



CORDIAL

***Man-machine oral and multimodal
communication***

Lannion

Activity Report

2010

1 Team

Head of project (Responsable scientifique)

Laurent Miclet [Profession=Enseignant] [Category=UnivFr] [Team Leader, Faculty member (Professor), Enssat] [HDR=Habilitation]

Administrative staff

Joëlle Thépault [Profession=Assistant] [Category=UnivFr] [Administrative assistant, Enssat, 20%]

Nelly Vaucelle [Profession=Assistant] [Category=UnivFr] [Administrative assistant, Enssat, 10%]

Faculty members (University of Rennes 1)

Nelly Barbot [Profession=Enseignant] [Category=UnivFr] [Associate Professor, Enssat]

Vincent Barraud [Profession=Enseignant] [Category=UnivFr] [Associate Professor, Enssat]

Olivier Boeffard [Profession=Enseignant] [Category=UnivFr] [Professor, Enssat] [HDR=Habilitation]

Arnaud Delhay [Profession=Enseignant] [Category=UnivFr] [Associate Professor, IUT]

Marc Guyomard [Profession=Enseignant] [Category=UnivFr] [Professor, Enssat] [HDR=Habilitation]

Damien Lolive [Profession=Enseignant] [Category=UnivFr] [Associate Professor, Enssat]

Arnaud Martin [Profession=Enseignant] [Category=UnivFr] [Professor, IUT] [HDR=Habilitation], since September 2010

Ph. D. Students

Anouar Ben Hassena [Profession=PhD] [Category=UnivFr] [Associate Professor, ATER, IUT, from October 2010 to August 2011]

Larbi Mesbahi [Profession=PhD] [Category=UnivFr] [Associate Professor, ATER, IUT from October 2009 to August 2010, PhD defended on October the 28th, 2010]

Sébastien Le Maguer [Profession=PhD] [Category=UnivFr] [bourse Conseil Général des Côtes d'Armor, from November the 1st, 2008]

Sylvie Saget [Profession=PhD] [Category=UnivFr] [from October the 15th, 2003]

Technical staff

Laurent Blin [Profession=Technique] [Category=UnivFr] [Research engineer, CPER MOB-ITS]

Laure Charonnat [Profession=Technique] [Category=UnivFr] [Research engineer, FNADT, until april 2010, Research grant from Orange Labs from may 2010]

Gaëlle Vidal [Profession=Technique] [Category=UnivFr] [Assistant engineer, FNADT, until may the 20th, 2010]

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 interpersonal 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 [dM98], 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 explicitly 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

[dM98] R. DE MORI (editor), *Spoken Dialogue with Computers*, Academic Press, 1998, ISBN 0122090551.

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 Processing

The research activities in speech processing accomplished by the Cordial project are embedded in the general scientific framework of *automatic speech transformation*. This framework, in particular, includes unit selection for speech synthesis, voice transformation, speech segmentation, etc.

What makes a voice specific, that we can recognize a familiar voice at phone for example, is a relatively complex concept to encircle and define. The main first concept in defining voice quality is certainly the perceived timbre of speech but it masks other suprasegmental factors (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 is divided along three different technological axis: text-to-speech synthesis (TTS), biometry, and the study of pathological voices.

3 Scientific Foundations

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 processing*.

3.1 Dialogue and modeling

Keywords: Speech Acts, planning, plan recognition.

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^[Aus70, Sea82] called in our context *dialogue acts* which support both the function of the act in the

[Aus70] J. AUSTIN, *Quand dire c'est faire*, Editions du seuil, Paris, France, 1970.

[Sea82] J. SEARLE, *Sens et expression*, Les éditions de minuit, 1982.

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^[All87,Lit85] 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 beliefs 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.

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 contains a lot of irregularities (problems of performance) which complicates the speech recognition task as well as the understanding and interpretation tasks.

System modeling

The second problem takes place in the use of the planning formalism [8] in order to associate three points of view^[NGS92]: 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

-
- [All87] J. ALLEN, *Natural language understanding*, Benjamin/Cummings Menlo Park, 1987.
[Lit85] D. J. LITMAN, *Plan Recognition and Discourse Analysis : An Integrated Approach for Understanding Dialogues*, PdD Thesis, University of Rochester, TR 170, 1985.
[NGS92] P. NERZIC, M. GUYOMARD, J. SIROUX, "Reprise des échecs et erreurs dans le dialogue homme-machine", *Cahiers de linguistique sociale* 21, 1992, p. 35–46.

have been found^[Lit85] 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.

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.

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.

3.1.1 Modeling the Communicative Common Ground (CCG)

This section is about modeling the Communicative Common Ground (CCG) in term of Collective Acceptance. This study has been 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 ([KDB⁺04]).

Garrod and Pickering [GP04] claim that considering spoken dialogue as a collaborative activity must lead to avoid or to modify the fundamental hypotheses which are responsible for complexity limitation in existing spoken dialogue systems. Generally speaking, a spoken dialogue 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. Moreover, in collaborative models of dialogues, utterance treatment (generation and interpretation) is notably based on the (Subjective) Common Ground (ie. mutual beliefs), among dialogue partners. Accommodation, then provides a way of ensuring the coherence of their epistemic state while solving coordination problems. Respecting this fundamental hypothesis constrains spoken dialogue systems to support rich epistemic states (containing mutual beliefs and nested beliefs) and the associated reasoning processes. On the whole, these considerations lead to practical limitations [TCM96] and theoretical incoherences.

The methodological approach follows in our work is:

1. To isolate a theoretical problem responsible of practical limitations.

The thesis claims that one of these fundamental hypothesis, considered by Garrod and Pickering, is the sincerity hypothesis, ie. considering dialog as a truth-oriented process. The sincerity hypothesis is partly responsible for the complexity of reaching mutual understanding and for the difficulty of modelling the corresponding collective decision process.

2. To propose an alternative view of dialog as a goal-oriented process where belief has an indirect role.

That is sincerity is a *possible but not necessary strategy* to reach mutual understanding.

3. To review existing literature on collaborative activity for expressing the preceding point.

The philosopher J. L. Cohen [Coh92] shows the properness of distinguishing between the pragmatic mental attitude which is acceptance from the context-free mental attitude which is belief. This allows several subjects to be handled such as, in pragmatics, cases where Moore's paradox ("It's raining and he believes that it is not raining") occur. This distinction is then suitable to distinguish goal-oriented process from truth-oriented process.

[KDB⁺04] G. KLEIN, W. D.D., J. BRADSHAW, R. HOFFMAN, P. FELTOVICH, "Ten Challenges for Making Automation a 'Team Player' in Joint Human-Agent Activity", *IEEE Intelligent Systems*, November/December 2004, p. 91–95.

[GP04] S. GARROD, M. J. PICKERING, "Why is conversation so easy?", *Trends in Cognitive Sciences* 8, 2004, p. 8–11.

[TCM96] J. TAYLOR, J. CARLETTA, C. MELLISH, "Requirements for belief models in cooperative dialogue", *User Modeling and User-Adapted Interaction* 6, 1, 1996, p. 23–68, <http://citeseer.ist.psu.edu/232252.html>.

[Coh92] J. L. COHEN, *An Essay on Belief and Acceptance*, Oxford University Press, Oxford, 1992.

4. To develop a collaborative model of dialog based on the distinction between belief and acceptance.

In this model, the notion of acceptance is used to capture the fact that an utterance is viewed as a tool allowing the speaker's communicative intention to be established. The process of understanding negotiation is then considered as the co-construction of this linguistic tool allowing dialogue partners to reach a point of mutual understanding which is sufficient for their current activities. The result of this co-construction is formalized by a collective acceptance.

5. To study practical consequences of this theoretical model for spoken dialog systems.

During 2006, the philosophical fundamentals of this approach have been developed [14] as well as a first version of the collaborative model of dialog [13].

During 2007, the collaborative model of dialog has been refined [SG07]. The model is notably based on F. Paglieri's distinction between beliefs and acceptance [Pag06]:

Belief	<i>vs</i>	Acceptance
$B_i(\phi)$	<i>vs</i>	$Acc_i(\phi)$
truth-oriented	<i>vs</i>	goal-oriented
ϕ is true		ϕ is suitable for the success of a certain goal
↓		↓
involuntary	<i>vs</i>	voluntary
gradual	<i>vs</i>	all-or-nothing
context-free	<i>vs</i>	context-dependant

We have also proposed first elements of a first formal model for acceptance as an individual attitude. Some of the practical consequences have been explored [Sag07].

3.1.2 Hability modeling

In the Russel and Norvig's book [RN03], an agent is defined by three main features:

- An agent perceives his environment, adapting himself and acting in consequence.

-
- [SG07] S. SAGET, M. GUYOMARD, "Doit-on dire la vérité pour se comprendre ? Principes d'un modèle collaboratif de dialogue basé sur la notion d'acceptation", *in: Modèles formels de l'interaction, MFI'07, Actes des quatrièmees journées francophones, Publiés dans les Annales du Lamsade, 8*, p. 239-248, mai 2007.
- [Pag06] F. PAGLIERI, *Belief dynamics: From formal models to cognitive architectures, and back again*, PdD Thesis, University of Siena, 2006.
- [Sag07] S. SAGET, "Using Collective Acceptance for modelling the Conversational Common Ground: Consequences on referent representation and on reference treatment", *in: 5th IJCAI's Workshop on Knowledge and Reasoning in Practical Dialog Systems*, p. 55-58, Hyderabad, India, 2007.
- [RN03] S. RUSSELL, P. NORVIG, *Artificial Intelligence: A Modern Approach*, edition 2nd edition, Prentice-Hall, Englewood Cliffs, NJ, 2003.

- 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 in [Dev07b, Dev07a], characterized 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 concretely our approach on an agent framework, called JADE (Java Agent Development Framework).

3.2 System and multimodality

Keywords:

multimodality, reference, educational software, 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.

[Dev07b] K. DEVOOGHT, "A Semantics for Changing Frames of Mind", *in*: *CONTEXT*, p. 192–205, 2007.
 [Dev07a] K. DEVOOGHT, "Modélisation de la capacité en fonction des activités d'un agent intentionnel", *in*: *conférence AFIA, Journée thématique Intelligence Artificielle Fondamentale*, 2007.

3.2.1 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 ^[May90], performed in a different context (the generation of communicative acts for the system output). 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^[Van86] 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)^[Vie91] 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^[Sun93,CS94] (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

-
- [May90] M. MAYBURY, "Communicative Acts for Explanation Generation", *International Journal of Man-machine studies* 37(2), 1990, p. 135-172.
- [Van86] C. VANDELOISE, *L'espace en français*, Éditions du seuil, Paris, 1986.
- [Vie91] L. VIEU, *Sémantique des relations spatiales et inférences spatio-temporelles : une contribution à l'étude des structures formelles de l'espace en langage naturel*, PdD Thesis, Université Paul Sabatier, Toulouse, 1991.
- [Sun93] SUNDIAL, "SUNDIAL, Prototype performance evaluation report", *Deliverable number D3WP8*, projet Sundial P2218, September 1993.
- [CS94] A. COZANNET, J. SIROUX, "Strategies for oral dialogue control", *in : Proceedings of International Conference on Spoken Language Processing (ICSLP)*, 2, p. 963-966, Yokohama, Japan, 1994.

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.

Recent progresses in speech recognition allow to plan new important developments inside the dialogue system GEORAL TACTILE^[SGMR97]. 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^[Bre98].

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^[Van86,Bor88] 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^[Aur93] 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

-
- [SGMR97] J. SIROUX, M. GUYOMARD, F. MULTON, C. RÉMONDEAU, "Multimodal References in Georal Tactile", *in: Proceedings of the workshop Referring Phenomena in a multimedia Context and their Computational Treatment, SIGMEDIA and ACL/EACL*, p. 39-44, Madrid, 1997.
- [Bre98] G. BRETON, "Modélisation d'un contexte cartographique et dialogique", *research report*, DEA Informatique de Rennes 1, 1998, ENSSAT.
- [Van86] C. VANDELOISE, *L'espace en français*, Éditions du seuil, Paris, 1986.
- [Bor88] A. BORILLO, "Le lexique de l'espace : les noms et les adjectifs de localisation interne", *Cahiers de grammaire 13*, 1988, p. 1-22.
- [Aur93] M. AURNAGUE, *A unified processing of orientation for internal and external localization*, Groupe Langue, Raisonnement, Calcul, Toulouse, France, 1993.

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 ^[VP00,Mit02] 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 (2006) 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) ^[CS07a]. 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, 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 salience of the entities on the map. This visual salience 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

-
- [VP00] R. VIEIRA, M. POESIO, "An empirically based system for processing definite descriptions", *Computational Linguistics* 26, 4, 2000, p. 539-545.
- [Mit02] R. MITKOV, *Anaphora resolution*, Longman, 2002.
- [Man03] H. MANUÉLIAN, *Description définies et démonstratives : Analyses de corpus pour la génération de textes*, PdD Thesis, LORIA - Université Nancy 2, 2003.
- [CS07a] A. CHOUMANE, J. SIROUX, "About several Reference Processings in Multimodal Human-Computer Communication", *Publication Interne number 1845*, IRISA, 2007.

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.

In 2007, we proposed a model for the visual context. This model is composed of three levels: the first one is the displayed map on the screen. The entities (towns, forests, rivers, ...) of the application are represented using dots, areas, lines, polylines. The second level is concerned with the internal representation of the entities of the first level; a vector of characteristics is assigned to each entity. The third level is a logical representation of the spatial relationships between entities. We also developed the linguistic model by adding a semantic network (from eurowordnet) and mapping functions between natural language and entities of the application. These developments provide users with flexibility. All these propositions have been presented in [CS07c]; they have been implemented using XML language.

We also developed methods and algorithms in order to deal with ER resolution and gestures. The first method is to deal with ER uttered without gesture but based on the visual context (example : the utterance «je veux les campings *le long de la rivière* » uttered without gesture). The resolution of ER is based on the visual saliance of the displayed entities. The algorithm is published in [Cho07]. We developed a second algorithm in order to deal with ERs produced both by speech and gesture. The algorithm uses two strategies: the first one is based on probabilities in order to find the entity designated by the gesture and the second one is based on the visual saliance. The algorithm is described in [CS07b]. First experiments in order to deal with complex and ambiguous gestures produced interesting results.

3.2.2 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.

3.3 Machine learning in dialogue systems

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

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

-
- [CS07c] A. CHOUMANE, J. SIROUX, "A Model for Multimodal Representation and Processing for Reference Resolution", *in: Workshop on Multimodal Interfaces in Semantic Interaction, ICMI 2007*, ACM, p. 39–42, Nagoya, Japan, 2007.
- [Cho07] A. CHOUMANE, "Traitement de désignations orales dans un contexte visuel", *in: 14ème conférence sur le Traitement Automatique des Langues (Résumé)*, p. 479–488, Toulouse, 2007.
- [CS07b] A. CHOUMANE, J. SIROUX, "Interpretation of Multimodal Designation with Imprecise Gesture", *in: 3rd International Conference on Intelligent Environments (IE)*, p. 232–238, Ulm - Germany, 2007.

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 on 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.

3.3.1 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^[dlH10].

[dlH10] C. DE LA HIGUERA, *Grammatical Inference*, Cambridge University Press, Cambridge, 2010.

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 in [VC04,OGV93]). We are interested in experimenting these techniques in the framework of the transformation of the outputs of a speech recognizer into a sequence of dialogue acts. In particular, we will consider the introduction of domain knowledge in the learning algorithm.

3.3.2 Analogical learning of sequences and 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, syntactic, 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^* which is the closest to x and adapts to x a prosody p_x directly deduced from p^* .

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 past years [Bli03,Bli02].

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

-
- [VC04] E. VIDAL, F. CASUBERTA, "Learning Finite-State Models for Machine Translation", in: *Proceedings of the 7th International Colloquium on Grammatical Inference*, 2004.
 - [OGV93] J. ONCINA, P. GARCÍA, E. VIDAL, "Learning subsequential transducers for pattern recognition and interpretation tasks", *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15, 1993, p. 448–458.
 - [Bli03] L. BLIN, "Génération de prosodie par apprentissage de structures arborescentes", in: *Actes de la Conférence Francophone sur l'Apprentissage Automatique (CAP)*, Laval, France, 2003.
 - [Bli02] L. BLIN, *Apprentissage de structures d'arbres à partir d'exemples : application à la prosodie pour la synthèse de la parole*, PdD Thesis, IRISA – Université de Rennes 1, 2002.

synthesize, look for a triple of trees (b, c, d) in \mathcal{S} such that x is to b as c is to d . Then compute the prosody p_x of x by solving the analogical proportion equation ' p_x is to p_b as p_c is to p_d ', which is usually written ' $p_x : p_b :: p_c : p_d$ '.

Defining an analogical proportion between sequences and trees

The classical definition of $a : b :: c : d$ as an analogical proportion requires the satisfaction of two axioms, expressed as equivalences of this primitive equation with two others equations^[LA96]:

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 $a : b :: c : d$.

Defining an analogical proportion 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 ^[Lep03] and the recent work of Yvon and Stroppa^[Str05, Yvo06].

We have defined the analogical proportion between sequences in generalizing the optimal alignment between two sentences ^[WF74] to an optimal alignment between four sentences, assuming that some analogical proportion (possibly trivial) is defined on the alphabet.

We have also defined the notion of *analogical dissimilarity*, AD , which expresses of how much four sentences "miss" the analogical proportion. $AD(a, b, c, d) = 0$ is equivalent to $a : b :: c : d$.

We have devised two algorithms ([11]) that make use of this notion. The first one, given four sentences, computes the AD between the four sentences. The second one, given three sentences, computes a sentence at minimal AD with the three sequences.

In 2008, 2009 and 2010, we have extended this research to trees. We have now a definition of the analogical proportion and the AD between four ordered trees, and two similar algorithms in the world of trees.

3.3.3 Miscellaneous works on analogy

Analogy in finite groups A research work has started on the study of analogical proportion between four elements of a finite group. This is done with respect to the axioms of analogy given by Lepage ^[Lep03] and using also the notion of factorization, the foundation of analogical

-
- [LA96] Y. LEPAGE, S.-I. ANDO, "Saussurian analogy: a theoretical account and its application", in: *Proceedings of COLING-96*, p. 717-722, København, 1996, <http://www.slt.atr.co.jp/~lepage/ps/coling96.ps.gz>.
- [Lep03] Y. LEPAGE, *De l'analogie rendant compte de la commutation en linguistique*, Université Joseph Fourier, Grenoble, 2003, Habilitation à diriger les recherches.
- [Str05] N. STROPPIA, *Définitions et caractérisations de modèles à base d'analogies pour l'apprentissage automatique des langues naturelles*, PdD Thesis, Ecole Nationale Supérieure des Télécommunications, 2005.
- [Yvo06] F. YVON, *Des apprentis pour le Traitement Automatique des Langues*, Habilitation à diriger des recherches, Université Paris 6, 2006.
- [WF74] R. A. WAGNER, M. J. FISHER, "The string-to-string correction problem", *Journal of the Association for Computing Machinery* 21(1), 1 1974, p. 168-173.

proportion according to Stroppa et Yvon [SY06]. We firstly have shown that the analogical equation has in general no solution in a non commutative group. We have explored the conditions that three elements must fulfill to define a fourth as being in analogical proportion with them. These conditions are presented according to different forms, that we show to be equivalent. Finally, we have defined the notion of analogical dissimilarity, provided that a distance is given on the group.

We have focused on two particular groups: the permutation group of a finite set and the group of invertible matrices. We have characterized for both the existence of an analogical proportion and shown how to build a set of analogical proportions when two elements are given. We have introduced the notion of analogical dissimilarity between four permutations and four invertible matrices.

The logic of analogy We have also worked, in collaboration with H. Prade (IRIT, Toulouse) on a logical modeling for providing a symbolic and qualitative representation of analogical proportion, and we have extended it to fuzzy logic in order to obtain a logical graded counterpart of numerical modelings.

This has been done firstly in exploring the existing background about postulates underlying the expected behavior of encodings of the analogical proportion, in its numerical form as well as in its set theoretic form. We have proposed a classical logic representation of analogical proportion, and finally extended it to fuzzy logics. Some investigations about potential applications in reasoning and learning respectively have been also studied. This work has produced a conference paper in 2009[MP09]. A journal article is currently under submission.

3.4 Speech Processing

Keywords: Speech processing, voice transformation, hidden Markov model, speech databases, set covering, kalman filter.

Our research activity in speech conversion is divided along three different technological axis: speech synthesis, 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? In

[SY06] N. STROPPIA, F. YVON, "Formal Models of Analogical Proportions", *Publication Interne number 2006D008*, Ecole Nationale Supérieure des Télécommunications, Paris, France., 2006.

[MP09] L. MICLET, H. PRADE, "Handling Analogical Proportions in Classical Logic and Fuzzy Logics Settings", in: *Lecture Notes in Computer Science 5590 Springer 2009. Proceedings of the 10th European Conference ECSQARU*, Springer-Verlag, p. 638–650, Verona, Italy, 2009.

other words, do the characteristics of the transformed voice corresponds to some human voice, even if a corresponding natural voice does not exist ?

The repercussions of the proposed studies are straightforward 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. The 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 attractiveness of such technologies.

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 [CfG⁺06]. 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.

3.4.1 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 ¹. This assumption leads to dynamic programming algorithms like

¹In the sense that any optimal solution carried out on a sub-sequence belongs to the optimal solution of the sequence.

[CfG⁺06] 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.

Viterbi or Dijkstra. 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 presently ill-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).

3.4.2 Optimal corpus design

In automatic speech recognition as well as speech synthesis fields, many technologies rely on models trained on large speech corpora. The quality of these models depends strongly on the linguistic content of these corpora. Therefore, 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 informations at phonologic and linguistic levels. These informations permit to characterize the sound elements that will be incorporated 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 procedure², or by a modification of some extra-linguistic factors proper to the speaker³. Consequently, it is necessary to minimize the global recording duration and so the size of the continuous speech database.

The simple strategy that consists in collecting randomly the acoustic materials turns out to be quickly expensive because of the exponential distribution of the linguistic events. Indeed, very few events take place very frequently compared with a considerable mass of rare events. This drawback becomes often acute owing to the need of many technologies to have several occurrences of a same event. Furthermore, this method does not guarantee the stability of the corpus content and its main characteristics as corpus size, sentence length, etc. This situation may influence the learning of the model parameters. One alternative consists in explicitly

²For example, the change of the microphone can introduce enough heterogeneousness, like phase problems, to alter a base.

³For example, the speaker catches a cold between two recording sessions.

controlling the content of the learning corpus according to the target application. The main difficulty is to assure the presence of units longer than a phoneme, given their heavy-tailed distribution. A solution is the automatic extraction, from huge text corpora, of a subset which covers all the descriptive attributes and minimizes the speech duration after recording.

Hence, the linguistic definition of a continuous speech database can be formulated as a Set Covering Problem (*SCP*). 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 algorithm 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 [FB01], their context [VSB97,BOD03], or a unit distribution to reach in the reduced database [KDY⁺07]. Some variants of the greedy algorithm have been also proposed, like agglomeration, spitting and pair exchange methods. In [Fra02], several combinations of these algorithms have been studied and applied to the construction of a speech synthesis corpus.

According to this study, the best compromise is an agglomerative greedy algorithm followed by a spitting greedy algorithm [5]. During the agglomerative phase, the score of a sentence corresponds to the number of its units that are missing in the ongoing covering and is divided by the sentence length. As regards the spitting phase, the longest redundant sentence is excluded of the covering. In order to clarify the rest of this paragraph, this algorithm is called *ASA, Agglomeration and then Spitting Algorithm*.

-
- [FB01] 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)*, p. 829–833, Aalborg, Denmark, 2001.
- [VSB97] J. VAN SANTEN, A. BUCHSBAUM, “Methods for optimal text selection”, *in: Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech)*, p. 553–556, Rhodes, Greece, 1997.
- [BOD03] 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)*, p. 277–280, Geneva, Switzerland, 2003.
- [KDY⁺07] A. KRUL, G. DAMNATI, F. YVON, B. C., T. MOUDENC, “Adaptive database reduction for domain specific speech synthesis”, *in: Proceedings of the 6th ISCA Tutorial and Research Workshop on Speech Synthesis (SSW6)*, W. P., A. J., H. W. (editors), p. 217–222, Bonn, Germany, 2007.
- [Fra02] H. FRANÇOIS, *Synthèse de la parole par concaténation d’unités acoustique : construction et exploitation d’une base de parole continue*, PdD Thesis, IRISA – Université de Rennes 1, 2002.

In [CBBD07], an alternative to a greedy strategy, which is sub-optimal, has been proposed by implementing of an integer programming approach. According to the combinatorial issue of the problem, this algorithm, called *LamSCP* for *Lagrangian based Algorithm for Multi-represented SCP*, uses Lagrangian-based heuristics in order to prune the search space and efficiently approach the optimal solution. *LamSCP* takes into account the constraints of multi-representation: a given minimal number of instances can be required in the covering for each attribute. It has been applied to extract multi-represented diphoneme coverings and triphoneme coverings from the large French text corpus *Le Monde* and from the English one *Gutenberg*. The results are better with *LamSCP* than the solutions found by *ASA*, offering a reduction in the cover size from 4 to 10 percents. Furthermore, *LamSCP* provides a lower bound to the *SCP* and enables to assess the real quality of the proposed solutions. Finally, the robustness of both algorithms to the perturbation of the matrix which represents the initial database has been studied in [CBBD08,ABB⁺08]: it turns out that both algorithms are robust relatively to their optimal solution sizes but *LamSCP* provides a sixfold improvement in stability as compared with *ASA*.

These works first anchor in the problematic of a continuous speech database for text-to-speech synthesis present some interest in other linguistic tasks like, for instance, speech recognition or speaker identification [9].

3.4.3 Automatic Speech Labeling and Segmentation

In speech processing, as in many other fields, automatic machine learning methodologies require databases of consequent size ⁴. 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 ⁵. An acceptable solution consists in supposing known the textual

⁴Even 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.

⁵Skills requested are those of an expert in acoustics and phonetics.

-
- [CBBD07] J. CHEVELU, N. BARBOT, O. BOËFFARD, A. DELHAY, “Lagrangian relaxation for optimal corpus design”, in: *Proceedings of the 6th ISCA Tutorial and Research Workshop on Speech Synthesis (SSW6)*, P. Wagner, J. Abresh, W. Hess (editors), p. 211–216, Bonn, Germany, 2007.
- [CBBD08] J. CHEVELU, N. BARBOT, O. BOEFFARD, A. DELHAY, “Comparing set-covering strategies for optimal corpus design”, in: *Proceedings of the 6th International Language Resources and Evaluation (LREC’08)*, E. L. R. A. (ELRA) (editor), Marrakech, Morocco, 2008.
- [ABB⁺08] P. ALAIN, N. BARBOT, O. BOEFFARD, J. CHEVELU, A. DELHAY, “Evaluation de méthodes de réduction de corpus linguistiques”, in: *Actes des XXVIIèmes Journées d’Etudes sur la Parole (JEP)*, Avignon, France, 2008.

transcription of the recorded message and to apply an automatic phonetic transcription system⁶. 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 [BFO99]. 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 [Rab89]. 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.

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 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.

⁶A human expertise is always necessary but the competence required is less accurated because it concerns only a checking of a textual transcription by listening.

[BFO99] F. BRUGNARA, D. FALAVIGNA, M. OMOLOGO, "Automatic segmentation and labeling of speech based on hidden Markov models", *Speech Communication* 12, 1999, p. 357-370.

[Rab89] L. RABINER, "A tutorial on hidden Markov models and selected applications in speech recognition", *Proceedings of the IEEE* 77, 2, 1989, p. 257-286.

3.4.4 Corpus annotation structure

Research in speech processing comes more and more linked to annotated speech corpora development. Annotated speech corpora are essential for a wide range of disciplines concerned with spoken human communication. Among many uses we can find the emergence of new theories in natural language processing; the training of linguistic and acoustic statistical models for speech technologies; the constitution of a large set of acoustic units for TTS.

An annotated corpus is made of raw language data and linguistic annotations. The basic data may be in the form of time functions, like audio recordings, or may be textual. Linguistic annotation covers any descriptive or analytic notations applied to that data and may concern transcriptions of all sorts, from phonetic features to discourse structures, including prosodic annotations, syllabification, part-of-speech and sense tagging, syntactic analysis, "named entity" identification, co-reference annotation, and so on.

Such databases are typically multidimensional, heterogeneous and dynamic. The annotations, obtained by manual or automatic process, come from various tools and are stored in many files usually in different formats. The growth in the use of corpora increases the need of tools for editing, manipulating, concentrating in a unique object and querying the annotations. As few discussions were about format of annotations file –reaching quite a consensus on XML format– some new architectures allowing manipulation and use of corpus annotations arose in the 2000s.

In the past years, few systems have been proposed to cope with the problem of the representation of several levels of speech annotations in a unique structure. Bird *et al.* [BL01] proposed a representation by a direct acyclic graph (Annotation Graph) with structured records on the arcs and optional time references on the nodes. The Annotation Graph has been implemented in two applications : ATLAS [BDG⁺00] and AGTK [BMM⁺02]. Another graph representation, HRG (Heterogeneous Relation Graph) has been developed by Taylor *et al.* [TBC01] in order to build a fast and efficient system for speech synthesis purposes. In HRG, linguistic information are represented by a graph where nodes do not contain any information in such but are linked to linguistic items (words, syllables, phonemes, ...) and where edges define the relationship between these items. Unlike Annotation Graph, HRG avoids replication of information to eliminate inconsistencies especially with respect to time. In [RK07], combined with finite-state machines, Heterogeneous Relation Graphs are used for linguistic information representation

-
- [BL01] S. BIRD, M. LIBERMAN, "A formal framework for linguistic annotation", *Speech Communication* 33, 01, 2001, p. 23–60.
- [BDG⁺00] S. BIRD, D. DAY, J. GAROFOLO, J. HENDERSON, C. LAPRUN, M. LIBERMAN, "ATLAS: A flexible and extensible architecture for linguistic annotation", *in: Proc. of LREC*, p. 1699–1706, 2000.
- [BMM⁺02] S. BIRD, K. MAEDA, X. MA, H. LEE, B. RANDALL, S. ZAYAT, "TableTrans, MultiTrans, InterTrans and TreeTrans: Diverse Tools Built on the Annotation Graph Toolkit", *in: Proc. of LREC*, p. 364–370, 2002.
- [TBC01] P. TAYLOR, A. W. BLACK, R. CALEY, "Heterogeneous relation graphs as a formalism for representing linguistic information", *Speech communication* 33, 2001, p. 153–174.
- [RK07] M. ROJC, Z. KACIC, "Time and space-efficient architecture for a corpus-based text-to-speech synthesis system", *Speech Communication* 49, 3, 2007, p. 230–249.

and feature construction. At the same period, Cassidy *et al.* [CH96,CH01] have developed a tool called Emu for manipulating speech corpus annotations. An Emu database is made of levels of annotation containing tokens associated or not with time information. Three kinds of relationships can exist between tokens both within and between levels: domination, sequence or association relations. More recently, Veaux *et al.* [VBR08], in the IrcamCorpusTools platform, organize the annotation representations into two classes: the first one gives annotation information and the second one the hierarchical and/or sequential relations between them.

In our opinion, an ideal annotation or transcription system for speech utterances should fulfil two basic requirements: to propose a functional split based on theoretical principles that ensure the fundamentals of a speech act (acoustics, phonetics, phonology, grammar and so on) and to evacuate all dependencies on the experimental settings or the application that might use this corpus. For instance, on the level of phonology, we have chosen to make an extensive use of the IPA system. Thus, a vowel such as the phoneme /a/ is defined by the whole set of descriptors including tongue position, aperture level or lips rounding gesture. The choice of a label comes next, and thus may be multiple in order to adapt to various descriptive constraints. It is thus possible to tie an IPA description, which by definition is unique, to various symbols (sampa alphabet, proprietary alphabet and so on). This point of view facilitates the extension of the data structure to various languages. The same applies for all levels of description of an utterance.

A second challenge deals with the interdependence of different levels of description for a speech utterance. In speech processing, and more generally for natural language processing, the use of corpora requires matching descriptors located on different levels of analysis. Thus, a speech synthesis system may process a set of acoustic segments characterized by a complex request jointly describing acoustic, grammatical, and syntactic features. For example, to build up candidate units corresponding to open syllables ending with sound [i] and located in the initial position of a word which is located itself at the end of a sentence. We could also extend this request by seeking the 10 closest candidates of a target sound characterized by a specific spectral profile. This situation requires a complex set of relations from the acoustic signal to its linguistic description. Offering different levels of description for the same statement also requires the use of different annotation tools (it may be an automatic system or a simple editor). It is hardly conceivable to have a unique tool to cover the different theoretical areas underlying the observed speech phenomenon. The use of such a combination of different annotation systems raises the problem of data consistency through these different levels and over time. The description system has to take into account this issue by providing a timestamp that will, in turn, enable a human expert or an automatic processing tool to determine the obsolescence of one information compared to another one.

Considering all these points and within the framework of speech processing, we propose an annotation structure that meets the previous requirements and also the following ones :

-
- [CH96] S. CASSIDY, J. HARRINGTON, "Emu: An enhanced hierarchical speech data management system", *Proc. of the 6th Australian Int. Speech Science and Technology Conf.*, 1996, p. 361-366.
 - [CH01] S. CASSIDY, J. HARRINGTON, "Multi-level annotation in the EMU speech database management system", *Speech Communication* 33, 01, 2001, p. 61-77.
 - [VBR08] C. VEAUX, G. BELLER, X. RODET, "IrcamCorpusTools: an extensible platform for speech corpora exploitation", *in: Proc. of LREC*, 2008.

- The structure should be multilingual;
- Annotation may be partial or incomplet;
- The structure should avoid redundant information;
- The information should be easily uploaded.

Our structure is based on a set of sequences of homogeneous linguistic data and relations linking the elements of these sequences.

3.4.5 Voice transformation

A Voice transformation system modifies a *source* speaker's speech utterance to sound as if a *target* speaker had spoken it [KM98]. This technology offers a number of useful applications in computer interfaces, health and biometrics. For instance, human/computer voiced interaction would be enhanced if a large variety of high quality synthesized voices were used [CCCL03]. On the biometrics level, a transformed voice could pose as a real voice in order to test a voice-based authentication system. Finally, in the health domain, some transformation techniques could give back their voice to handicapped persons.

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. 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 (model) the voice characteristics to be transformed. Second, the transformation should be found (computed).

The segmental acoustic space of a speaker is hard to model and is strongly linked to the phonetic characteristics of the language [TBT05]. A segmental voice transformation technique, should rather take into account the perceptive differences between two utterances of a same sound by two speakers than the phonetic characteristic of the sound itself. These differences are due to variation in the speakers' physiology and their sociological background. 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

-
- [KM98] A. KAIN, M. MACON, "Spectral voice conversion for text-to-speech synthesis", *in: International Conference on Acoustics, Speech, and Signal Processing, 1*, p. 285-288, 1998.
- [CCCL03] Y. CHEN, M. CHU, E. CHANG, J. LIU, "Voice conversion with smoothed GMM and MAP adaptation", *in: Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech)*, p. 2413 - 2416, Geneva, Switzerland, 2003.
- [TBT05] T. TODA, A. BLACK, K. TOKUDA, "Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter", *in: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, p. I9-I12, Philadelphia, USA, 2005.

the acoustic space of a speaker is segmented. Then, a specific transformation is separately implemented on each segment of this acoustic space.

The first step of our study consisted in comparing several GMM-based voice conversion systems (VCS-GMM). This statistical modeling is robust and is often used in VCS [SCM98] [KM98] [CCCL03] for it conducts a fair classification and captures speakers' specificities. A set of conversion functions based on GMM models has been derived and their efficiency compared on identical corpora. During a second step, we studied the influence of the degradation of learning data on the efficiency of GMM-based conversion functions. The pursued purpose was to maintain fair conversion quality for voice-conversion application when very few training data is available for both source and target speakers. Then, we quantified the relationship between the amount of training data and the optimal number of parameters of the VCS. For all the proposed systems, the learning process showed *over-fitting* issues. This default was observed (and measured) thanks to the divergence of the results between the learning corpus and the test corpus as the number of gaussian component of the model increase. At the conclusion of this study, we proposed three alternatives for robust conversion functions in order to minimize this risk. Further works were then led to assess the influence of the degradation of the learning process on the GMM-based VCS, such as reduction of the training corpus and the lack of parallel training data.

On an other level, we studied alternate methods to the linear conversion transform classically used in GMM-based VCS. This resulted in the use of a neural network that learns a conversion function based on Radial Basis Functions (RBF). The use of RBF was justified by the fact that it is a universal estimator able to model speakers traits with a reduced set of parameters.

3.4.6 Prosody transformation

A voice transformation system has to satisfy two main requirements: a transformation of the segmental acoustic features and the one of supra-segmental features. We focus here on prosody transformation and more particularly on the duration and the fundamental frequency, F_0 . Usually, such a transformation system can be decomposed into three stages: stylization, classification and then transformation.

Concerning F_0 stylization, numerous models of such contours are introduced in the literature, like symbolic models based on tags^[Mer02], quantitative generative models using phrase and accent commands^[Fuj04], dynamical state space model which consider that the observation of a portion of the melody is explained by a stochastic state variable^[RO99]. Especially, a wide range of publications deals with stylization of F_0 contours using polynomial functions. We can

-
- [SCM98] Y. STYLIANOU, O. CAPPE, E. MOULINES, "Continuous probabilistic transform for voice conversion", *IEEE Transactions on Speech and Audio Processing* 6, 2, March 1998, p. 131 – 142.
- [Mer02] P. MERTENS, "Synthesizing elaborate intonation contours in text-to-speech for french", *in: Proceedings of the Speech Prosody Conference*, p. 499–502, Aix-en-Provence, France, 2002.
- [Fuj04] H. FUJISAKI, "Information, prosody, and modeling - with emphasis on tonal features of speech", *in: Proceedings of Speech Prosody Conference*, p. 1–10, Nara, Japan, 2004.
- [RO99] K. N. ROSS, M. OSTENDORF, "A dynamical system model for generating fundamental frequency for speech synthesis", *IEEE Transactions on Speech and Audio Processing* 7, 3, 1999, p. 295–309.

cite models like MoMel^[HCE00], Tilt^[Tay00], B-spline model ^[LBB06b,LBB06a], as well as Sakai and Glass's model^[SG03] based on regular spline functions.

In addition to the modeling issue, the model has to be appropriate to a classification providing a melodic space characterization of the speaker by a class set. As for the melodic contour classification issue, few works deal with an unsupervised F0 clustering ^[YIS03,Rei07]. The problem is to derive a set of basic melodic patterns from a set of sentences from which F0 has been previously computed. The idea is that concatenation of elementary F0 contours can characterize a complete melodic sentence. The major difficulty is to take into account the syllable duration. Two melodic contours with different temporal supports can represent the same elementary melodic pattern. A recent approach based on Hidden Markov Models (HMM) enables a time independent comparison ^[LBB07b,LBB07a].

In the literature, a great amount of recent works deals with prosody transformation and more particularly F0 ^[GK03]. A standard approach consists in modifying the F0 by applying a linear or polynomial transformation which is based on global parameters of the source and the target voices ^[CH98,CVW02]. Some other approaches decompose that complex transformation problem into subproblems doing a partition of the feature space, as done for example by the

-
- [HCE00] D. HIRST, A. D. CRISTO, R. ESPESER, "Levels of representation and levels of analysis for the description of intonation systems", *Prosody : Theory and Experiment 14*, 2000, p. 51-87.
- [Tay00] P. TAYLOR, "Analysis and synthesis of intonation using the Tilt model", *Journal of the Acoustical Society of America 107*, 2000, p. 1697-1714.
- [LBB06b] 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)*, p. 333-338, Saint Petersburg, Russia, 2006.
- [LBB06a] 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*, P. Sojka, I. Kopeček, K. Pala (editors), 4188, Springer Verlag, p. 423-430, Berlin, Heidelberg, 2006.
- [SG03] 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*, p. 712-717, St. Thomas, U.S. Virgin Islands, 2003.
- [YIS03] Y. YAMASHITA, T. ISHIDA, K. SHIMADERA, "A stochastic F0 contour model based on clustering and a probabilistic measure", *IEICE Transactions on Information and Systems E86-D*, 3, 2003, p. 543-549.
- [Rei07] U. D. REICHEL, "Data-driven extraction of intonation contour classes", in: *Proceedings of the 6th ISCA Tutorial and Research Workshop on Speech Synthesis (SSW6)*, W. P., A. J., H. W. (editors), p. 240-245, Bonn, Germany, 2007.
- [LBB07b] D. LOLIVE, N. BARBOT, O. BOËFFARD, "Unsupervised HMM classification of F0 curves", in: *Proceedings of Interspeech'2007*, p. 478-481, Antwerp, Belgium, 2007.
- [LBB07a] D. LOLIVE, N. BARBOT, O. BOËFFARD, "Clustering algorithm for F0 curves based on hidden Markov models", in: *Proceedings of the 6th ISCA Tutorial and Research Workshop on Speech Synthesis (SSW6)*, P. Wagner, J. Abresh, W. Hess (editors), p. 85-89, Bonn, Germany, 2007.
- [GK03] B. GILLET, S. KING, "Transforming F0 contours", in: *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech)*, p. 1713-1716, Geneva, Switzerland, 2003.
- [CH98] D. T. CHAPPELL, J. H. L. HANSEN, "Speaker-specific pitch contour modeling and modification", in: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2, p. 885-888, Seattle, Washington, USA, 1998.
- [CVW02] T. CEYSSENS, W. VERHELST, P. WAMBACQ, "On the construction of a pitch conversion system", in: *Proceedings of EUSIPCO*, p. 1301-1304, Toulouse, France, 2002.

codebook solution [CH98,HN07].

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

- To propose a method to stylize and cluster F_0 melodic contours in order to characterize, in an unsupervised manner, the melodic space of a speaker.
- To relax the constraint of having parallel corpora in prosody transformation systems which would soften software design.

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 one, booking services by telephone, computer assisted language learning or even intelligent transport systems. Concerning this last point, we are participating in a study of feasibility for a voice commanded jacket that is designed for signalling the intentions of cyclists and increase their security.

5 Software

The major development of this year is first presented and is a data structure developed to represent speech corpora. Then we present the other softwares we are developing to demonstrate progress in the field of speech synthesis (Talking Head), to evaluate new techniques (Web based listening test system) and to segment speech corpora. The last point which is presented is dedicated to the DORIS platform that is used to promote collaborative projects with industrial partners.

5.1 ROOTS

Participants: Nelly Barbot, Vincent Barraud, Olivier Boeffard, Laure Charonnat, Arnaud Delhay, Sébastien Le Maguer, Damien Lolive,.

ROOTS is an integrated object oriented software for describing speech utterances. Defining complex objects by means of encapsulation is relatively easy, for example see further the class Utterance that models a speech statement. The set of classes can be used in a comprehensive manner or fractionally. The design of this software is based on two fundamental notions: firstly, sequences of items, for example a sequence of words, a sequence of syllables, or a sequence of

[HN07] E. HELANDER, J. NURMINEN, "A novel method for prosody prediction in voice conversion", *in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 4, p. 509–512, 2007.

melodic patterns and then relations between sequences. As in [TBC01], ROOTS focuses on the non-redundancy of information in order to simplify sequences and relations handling. Only sequences hold descriptive information. A relation is a correspondence between two sequences based on the ranking of elements present in the sequences. The proposed system ensures a clear separation between description and treatment. The limit of ROOTS, and its strength, is that it offers only a descriptive function, information processing runs outside.

By the availability of different classes, ROOTS provides a structural framework that covers the main features describing a spoken statement. A specific class will define what is each item, for example: a word, a syllable, a Part-Of-Speech (POS), an acoustic segment, etc. The important point to consider is that the semantic interpretation of these objects is left to the application. For example, the coexistence of several sequences of words is possible, a sequence of raw words, a sequence of corrected words, a sequence of words including groupings such as locutions, etc. All these semantic variations are unknown for ROOTS, it only provides the mean to describe a sequence by a label. ROOTS can be extended easily with new types that inherit from sequences and new types of items.

ROOTS proposes, for each of its objects, a serialization operator towards an XML external description. Each object is then responsible for its own external descriptors and is capable of loading such an XML description when it is created. This serialization mechanism guarantees the encapsulation hierarchy for complex objects. In the same way, we have added a graphical output mechanism in order to build figures in the L^AT_EX/PGF format. This visualization tool is very convenient to expertise and analyse the content of the structure. At the same time, an import/export mechanism towards non XML files is provided. It relies on the most common practices, for example the HTK format in order to describe acoustic segments and labels and to guarantee the possible usage of common tools like Wavesurfer.

Currently, a complete ROOTS prototype has been written in PERL. This prototype is divided into three main packages that correspond to the different description levels: *linguistic*, which contains classes used to represent syntax and part-of-speech; *phonology*, which contains phoneme and syllable specifications; *acoustic*, which contains low level items (allophones, non-speech-sounds, F_0 and segments). The API software documentation is available online from the homepage of ROOTS at <http://www.irisa.fr/cordial/roots>.

5.2 TALKING HEAD

Participants: Vincent Barreaud, Olivier Boeffard.

This software is dedicated to exhibit advances of CORDIAL in speech technology. Its purpose is to produce a visual output (a web page, a stand-alone application for a scientific event, ...) for non-specialist audience.

The core of the system consists in three parts. The first part is a 3D model of a face with control points located on the lips, the tongue, the jaw. This model can be derived on a specific software such as Blender or DAZ or simply fetch on one of many free open source repositories. The second part is a physical engine which purpose is to animate the control points from and

[TBC01] P. TAYLOR, A. W. BLACK, R. CALEY, "Heterogeneous relation graphs as a formalism for representing linguistic information", *Speech communication* 33, 2001, p. 153–174.

to predefined positions ("viseme" for "éléments de visage" or facexel for "face elements"). The last part is a module that analyzes the audio segment and defines, according to the sequence of constituting phonemes which visemes (actually di-visemes) is to be played.

The main purpose of this system is to synchronously play an audio stream and animate the face accordingly. Models can be described in XML and can be modified. We used a Java based engine (JMonkey) in order to easily interface it with the DORIS software. Indeed, future version of this Talking Head should integrate a MRCP client that would interface with DORIS. For future development, we plan to interact with other research teams of the Media and Interaction Department of IRISA.

5.3 WEB BASED LISTENING TEST SYSTEM

Participants: Laurent Blin, Vincent Barraud, Olivier Boeffard.

In order to perform subjective tests for voice transformation and Text-to-Speech techniques, we developed a web based listening test system that can be used with a large range of testees (listeners). Our main goals were to make the configuration of the tests as simple and flexible as possible, to simplify the recruiting of the testees and, of course, to keep track of the results using a relational database.

This first version of our system can perform the most widely used listening tests used in the speech processing community (AB-BA, ABX and MOS tests). It can also easily perform other tests derived by the tester by means of a test configuration file [BBB08]. This system has been proposed to the speech processing community and used for several perceptual tests.

5.4 AUTOMATIC SEGMENTATION SYSTEM

Participants: Laure Charonnat, Olivier Boeffard.

The automatic segmentation system consists of a set of scripts aligning the phonetic transcription of a text with its acoustic signal counterpart. The system is made of two parts: the first one generates a phonetic graph including phonological variants (pauses, liaisons, schwas,...), the second one, based on HMM modelling, performs a Viterbi decoding determining the most likely phonetic sequence and its alignment with the acoustic signal.

To be efficient, the system should be applied to texts that have been manually checked (compliance with the recording, spelling, syntax) and annotated. The annotation stage consists in adding tags indicating excerpts in foreign language, non standard pronunciation and noises (breathing, laughter, coughing, sniffing, snorting, etc.). It is also possible to improve the decoding performances by adding a list of phonetisation of proper names and foreign pronunciations.

A new step of development of this automatic segmentation system has been reached within the framework of a collaboration with France Telecom (see 7.4). The segmentation learning step can be incremental on each new sentence (instead of working with a batch of pre-recorded

[BBB08] L. BLIN, O. BOEFFARD, V. BARREAU, "WEB-Based listening test system for speech synthesis and speech conversion evaluation", in: *Proceedings of the 6th International Language Resources and Evaluation (LREC'08)*, E. L. R. A. (ELRA) (editor), Marrakech, Morocco, 2008.

files). This new scenario needs speaker adaptation techniques and allows the creation of new TTS voices on the fly.

5.5 DORIS platform

Participants: Laurent Miclet [*correspondant*], Olivier Boeffard, Laurent Blin, Vincent Barreaud, Damien Lolive,.

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 of vocal and dialogue technologies. The main funding comes from a CPER, namely through the platform INVENT'IST MOB-ITS which is common to the three Irisa projects in Lannion (Cordial, Cairn and Pilgrim). For the Cordial project, DORIS is the core of MOB-ITS.

DORIS is concerned by the different research projects like GEORAL and ORDICTÉE. In November 2005, a research engineer (initially founded by FNADT, now by CPER) has been installed full time on the DORIS/MOB-ITS platform to manage the technical aspects and to develop new softwares for the previously quoted projects.

5.5.1 Hardware architecture

On the powerhouse systems side, a cluster of Dell PowerEdge systems 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 1To.

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 speed internet connection with VPN and access to collaborative software available as to make the Cordial project members' work easier.

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 5.5.3).

5.5.2 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-agent⁷ architecture. It satisfies the needs for modularity, quick and clean development and interoperability.

⁷An 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.

To fulfill this role, we chose and installed JADE⁸, a software framework fully implemented in Java language. It allows the implementation of multi-agent systems through a middle-ware that complies with the FIPA⁹ 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.

5.5.3 New steps with DORIS

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) or SIP (Session Initiation Protocol) which help to establish control connections to external media streaming devices, and media delivery mechanisms like RTP (Real Time Protocol).

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. An API for MRCP clients has been developed in Java.

6 New Results

New results in 2010 are presented here according to the paragraphs' numbers in section *Scientific Foundations*.

6.1 Dialogue and modeling

6.1.1 Logical modeling for dialogue processing

This research topic has no new results in 2010.

⁸Java Agent Development Framework, a free software distributed by Telecom Italia Lab (TILAB).

⁹Foundation 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.

6.1.2 Modeling the Communicative Common Ground (CCG)

Participants: Sylvie Saget, Marc Guyomard.

This research topic has no new results in 2010. The Ph.D. student is currently writing her manuscript.

6.1.3 Hability modeling

Participants: Karl DeVooght, Marc Guyomard.

This research topic has no new results in 2010.

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 Multimodal interactive system: GEORAL TACTILE and reference

This research topic has no new results in 2010.

6.2.2 ORDICTÉE

Participants: Marc Guyomard.

This research topic has no new results in 2010.

6.3 Machine learning in dialogue systems

6.3.1 Grammatical inference

This research topic has no new results in 2010.

6.3.2 Analogical learning of sequences and tree structures

Participants: Anouar Ben Hassena, Laurent Miclet.

A Ph.D. thesis work has begun in october 2007 with Anouar Ben Hassena. The bibliographical work has been oriented firstly to study the work by Stroppa and Yvon on the definition of an analogical proportion between trees (*analogy by factorisation*), secondly to the study of edit distances between trees. A particular interest has been given to constrained edit distances and alignment of trees.

Starting from this basis, a first work has been realised: the modification of the tree alignment algorithm by Jiang^[JWZ94] to enable the alignment of more than two trees. An algorithm has been devised and implemented (*analogy by alignment*), which finds the best alignment between four labelled trees, assuming that some analogical proportion (and an associated analogical dissimilarity) exists on the set of labels.

This gives an alternative definition to the analogical proportion between trees. The relationship between the two definitions has been explored, as well as the computational issues. Our algorithm can compute the analogical dissimilarity by alignment between four trees in polynomial time whereas the factorization technique checks in exponential time whether four trees are in analogical proportion. This difference is the consequence of the fact that some factorizations cannot be realised in the framework of alignment: our method gives a more restrictive definition of analogical proportion between trees.

This work has produced in 2010 a series of conference papers ([4, 7, 6, 5]). Experiments on learning by analogy the syntactic structure of natural language sentences are currently undertaken. Preliminary results are promising.

6.3.3 Miscelleaneous works on analogy

Participants: Nelly Barbot, Laurent Miclet, Henri Prade (IRIT).

Analogy in finite groups This research topic has no new results in 2010.

The logic of analogy This research topic has no new results in 2010.

6.4 Speech Processing

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

Participants: Sébastien Le Maguer, Nelly Barbot, Olivier Boeffard.

This work is covered within the framework of the PhD thesis of Sébastien Le Maguer, funded by le Conseil Général des Côtes d’Armor (7.3), started on 20th of October 2008.

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

[JWZ94] T. JIANG, L. WANG, K. ZHANG, “Alignment of Trees - An Alternative to Tree Edit”, *in: CPM ’94: Proceedings of the 5th Annual Symposium on Combinatorial Pattern Matching*, Springer-Verlag, p. 75–86, London, UK, 1994.

sequences. To limit the search space for the best acoustic sequence, current TTS systems impose a search heuristic usually based on a dynamic principle (additive cost functions, etc).

The unit selection process has to answer a double objective. On the one hand, one needs to find a good match between a subsequence of the phonemic sequence and an arbitrary sequence of phonemes in the speech database. A supplementary mechanism defining concatenation costs is needed in order to judge potential acoustic mismatches at a sequence level. The difficulty of the problem lies in the fact that both criteria are combined. The choice of a matching sub-sequence of acoustic units depends on its left (past) and right (future) contexts. This combinatorial issue can be solved as a best path search within a graph. The majority of TTS systems consider the Viterbi algorithm. This algorithm, effective in both spatial and temporal complexities, is justified by the fact that the expression of the global cost of a sequence of units is defined by an additive recurrence. This justification is widely approved by the speech community as for the expression of concatenation and target costs. But considering global costs like ones of prosodic nature, a phone-based additive recurrence is much harder to justify because these phenomena take place on the scale of the intonative group and of the sentence.

We consider that it is possible to enhance the quality of the current TTS systems by taking into consideration supra-segmental criteria during the search for the optimal acoustic sequence. Taking into account these global criteria is not easy as it is necessary to define new descriptive models for the segmental and supra-segmental costs of a sequence. New techniques of selection should be capable to propose synthetic voices with more contrast or expressiveness while maintaining a very good sound quality. During this year, we have considered a speech parametrical synthesis system based on HMM, called HTS. Indeed, since few years, parametrical speech synthesis methods are considered to help the acoustic unit selection [KB06]. This hybrid approach, between parametrical synthesis and concatenative TTS systems, has been used in Ximera system to generate prosodic parameters that are used as targets for the segment selection [KTN⁺04,TKH⁺06]. In [RR05], an acoustic target is generated from a HMM-based speech synthesis system and is explicitly used in a cost function during the selection process. One of the objectives of the thesis is a theoretical proposal in order to use global dependences when searching for the best sequence of acoustic units, and to combine of a HMM-based and unit-selection based algorithms.

The use of HTS includes a training stage. The training corpus is composed by the signal and its description. This description contains information from different levels (acoustic, phonology, linguistic). Actually our work is focused on the creation of such a corpus for French. We are

-
- [KB06] J. KOMINEK, A. W. BLACK, “The Blizzard challenge 2006 CMU entry introducing hybrid trajectory-selection synthesis”, *research report*, CMU, 2006.
- [KTN⁺04] H. KAWAI, T. TODA, J. NI, M. TSUZAKI, K. TOKUDA, “Ximera: A new TTS from ATR based on corpus-based”, *in: Proceedings of the ISCA Tutorial and Research Workshop on Speech Synthesis (SSW5)*, p. 179–184, 2004.
- [TKH⁺06] T. TODA, H. KAWAI, T. HIRAI, J. NI, N. NISHIZAWA, J. YAMAGISHI, K. TOKUDA, S. NAKAMURA, “Developing a test bed of english Text-to-Speech system XIMERA for the Blizzard challenge 2006”, *in: Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, 2006.
- [RR05] S. ROUBIA, O. ROSEC, “Unit selection for speech synthesis based on a new acoustic target cost”, *in: Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, p. 2565–2568, 2005.

conducting experiments in order to assess the influence of the descriptive features on the speech parameter generation derived from HTS.

6.4.2 Optimal corpus design

Participants: Nelly Barbot, Arnaud Delhay, Olivier Boeffard.

In order to pursue the comparison between the greedy strategy and the Lagrangian based algorithm^[CBB08] to design a rich linguistic corpus, complementary experiments are currently carried on. The objective is to compare different statistics of both algorithm solutions (size, phonetic composition and distribution, etc.) in terms of confidence intervals. According to the matrix representation of the initial corpus, sentences match up to matrix columns. The impact on computation time and solution quality of the perturbation of the initial corpus and its associated matrix has been studied. Indeed, if both algorithms seem to be robust to the sentence ranking perturbation in the corpus, we have observed that the mixing of the matrix columns implies a severe increase of the computation time.

6.4.3 Automatic speech labeling and segmentation

Participants: Laure Charonnat, Gaëlle Vidal, Vincent Barraud, Olivier Boeffard.

New results in the field of automatic speech segmentation are linked to a collaboration with France Telecom (7.4) about the development of tools for personalized voices creation. The context is a web application allowing people to record themselves and create their own voice for speech synthesis. This context implies new constraints for the segmentation task. First, the segmentation, done on each recorded sentence, must be fast enough to enable a new computation of the database and rapidly produce speech synthesis; secondly, as the conditions of recording are not monitored, the segmentation system has to assert that the recorded sentence is consistent with the phonetic transcription. Some new Hidden Markov Models (HMM) have been initialised on our manually segmented corpora and reestimated on automatic segmented speech of 55 speakers. These multispeakers models have improved the results of the automatic segmentation system. Some speaker adaptation mechanisms have been studied. The new segmentation software includes MLLR^[LW95] and MAP^[GLH92] speaker adaptation. Finally, a system of decision based on duration models is under consideration to validate the phonetic transcription.

[CBB08] J. CHEVELU, N. BARBOT, O. BOEFFARD, A. DELHAY, “Comparing set-covering strategies for optimal corpus design”, in: *Proceedings of the 6th International Language Resources and Evaluation (LREC’08)*, E. L. R. A. (ELRA) (editor), Marrakech, Morocco, 2008.

[LW95] C. LEGGETTER, P. WOODLAND, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models”, *Computer Speech & Language* 9, April 1995, p. 171–185.

[GLH92] J.-L. GAUVAIN, C.-H. LEE, M. HILL, “MAP Estimation of Continuous Density HMM : Theory and Applications”, in: *Workshop on Speech and Natural Language*, p. 185–190, 1992.

6.4.4 Corpus annotation structure

Participants: Nelly Barbot, Vincent Barreaud, Olivier Boeffard, Laure Charonnat, Arnaud Delhay, Sébastien Le Maguer, Damien Lolive.

Within this topic, we have developed a software that is an original contribution that allows to represent in a joint manner several description levels of a speech utterance. ROOTS is particularly well adapted to the description of speech corpora, is a priori independent from the language, and can be easily extended to new descriptions. Two significant notions support the software architecture: first, sequences guarantee a time-based order relation, and second, relations allow to connect the set of sequences. One originality of ROOTS resides in an algebraic modeling of relations between sequences (technically, using a relation is just accessing a matrix structure). Such a modeling allows to compose very efficiently (both in terms of spatial and time complexity) new relations that can be missing from the description. Thus, for example, it is possible to create a relation from a word sequence towards an allophone sequence only by using two intermediate relations which are word/syllable and syllable/allophone relations. During the software specification, focus has been put on a serialization process towards XML. Each object is then responsible for its own data. This solution enables us to think about the use of ROOTS in applications with a data flow architecture where processes feed a unified and coherent structure.

In terms of perspectives, we plan to continue the extension of certain levels, notably for the acoustic description, and to improve the unification of other levels like the linguistic one, in particular by getting closer to ISO standards in order to facilitate a multilingual usage.

An article describing all this work has been submitted to the special issue on “New Frontiers in Rich Transcription” of the IEEE Transactions on Audio, Speech and Language Processing.

6.4.5 Voice transformation

Participants: Larbi Mesbahi, Vincent Barreaud, Olivier Boeffard.

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 (7.2). This thesis started on November 2006 and was defended on October the 28th, 2010 [2].

The latest research phase concerns the enrichment of the learning sets with phonetic informations. Two goals are pursued in this scope. The first of them is to reduce the pairing errors in the DTW alignment for they result in an imprecise transformation function. In that case, the purpose is to avoid the pairing of a source vector from a phonetic class with a target vector of an other phonetic class. This problem often occurs and is referred as the "one-to-many" problem [GRC09]). The second objective is to find a method to design a minimal learning corpus. The underlying question here is to know if every phonetic class is needed during the learning phase.

[GRC09] E. GODOY, O. ROSEC, T. CHONAVEL, “Alleviating the One-to-Many Mapping Problem in Voice Conversion with Context-Dependent Modeling”, *in: Interspeech 2009*, p. 1627–1630, 6-10 September 2009.

To get around the one-to-many issue, we suggest to guide the DTW by restricting the pairing of a source vector to a target vector of the same phonetic class. This is possible thanks to efficient (automatic and manual) tagging methods that can match a phonetic class with a vector with little error. As in our previous work on SVQ-Tree, the goal here is to reduce the number of possible target vectors to be paired with a source vector. Phonetic information can reduce this set of vectors from a whole sentence to a dozen. Source and target sentences undergo a labeling phase and then, DTW is performed on the obtained phonetic classes.

At that point, two learning strategies can be applied. The first one consists in gathering all the paired vectors in a single training corpus and to learn a transformation function on it. The second strategy is to form several training sub-corpora from neighbouring (one or more) phonetic classes and to learn distinct transformation functions on those sub-corpora.

In parallel with this effort, a preliminary study on alternate parametrization has been led. Indeed, we believe that, the parametrization step contributes to the quality of the converted voice. To do so, a maximum of the speaker characterization must pass on the parametrized data. In this scope, we chose to use the True-Envelope characterization. But, as have shown previous studies, the dimensionality of this parametrization must be reduced for the data to be used as training material. To achieve that, Principal Component Analysis is used. This solution is even more efficient when used to derive phone-specific conversion functions.

6.4.6 Prosody transformation

Participants: Nelly Barbot, Damien Lolive, Olivier Boeffard.

Concerning the transformation of a voice at a supra-segmental level, this work begun within the framework of the PhD thesis of Damien Lolive, funded by the French Ministry of National Education and Research, started on October 2005 and defended on 27th of November 2008. During 2009, Damien Lolive has continued with working on this subject as Associate Professor. In this study, we are particularly interested in the melodic contour transformation from a source speaker into a target speaker ^[Lol08]. One of the main developed axis deals with melody contour stylisation which is a necessary step before tackling the problem of melody conversion between two speakers. Several papers presenting this model have been published

[Lol08] D. LOLIVE, *Transformation de l'intonation. Application à la synthèse de la parole et à la transformation de voix*, PdD Thesis, Université de Rennes 1, 2008.

previous years [BBL05,LBB06d,LBB06c,LBB06b,LBB06a]. The goal of this work was to propose an efficient model for melody stylisation and a b-spline model appears to be efficient enough to model melody contours. The second point addressed was the automatic selection of the number of parameters of the model. An Minimum Description Length -MDL- methodology has been proposed to optimise the size of the model. A paper [3] that sums up this work has been published this year in the special issue entitled “Model Order Selection in Signal Processing Systems” of the IEEE Journal of Selected Topics in Signal Processing.

7 Contracts and Grants with Industry

7.1 SEQUANA

A contract between Région Bretagne, Département des Côtes d’Armor and Université de Rennes 1 - IRISA allows a PhD student, Anouar Ben Hassena, to work on the field of learning prosody by analogy for a three year period.

7.2 TRANSPAR

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

7.3 UNIT SELECTION

A contract between le Conseil Général des Côtes d’Armor and the l’Université de Rennes 1 - ENSSAT allows a PhD student, Sébastien Le Maguer, to work on the field of unit selection for speech synthesis systems.

-
- [BBL05] 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)*, p. 325–328, Lisbon, Portugal, 2005.
- [LBB06d] 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, p. 219–234, 2006.
- [LBB06c] 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*, p. 499–502, Dinard, France, 2006.
- [LBB06b] 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)*, p. 333–338, Saint Petersburg, Russia, 2006.
- [LBB06a] D. LOLIVE, N. BARBOT, O. BOËFFARD, “Comparing B-spline and spline models for F_0 modelling”, *in: Lecture Notes in Artificial Intelligence - Proceedings of the 9th International Conference on Text, Speech and Dialogue - Brno, Czech Republic*, P. Sojka, I. Kopeček, K. Pala (editors), 4188, Springer Verlag, p. 423–430, Berlin, Heidelberg, 2006.

7.4 EXTERNAL RESEARCH CONTRACT : PERSONALIZED VOICES CREATION

The contract between France Telecom and the Université de Rennes 1 - ENSSAT concerns the realization of research works within the framework of corpus-based speech synthesis. The aim of the study is to supply tools and models for a web application dedicated to the creation of personalized voices for speech synthesis.

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 Elsnat, and of the French-speaking network FRANCIL (Réseau FRANCOphone d'Ingénierie de la Langue).

8.2 Intelligent Transport Systems

The Cordial team takes part in the scientific interest group (GIS) on Intelligent Transport Systems located in Brittany. In particular, we are participating in the feasibility study of a voice commanded signalling jacket for cyclists. This project has been initiated by the “**mc créativité +**” society who has a patent on it. This study is realized in collaboration with the Cairn project team of Irisa and has led to the recruitment of a trainee for one month and a half during the summer 2010.

9 Dissemination

9.1 Leadership within scientific community

Marc Guyomard has been a member of the Scientific Advisory Committee of YRRSDS 2010 (Young Researcher's Roundtable on Spoken Dialogue Systems), organised by Waseda University (<http://www.waseda.jp/top/index-e.html>), in Tokyo, Japan, on September 22nd and 23rd, 2010. Mar Guyomard has been a reviewer of The B method: from Research to Teaching Conference.

Laurent Miclet has been member of the Scientific Committees for the Congresses ICGI 2010, RFIA 2010 and CAp 2010.

Olivier Boeffard is an elected member of the board of SynSIG, the international ISCA special interest group on speech synthesis, http://www.synsig.org/index.php/Main_Page

9.2 Teaching at University

Olivier Boeffard teaches the course *Speech Synthesis* in the Master SISEA, 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 teaches 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 Supervisé* (AS) 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 co-supervised the french translation of Artificial Intelligence, a Modern Approach [1].

9.3 Conferences, workshops and meetings, invitations

Olivier Boeffard has been President of the jury of the PhD thesis of F. Villanvicencio : *Transformation of voice identity*. Thèse de Université Pierre et Marie Curie, ParisTech, 22th of March 2010.

Olivier Boeffard has been member of the jury of the PhD thesis of G. Degottex : *Glottal source and vocal-tract separation: Estimation of glottal parameters, voice transformation and synthesis using a glottal model*. Thèse de Université Pierre et Marie Curie, ParisTech, 16th of November 2010.

Laurent Miclet has been President of the jury of the PhD thesis of R. Lefort : *Apprentissage et classification faiblement supervisée : application en acoustique halieutique*. Thèse de Télécom Bretagne en habilitation conjointe avec l'Université de Rennes 1, Traitement du Signal et Télécommunications, le 29 Novembre 2010.

Vincent Barreaud and Olivier Boeffard have been members of the jury of the PhD thesis of L. Mesbahi : *Transformation automatique de la parole : étude des transformations acoustiques*. Thèse de Université de Rennes 1, 22th of October 2010 [2].

9.4 Graduate Student and Student intern

Cordial hosted the internship of Raphaël Chevalier (IUT INFO Lannion) from April to June 2010. The purpose of this project was to assess which technologies to use for the Talking Head demonstrator. This work was directed by Vincent Barreaud, Olivier Boeffard and Arnaud Delhay (as academic part).

10 Bibliography

Major publications by the team in recent years

- [1] O. BOEFFARD, C. D'ALESSANDRO, *Synthèse de la parole*, Hermès Science, New-York, 2002, ch. 3, p. 115–154.
- [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.”, *Revue d'Intelligence Artificielle*. 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*, 862, Springer Verlag, Berlin, Heidelberg, sep 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)*, 5, 2002.
- [6] H. FRANÇOIS, O. BOËFFARD, “Evaluation of units selection criteria in corpus-based speech synthesis”, in : *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech)*, p. 1325–1328, Geneva, Switzerland, 2003.
- [7] H. FRANÇOIS, *Synthèse de la parole par concaténation d’unités acoustiques : construction et exploitation d’une base de parole continue*, PhD Thesis, Université de Rennes 1, 2002.
- [8] M. GUYOMARD, P. NERZIC, J. SIROUX, “Plans, métaplans et dialogue”, *research report number 1169*, Irisa, September 1998.
- [9] 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”, *Speech Communication* 48, 10, 2006, p. 1319–1348.
- [10] S. KRSTULOVIC, F. BIMBOT, O. BOËFFARD, D. CHARLET, D. FOHR, O. MELLA, “Selecting Representative Speakers for a Speech Database on the Basis of Heterogeneous Similarity Criteria”, *Speaker Classification II, Selected Projects. Lecture Notes in Computer Science 4441*, 2007, p. 276–292.
- [11] L. MICLET, S. BAYOUDH, A. DELHAY, “Analogical Dissimilarity: Definition, Algorithms and Two Experiments in Machine Learning”, *JAIR* 32, August 2008.
- [12] S. NEFTI, O. BOËFFARD, T. MOUDENC, “Confidence measure for phonetic segmentation of continuous speech”, in : *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech)*, p. 897–900, Geneva, Switzerland, 2003.
- [13] 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)*, p. 131–138, University of Potsdam, Germany, 2006.
- [14] S. SAGET, “In favour of collective acceptance: Studies on goal-oriented dialogues”, in : *Proceedings of Collective Intentionality V*, Helsinki, Finland, 2006.
- [15] 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*, John Lee (ed), AAAI Press, 1998.
- [16] F. VIOLARO, O. BOËFFARD, “A hybrid model for text-to-speech synthesis”, *IEEE Transactions on Speech and Audio Processing* 6, 5, 1998, p. 426–434.

Books and Monographs

- [1] S. RUSSEL, P. NORVIG, *Intelligence Artificielle, 3ème édition*, Pearson, 2010, Laurent Miclet has co-supervised the french translation.

Doctoral dissertations and “Habilitation” theses

- [2] L. MESBAHI, *Transformation automatique de la parole, étude des transformations acoustiques*, PhD Thesis, Université de Rennes 1, 2010.

Articles in referred journals and book chapters

- [3] D. LOLIVE, N. BARBOT, O. BOEFFARD, “B-spline model order selection with optimal MDL criterion applied to speech fundamental frequency stylisation”, *IEEE Journal of Selected Topics in Signal Processing* 4, 3, 2010, p. 571–581.

Publications in Conferences and Workshops

- [4] A. BEN HASSENA, L. MICLET, “Analogical learning using dissimilarity between tree-structures”, *in: ECAI*, p. 1039–1040, Lisbon, Portugal, August 16-20, 2010.
- [5] A. BEN HASSENA, L. MICLET, “Appariement analogique de squences et d’arbres : Application au parsing automatique”, *in: GAOC’10 : atelier Graphe et Appariement d’Objets Complexes, en conjonction avec la confrence EGC’10*, p. A7–3 – A7–14, 2010.
- [6] A. BEN HASSENA, L. MICLET, “Tree analogical learning. Application in NLP”, *in: TALN*, 2010.
- [7] A. BEN HASSENA, L. MICLET, “Tree Analogical Matching : Definitions, Algorithms ans Applications”, *in: ISA 2010 : IADIS International Conference Intelligent Systems and Agents 2010*, p. 59–66, Freiburg, Germany, July 29-31, 2010.