustic information in a ASR system, tuning text-to-speech synthesis systems with *ad hoc* prosodic models.

We propose a model making it possible to describe a melody contour at the sentence level built over a sequence of elementary melody contours. The difficulty of modeling is that one does not *a priori* know an alphabet made of classes of elementary contours. They thus should be estimated starting from the observations of the sentence level all while being based on assumptions of parsimony [21].

A melody contour is a mono-dimensional signal with real values evolving according to time. We propose to take for methodological assumption the class of dynamical state space models [42]. We consider that the observation of a portion of the melody is explained by a stochastic state variable defining the equation of a standard Kalman filter under gaussian and linearities assumptions. We suppose that a portion of the melody curve followed by a Kalman filter corresponds to a class. The complete observation of the sentence level is governed by a time switching of Kalman filters. This switching process is modeled by an hidden Markov chain.

3.4.8. 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 7.2.

3.5. Language learning

Keywords: *Educational software*, *teaching and learning languages*.

The aim of this study is to design and to develop educational software for helping to teach and to learn languages.

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

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

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.

4. 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, Johann L'Hour.

The Cordial project aims to promote its research activities by means of technological demonstrations. To achieve this point, hardware and software ressources 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 contrat 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), SEMANTIC PARROT (see section 5.5), and ORDICTÉE (see section 5.4). An INRIA research engineer manages the technical aspects of the platform and develops 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 our 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[2] architecture. It satisfies our needs for modularity, quick and clean development and interoperability. To fulfill this role, we chose and installed JADE[3], 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[4] 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 our scientific interests.

Today, two projects have taken place inside the DORIS platform: GEORAL (see sections 5.3 and 6.2) and the SEMANTIC PARROT (see section 5.5).

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 our 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 at the Interspeech 2004 conference in Korea [15].

5.3. Georal

Participants: Jacques Siroux [*correspondant*], Marc Guyomard, Johann L'Hour.

[2] 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.

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

[4] 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. Improvements 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 nomralization). The dialogue 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 3.5, 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. Epigram

Participant: Laurent Miclet [*correspondant*].

The software library called EPIGRAM (Environnement de Programmation pour l'Inférence GRAMmaticale), has been developed between 1997 and 2001. EPIGRAM is a library of high level modules enabling the development of grammatical inference programs and applications. It has been written around a C++ environment called LEDA, a library of data types and combinatoric algorithms developed at the Max-Planck-Institut für Informatik, Saarbrücken, Germany. EPIGRAM has been written together with the team EURISE of the University of Saint-Étienne and the former IRISA project Aïda (F. Coste, now in the Symbiose project), as a part of a contract with France Telecom FT R&D (CTI 97 1B 004).

6. New Results

6.1. Dialogue and modeling

6.1.1. Logical modeling for dialogue processing

Participant: Jean-Christophe Pettier.

Dialogue systems have to model a world description to answer user requests. In order to avoid the system to start an infinite loop on a query it can not answer, a finite first-order dialogue logic has been devised which permits to envision the computation of world model backbones (assignments that pertain to every model). To assess our approach on model inference mechanisms, the propositional case seems to be the relevant testbed. Indeed, in this context, comparison to state-of-the-art solvers for which performance is an highly competitive issue should provide valuable information before lifting our strategy to the first-order case.

Roughly speaking, this strategy postpones space search while the input formula can be contracted on inner conflict detections. The widely spread Conjonctive Normal Form (CNF) format is then inappropriate as it looses the instance inner structure. Another drawback of this format is that it makes implicitly solvers sensitive to encoding strategies for natural structured problems. A rapid search for a format with no encoding requirements and publicly available benchmarks provided us with the ISCAS format in the context of circuit verification. This format was originally thought of one among others on which we would assess our propositional solver but a recent publication [34] attested that this format was in fact the only one with public benchmarks that keeps track of structural information. As CNF equivalent benchmarks are also provided for a direct comparison with state-of-the-art SAT solvers, we decided to focus on it.

Consequently, we stopped the development of our general solver to consider in which way the ISCAS inner variables corresponding to sub-formula factorizations could speed up SAT solving. Our conclusion is that these inner variable definitions amplify boolean propagation and may consequently provide additional formula contractions. Hypothesizing performance improvement, we decided to add an additional polynomial phase before exponential space search, the solver development has since then resumed.

6.1.2. Dialogue systems evaluation

Participants: Jacques Siroux, Johann L'Hour.

This activity was developed in the framework of the AUPEL-AUF agency and with a collaboration with several laboratories. The design of the new GEORAL system will allow to work on this topic by recording real corpora and by testing different methods and algorithms. This year we designed and implemented a software for recovery and administration of traces in Georal tactile system. The software allows an administrator to choose the modules of Georal tactile system to be "spied" and to store in a database the exchanged messages using different formats (surface form, deep structure, XML format,...). These formats have been chosen in order to exploit both user and Georal activities at different levels and purposes: speech recognition (training), speech understanding and interpretation (linguistic processing), dialogue management, ergonomic studies, system evaluation, ... The software, developped using JAVA language, has to run within the JADE framework and has to deal with some hints of a multimodal dialogue system: how to determine a speech turn (taking into account overlapps in speech), how to deal with parallel activities (touch screen and speech-based activities).

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, Johann L'Hour.

Recent progresses in speech recognition allow to plan new important developments inside the dialogue system GEORAL TACTILE [57]. 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 [31].

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 [60][28] which take into consideration the spatial characteristics of the handled elements. Next we 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 [23] 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 labs. 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.

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 been proposed this year 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.

As a first phase, a study is conducted to test the efficiency of the learning of subsequential transducers realizing the transformation of the outputs of a speech recognizer into a sequence of dialogue acts.

The work undertaken so far on this subject lead to the human labeling of dialogue corpora containing approximately 3000 sentences. A technological review on the inference of transducers has been carried during this year.

6.3.2. Lazy learning of tree structures

Participant: Laurent Miclet.

This research topic has no new results in 2004, since its main contributor has completed his thesis had its defense on December 2002. The topic is described at section 3.4.2. We have maintained the discussion with France Telecom Recherche et Developpement and set up a new collaboration on this topic. This could be finalized through another CRC (see section 7.2). Secondly, we have started more fondamental studies on learning by analogy on sequences and trees, to generalize the "nearest-neighbour" technique used in the previous approach. This is described in the next section.

6.3.3. Learning by analogy on sequences

Participants: Sabri Bayoudh, Arnaud Delhay, Laurent Miclet.

The thesis of L. Blin has studied how to learn the prosody of a sentence by using a distance between trees (the sentences are represented by trees) and the nearest neighbour technique. It has been concluded at the end of year 2002, and given last results in 2003 [27].

We examine how the nearest neighbor method could be extended to learning by analogy. Our first aim is to define what is analogy on sequences, then to define learning by analogy on sequences. Future work will extend the study on other structures.

An analogy is described as follows : *find x from a triple a, b and c such that x is to c as b is to a* and is often written by

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

Our approach is based on the use of the edit distance between sequences to define the relation "is to". This approach does not use the inclusion property, like Lepage and Yvon do, and enables substituions in sequences.

We have mostly worked on defining what is solving an analogical equation on sequences when the edit distance is introduced, to generalize the works by Lepage [40] and Yvon [63] (for whom the edit distance is a trivial case).

We have produced an algorithm which is consistent with those of Lepage and Yvon, and shown that it can give an unique solution to any analogical equation on sequences. A sufficient condition is that the alphabet on which the sequences are written has itself a inner distance relation and that any analogical equation can be solved on its letters. This is for example the case if this alphabet is a cyclic group. This algorithm is generalized with the use of transducers in our recent publications [14][13][12].

We have proposed an other way to consider solving analogies on letters by defining a letter with a set of features[12][18]. Solving analogies on these sets is straightforward and the basic technique has been explained by Lepage in [40]. Considering this approach, we can now consider an alphabet either as a cyclic group with a constrained distance or as a set of elements that are defined by binary features. It is possible to define a distance between these sets of features; one of the simplest is the Hamming distance.

Current studies are focused on learning by analogy on sequences and a PhD student, Sabri Bayoudh, has begun his thesis on the 1st of October 2004. A master student will help in improving the solving techniques that we have proposed.

6.3.4. Learning speech units for speech synthesis

Participant: Olivier Boëffard.

This study is covered within the framework of a PhD thesis funded by a research contract with FTR&D Lannion (FTR&D/DIH/ISP). Work began on October 1, 1999 and was completed by the defence of Helene François' PhD thesis in December 2002. Since December 2002, we are involved in national research project named NÉOLOGOS, 2003 TECHNOLANGUE call for proposal 7.1. Our essential participation consists in applying the methodologies developed during the Hélène François' thesis work to define speech corpora usable at the same time for speech recognition and speech synthesis. Currently, the content of the corpora is defined, their collections take place during 2004.

The NEOLOGOS project is a speech databases creation project for the French language subsidized by the French ministry for research in the framework of the Technolangues program. Academic laboratories (LORIA and IRISA) and industrial companies (France Telecom, ELDA and TELISMA, coordinator of the project) are collaborating in the field of speech recognition for the creation of two new kinds of speech databases : a SpeechDat-like speech database for children's voices (PAIDIALOGOS sub-project); a speech database with a novel kind of structure for adult voices (IDIOLOGOS sub-project). In both subprojects the goal is to bring to the research community new sources of telephone speech data likely to improve ASR performance: on the one hand, to significantly improve speech recognition for children (with PAIDIALOGOS), on the other hand to provide speech data to support the development of advanced ASR techniques such as eigenvoices (with IDIOLOGOS). IDIOLOGOS should also provide the means of advanced studies on speakers characteristics, with a significant panel of reference speakers, including in the area of speech synthesis and speaker identification.

During 2004, Neologos focused on the design of the linguistic databases : a bootstrap database recorded by 1,000 speakers; each speaker utterates the same set of 50 phonetically well-covered sentences and a full database recorded by 200 reference speakers chosen among the first 1,000. These databases are designed to maximise the phonological diversity of the speech material. They are technically built to allow text-to-speech synthesis. Considering the bootstrap database, one can find 1,000 small inventories, in order to synthesise 8 of the 50 sentences of each small database. We think that these 1,000 small unit dictionaries can help in finding a relation between the quality of the synthetic speech and the natural voice quality of the speaker. The 200 reference speakers database achieves the same goal, but now we can, thanks to the 500 recorded utterances for each speaker, we can build a full diphone TTS system.

Both corpora of phonetically rich sentences were constructed by processing and simplifying sentences from large publicly available newspaper corpora in French. Automatic corpora reduction methods such as the greedy algorithm reported in [38] were used to extract a subset of sentences meeting a criterion of minimal

representation of all phonemes as well as a criterion of minimum representation of diphone classes. There were 99 diphone classes constructed from 10 broad phonetic classes including the silence [17].

6.3.5. Automatic speech labeling

Participants: Samir Nefti, Olivier Boëffard.

This study is covered within the framework of a PhD thesis funded by a research contract with FTR&D Lannion (FTR&D/DIH/ISP). Work began on January 1, 2000.

This work relates to the automatic segmentation of speech corpora, read or spontaneous, into phone units. Text-to-Speech synthesis systems based on concatenated acoustic units need this process. The general framework of this study is quoted in section 3.4.4.

Considering the wide scope of this topic, we have addressed only the detection problem. We turn out attention towards methods of scoring the confidence of this acoustic to phonologic mapping. We have conducted experimental studies to validate our choices. Compared to state-of-the-art scientific background, the original confidence measure we proposed within the HMM framework yields experimentally the best scores evaluated through DET curves – 12% Equal Error Rate, EER, for a randomly blurred test database. The next step concerns the treatment of rejected models given the speech. Within an area delimited by the confidence measure we propose to substitute the wrong sequence of phones by a local language model built on sequences of phonemes. Experiments were conducted and we concluded that this strategy can improve the performance of the speech alignment process [11], [10].

6.3.6. Learning prosody from speech

Participants: Salma Mouline, Nelly Barbot, Olivier Boëffard.

This study is covered within the framework of a PhD thesis funded by FTR&D. Work began on January 1, 2003 and the participation of two Master students.

The approach suggested within the framework of this thesis is based on the modeling of the prosody by a set of forms representative of the various realizations of the melody present in a reference speech corpus 3.4.7. Once this representation defined, we plan to segment automatically the database using these elementary forms. Each one of these segments is annotated using syntactic, phonetic and phonological labels obtained during the linguistic analysis of the corpus. Next, we want to answer the question of mapping the tagged prosodic elementary forms and the associated linguistic characteristics. Taking into account the correspondence between linguistic and prosodic parameters should make it possible to restore the style of elocution actually recorded by a speaker. Moreover, at the synthesis stage, during the selection process of the acoustic units, the prosodic targets resulting from the proposed system should better correspond to the true prosodic parameters of the recorded speaker. During 2004, we studied the MoMel parametric model and propose, within an automatic learning framework, a solution to adapt the parameters of this model to new speech corpora (different voices), [16].

In relation to this scientific topic 3.4.7, we welcome a student during the 2003/04 academic year. We focus on modeling the F_0 evolution and propose a parametric representation based on B-splines 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, that are the degree of the B-splines, 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 sucessive knots, these B-spline functions are non-negative polynomial functions of degree m and its degrees of continuity at a knot depend of 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. The main difficulty is to discover an optimal number of knots. Experiments show that, for too few knots, the error is too high, and for too many, the model complexity is overestimated. A first means is an experimental stopping when this error is less than a given threshold, and we plan to consider this issue in a better theoretic

way applying a true bayesian framework or a simpler solution like MDL (Minimum Description Length) principle.

In relation to this scientific topic 3.4.7, we welcome a Master student during the 2003/04 academic year. Another representation of the fondamental frequency have been considered during this training period. This new approach represents the F_0 contours with a state-space dynamical system model. More precisely, the sequence of F_0 observations is represented in terms of a sequence of unobserved states and the relation between the two processes are non-linear. The main issue is the automatical learning of parameters of this extended Kalman filter model. This work uses an iterative method based on the two-step EM algorithm (Estimation Maximization) for the maximum log likelihood estimation of the joint processes. This classic algorithm from systems engineering uses local linearization, and so approximation in every step, which can introduce unstability of the learning process. We plan to improve stability and accuracy using other approximate inference techniques such as the unscented filter or particular filter.

6.3.7. Learning semantics from speech

Participants: Pierre Alain, Nelly Barbot, 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) 7.2. Work began on October 1, 2003.

We adopt a methodology based on the observation of corpus of examples within a statistical theoretical framework. However, it is difficult to find corpora annotated by semantic elements (particularly in French). We propose to handle three kind of semantic sources :

- Eurowordnet, an ontology which comprises a hierarchical network of concept nodes, populated with words. The nodes in WordNet networks are termed synsets, as they contain sets of synonymous words representing a common underlying concept. Synsets offer a means of semantic generalization, both over the component words within a given synset and between synsets via hierarchical relations such as hyponymy and hypernymy.
- Dictionnary of the Language. With a classical dictionary, each word is explained by one definition or more if the word is polysemic. We propose to exploit the sentence given as an example to illustrate the correct use of a concept.
- The web. Recently, several publications refer to a methodological framework which consists in web searching to find semantic correlates and use it as learning sentences.

During 2004, we focus on the representation of the lingustic data. We adopt an XML frawework due to the great activity of this technological field. Our next step will focus on the use of eurowordnet to tag semantically text corpora with hierarchical concepts. The corollary of this work is to propose a solution to automatically enrich eurownet with new related synsets.

6.3.8. Learning to improve the dialogue management

As indicated in section 7.2, some collaborative work has started this year 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.

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.

The 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. Dialogue and Semantics

In 2003 has been finalized the CRC[5] "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 3.4.6, another thesis at FTRD DIH/DII, started Feb 2002 and the thesis of E. Livolant started in January 2004. 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).

Cordial is also part of a "pre-projet" in the interdisciplinary program TCAN of the CNRS, called ANALANGUE (analogies in sequences).

9. Dissemination

9.1. Leadership within scientific community

Olivier Boëffard has been reviewer for the IEEE transactions on Speech and Audio Processing, the Signal Processing journal, and the Speech Communication journal.

Laurent Miclet has been a member of the scientific committee of the French Machine Learning Congress, Conférence d'Apprentissage *CAP 2004* and International Colloquium on Grammatical Inference (ICGI'04).

Jacques Siroux has been reviewer for the journal *Intellectica* and for the CARI2004, IJCNLP2004, LREC2004 conferences.

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

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*).

9.3. Conferences, workshops and meetings, invitations

Laurent Miclet has been a member of the jury for the thesis: A. Habrard. *Modèles et techniques en inférence grammaticale probabiliste : de la gestion du bruit à l'extraction de connaissances*. Thèse de l'Université Jean Monnet, St Etienne, Octobre 2004.

Laurent Miclet will be a member of the jury for the HDR: O. Boëffard. *Contributions à la synthèse de la parole*. HdR Université de Rennes 1, Décembre 2004.

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] P. DUPONT, L. MICLET, E. VIDAL. *What is the search space of the regular inference ?*, in "Grammatical Inference and Applications, Lecture notes in AI 862", Springer Verlag, September 1994.

[4] H. FRANÇOIS, O. BOËFFARD. *Evaluation if units selection criteria in corpus-based speech synthesis*, in "proceedings of the Eurospeech Conference, Geneva, Switzerland", 2003, p. 1325–1328.

[5] M. GUYOMARD, P. NERZIC, J. SIROUX. *Plans, métaplans et dialogue*, Technical report, n° 1169, Irisa, September 1998.

[6] L. MICLET, A. DELHAY. *Analogy on Sequences: a Definition and an Algorithm*, Technical report, n° 4969, INRIA, October 2003, http://www.inria.fr/rrrt/rr-4969.html.

[7] S. NEFTI, O. BOËFFARD, T. MOUDENC. *Confidence measure for phonetic segmmentation of continuous speech*, in "proceedings of the Eurospeech Conference, Geneva, Switzerland", 2003, p. 897–900.

[8] 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.

[9] F. VIOLARO, O. BOËFFARD. *A Hybrid Model for Text-to-Speech Synthesis*, in "IEEE transactions on Speech and Audio Processing", vol. 6, n° 5, 1998, p. 426–434.

Doctoral dissertations and Habilitation theses

[10] O. BOËFFARD. *Contributions à la synthèse de la parole*, Ph. D. Thesis, Habilitation à diriger les recherches, IRISA – Université de Rennes 1, December 2004.

[11] S. NEFTI. *Segmentation automatique de parole en phones*, Ph. D. Thesis, IRISA – Université de Rennes 1, December 2004.

Articles in referred journals and book chapters

[12] A. DELHAY, L. MICLET. *Analogie entre séquences : définition, calcul et utilisation en apprentissage supervisé*, in "Revue d'Intelligence Artificielle", à paraître, 2005.

Publications in Conferences and Workshops

[13] A. DELHAY, L. MICLET. *Analogical Equations in Sequences: Definition and Resolution*, in "Proceedings of the 7th International Colloquium on Grammatical Inference", October 2004, p. 127–138.

[14] A. DELHAY, L. MICLET. *Solving analogical equations for learning by analogy with sequences*, in "Actes de la Conférence sur l'Apprentissage, CAp 2004", June 2004.

[15] J. L'HOUR, O. BOËFFARD, J. SIROUX, L. MICLET, F. CHARPENTIER, T. MOUDENC. *DORIS, a multia-gent/IP platform for multimodal dialogue applications*, in "proceedings of the International Conference on Spoken Language Processing (ICSLP'04)", 2004.

[16] S. MOULINE, O. 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'04)", 2004.

[17] E. PINTO, D. CHARLET, H. FRANÇOIS, D. MOSTEFA, O. BOËFFARD, D. FOHR, O. MELLA, F. BIMBOT, K. CHOUKRI, Y. PHILIP, F. CHARPENTIER. *Development of new telephone speech databases for French : the NEOLOGOS Project*, in "proceedings of the International Conference on Language Resources and Evaluation (LREC'04)", 2004.

Internal Reports

[18] L. MICLET, A. DELHAY. *Relation d'analogie et distance sur un alphabet défini par des traits*, Technical report, n° 5244, INRIA, July 2004, http://www.inria.fr/rrrt/rr-5244.html.

Bibliography in notes

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

[20] J. ALLEN. *Natural Language Understanding*, Benjamin/Cummings Menlo Park, 1987.

[21] C. ANDRIEU, M. DAVY, A. DOUCET. *Efficient particle filtering for jump Markov systems. Application to time-varying autoregressions*, in "IEEE transactions on Signal Processing", vol. 51, n° 7, 1770 2003, 1762.

[22] K. ARAI, J. WRIGHT, G. RICCARDI, A. GORIN. *Grammar fragment acquisition using syntactic and semantic clustering*, in "Speech Communication", vol. 27, n° 1, 1999, p. 43–62.

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

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

[25] D. BIGORGNE, O. BOËFFARD, B. CHERBONNEL, F. EMERARD, D. LARREUR, J. L. L. 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.

[26] 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, December 2002.

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

[28] 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.

[29] O. BOËFFARD. *Variable-length acoustic units inference for text-to-speech synthesis*, in "proceedings of the Eurospeech Conference", 2001.

[30] O. BOËFFARD, F. EMERARD. *Application-dependent prosodic models for text-to-speech synthesis and automatic design of learning database corpus using genetic algorithm*, in "Eurospeech'97, Rhodes, Greece", vol. 5, September 1997, p. 2507–2510.

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

[32] 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.

[33] E. CAMPIONE, D. HIRST, J. VÉRONIS. *Intonation: Models and Theories*, chap. Stylisation and symbolic coding of F0 : comparison of five models, Kluwer Academic Publishers, 2000, p. 185–208.

[34] F. B. CHRISTIAN THIFFAULT, T. WALSH. *Solving Non-clausal Formulas with DPLL search*, in "Seventh International Conference on Theory and Applications of Satisfiability Testing (SAT 2004), Vancouver, Canada", May 2004, p. 147-156.

[35] A. COZANNET, J. SIROUX. *Strategies for Oral Dialogue Control*, in "Proceedings of International Conference on Spoken Language Processing (ICSLP) 94, Yokohama, Japon", vol. 2, 1994, p. 963–966.

[36] E. DRENTH, B. RUBER. *Context-dependent probability adaptation in speech understanding*, in "Computer Speech and Language", vol. 11, n° 3, 1997, p. 225–252.

[37] P. DUPONT, L. MICLET. *L'inférence grammaticale régulière : fondements théoriques et principaux algorithmes*, Technical report, n° 3449, INRIA, July 1998, http://www.inria.fr/rrrt/rr-3449.html.

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

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

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

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

[42] A. LOGOTHETIS, V. KRISHNAMURTHY. *Expectation maximization algorithms for MAP estimation of jump Markov linear systems*, in "IEEE transactions on Signal Processing", vol. 47, n° 8, 1999, p. 2139–2156.

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

[44] M. MCTEAR. *Spoken Dialogue Technology : Enabling the Conversational User Interface*, in "ACM Computing surveys", vol. 34, n° 1, 2002, p. 90–169.

[45] P. MERTENS. *Synthesizing elaborate intonation contours in text-to-speech for french*, in "Proccedings of the Speech Prosody Conference", 2002.

[46] 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° 1, 1992, p. 9-17.

[47] W. MINKER, A. WAIBEL, J. MARIANI. *Stochactically-based semantic analysis*, Kluwer Academic Publishers, 1999.

[48] S. NEFTI, O. BOËFFARD. *Acoustical and topological experiments for an HMM-based speech segmentation system*, in "proceedings of the Eurospeech Conference, Aalbord, Denmark", 2001.

[49] 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.

[50] K. NIGAM. *Using unlabeled data to improve text classification*, Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA 15213, 2001.

[51] D. O'SHAUGHNESSY. *Interacting with computers by voice: automatic speech recognition and synthesis*, in "Proceedings of the IEEE", vol. 91, n° 9, 2003, p. 1272–1305.

[52] 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.

[53] L. RABINER. *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*, in "proceedings of the IEEE", vol. 77, n° 2, 1989, p. 257–286.

[54] S. RAY, M. CRAVEN. *Representing Sentence Structure in Hidden Markov Models for Information Extraction*, in "proceedings of the IJCAI conference", vol. 2, 2001, p. 1273–1279.

[55] J. P. H. V. SANTEN. *Combinatorial issues in TTS synthesis*, in "proceedings of the Eurospeech Conference", 1997.

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

[57] J. SIROUX, 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", July 1997, p. 39–44.

[58] SUNDIAL. *SUNDIAL, Prototype performance evaluation report*, Deliverable, n° D3WP8, projet Sundial P2218, September 1993.

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

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

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

[62] 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.

[63] F. YVON. *Analogy-based NLP : Implementation Issues.*, Technical report, Ecole Nationale Supérieure des Télécommunications, 2003.