RESEARCH CENTER
Nancy - Grand Est

FIELD

Activity Report 2014

Section Scientific Foundations

Edition: 2015-03-24

CAMUS Team

3. Research Program

3.1. Research directions

The various objectives we are expecting to reach are directly related to the search of adequacy between the sofware and the new multicore processors evolution. They also correspond to the main research directions suggested by Hall, Padua and Pingali in [28]. Performance, correction and productivity must be the users' perceived effects. They will be the consequences of research works dealing with the following issues:

- Issue 1: Static parallelization and optimization
- Issue 2: Profiling and execution behavior modeling
- Issue 3: Dynamic program parallelization and optimization, virtual machine
- Issue 4: Object-oriented programming and compiling for multicores
- Issue 5: Proof of program transformations for multicores

Efficient and correct applications development for multicore processors needs stepping in every application development phase, from the initial conception to the final run.

Upstream, all potential parallelism of the application has to be exhibited. Here static analysis and transformation approaches (issue 1) must be processed, resulting in a *multi-parallel* intermediate code advising the running virtual machine about all the parallelism that can be taken advantage of. However the compiler does not have much knowledge about the execution environment. It obviously knows the instruction set, it can be aware of the number of available cores, but it does not know the effective available resources at any time during the execution (memory, number of free cores, etc.).

That is the reason why a "virtual machine" mechanism will have to adapt the application to the resources (issue 3). Moreover the compiler will be able to take advantage only of a part of the parallelism induced by the application. Indeed some program information (variables values, accessed memory adresses, etc.) being available only at runtime, another part of the available parallelism will have to be generated on-the-fly during the execution, here also, thanks to a dynamic mechanism.

This on-the-fly parallelism extraction will be performed using speculative behavior models (issue 2), such models allowing to generate speculative parallel code (issue 3). Between our behavior modeling objectives, we can add the behavior monitoring, or profiling, of a program version. Indeed current and future architectures complexity avoids assuming an optimal behavior regarding a given program version. A monitoring process will allow to select on-the-fly the best parallelization.

These different parallelizing steps are schematized on figure 1 .

The more and more widespread usage of object-oriented approaches and languages emphasizes the need for specific multicore programming tools. The object and method formalism implies specific execution schemes that translate in the final binary by quite distant elementary schemes. Hence the execution behavior control is far more difficult. Analysis and optimization, either static or dynamic, must take into account from the outset this distortion between object-oriented specification and final binary code: how can object or method parallelization be translated (issue 4).

Our project lies on the conception of a production chain for efficient execution of an application on a multicore architecture. Each link of this chain has to be formally verified in order to ensure correction as well as efficiency. More precisely, it has to be ensured that the compiler produces a correct intermediate code, and that the virtual machine actually performs the parallel execution semantically equivalent to the source code: every transformation applied to the application, either statically by the compiler or dynamically by the virtual machine, must preserve the initial semantics. They must be proved formally (issue 5).

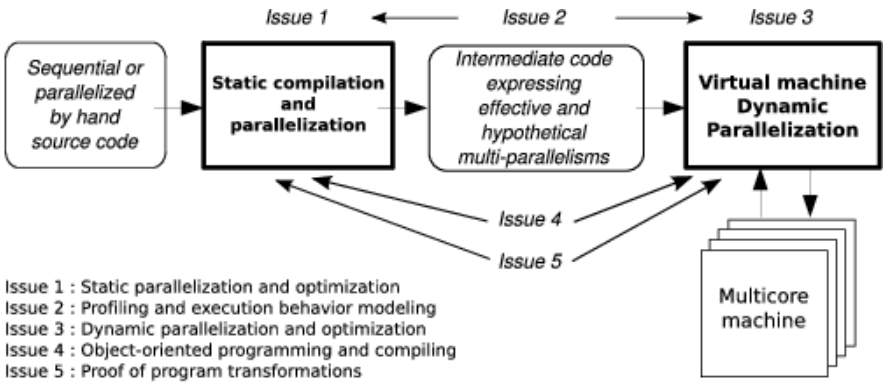

Figure 1. Automatic parallelizing steps for multicore architectures

In the following, those different issues are detailed while forming our global and long term vision of what has to be done.

3.2. Static parallelization and optimization

Participants: Vincent Loechner, Philippe Clauss, Éric Violard, Jean-François Dollinger, Aravind Sukumaran-Rajam, Juan Manuel Martinez Caamaño.

Static optimizations, from source code at compile time, benefit from two decades of research in automatic parallelization: many works address the parallelization of loop nests accessing multi-dimensional arrays, and these works are now mature enough to generate efficient parallel code [27]. Low-level optimizations, in the assembly code generated by the compiler, have also been extensively dealt for single-core and require few adaptations to support multicore architectures. Concerning multicore specific parallelization, we propose to explore two research directions to take full advantage of these architectures: adapting parallelization to multicore architecture and expressing many potential parallelisms.

3.3. Profiling and execution behavior modeling

Participants: Alain Ketterlin, Philippe Clauss, Aravind Sukumaran-Rajam.

The increasing complexity of programs and hardware architectures makes it ever harder to characterize beforehand a given program's run time behavior. The sophistication of current compilers and the variety of transformations they are able to apply cannot hide their intrinsic limitations. As new abstractions like transactional memories appear, the dynamic behavior of a program strongly conditions its observed performance. All these reasons explain why empirical studies of sequential and parallel program executions have been considered increasingly relevant. Such studies aim at characterizing various facets of one or several program runs, *e.g.*, memory behavior, execution phases, etc. In some cases, such studies characterize more the compiler than the program itself. These works are of tremendous importance to highlight all aspects that escape static analysis, even though their results may have a narrow scope, due to the possible incompleteness of their input data sets.

3.4. Dynamic parallelization and optimization, virtual machine

Participants: Aravind Sukumaran-Rajam, Juan Manuel Martinez Caamaño, Jean-François Dollinger, Alexandra Jimborean, Philippe Clauss, Vincent Loechner, Alain Ketterlin.

This link in the programming chain has become essential with the advent of the new multicore architectures. Still being considered as secondary with mono-core architectures, dynamic analysis and optimization are now one of the keys for controling those new mechanisms complexity. From now on, performed instructions are not only dedicated to the application functionalities, but also to its control and its transformation, and so in its own interest. Behaving like a computer virus, such a process should rather be qualified as a “vitamin”. It perfectly knows the current characteristics of the execution environment and owns some qualitative information thanks to a behavior modeling process (issue 2). It appends a significant part of optimizing ability compared to a static compiler, while observing live resources availability evolution.

3.5. Proof of program transformations for multicores

Participants: Éric Violard, Julien Narboux, Nicolas Magaud.

Our main objective consists in certifying the critical modules of our optimization tools (the compiler and the virtual machine). First we will prove the main loop transformation algorithms which constitute the core of our system.

The optimization process can be separated into two stages: the transformations consisting in optimizing the sequential code and in exhibiting parallelism, and those consisting in optimizing the parallel code itself. The first category of optimizations can be proved within a sequential semantics. For the other optimizations, we need to work within a concurrent semantics. We expect the first stage of optimizations to produce data-race free code. For the second stage of optimizations, we will first assume that the input code is data-race free. We will prove those transformations using Appel’s concurrent separation logic [29]. Proving transformations involving program which are not data-race free will constitute a longer term research goal.

CARAMEL Project-Team

3. Research Program

3.1. Cryptography, Arithmetic: Hardware and Software

One of the main topics for our project is public-key cryptography. After 20 years of hegemony, the classical public-key algorithms (whose security is based on integer factorization or discrete logarithm in finite fields) are currently being overtaken by elliptic curves. The fundamental reason for this is that the best algorithms known for factoring integers or for computing discrete logarithms in finite fields have — at best — a subexponential complexity, whereas the best attack known for elliptic-curve discrete logarithms has exponential complexity. As a consequence, for a given security level 2^n, the key sizes must grow linearly with n for elliptic curves, whereas they grow like n^3 for RSA-like systems. As a consequence, several governmental agencies, like the NSA (National Security Agency, USA) or the BSI (Bundesamt für Sicherheit in der Informationstechnik, Germany), now recommend to use elliptic-curve cryptosystems for new products that are not bound to RSA for backward compatibility.

Besides RSA and elliptic curves, there are several alternatives currently under study. There is a recent trend to promote alternate solutions that do not rely on number theory, with the objective of building systems that would resist a quantum computer (in contrast, integer factorization and discrete logarithms in finite fields and elliptic curves have a polynomial-time quantum solution). Among them, we find systems based on hard problems in lattices (NTRU is the most famous), those based on coding theory (McEliece system and improved versions), and those based on the difficulty to solve multivariate polynomial equations (UOV, for instance). None of them has yet reached the same level of popularity as RSA or elliptic curves for various reasons, including the presence of unsatisfactory features (like a huge public key), or the non-maturity (system still alternating between being fixed one day and broken the next day).

Returning to number theory, an alternative to RSA and elliptic curves is to use other curves and in particular genus-2 curves. These so-called hyperelliptic cryptosystems have been proposed in 1989 [32], soon after the elliptic ones, but their deployment is by far more difficult. The first problem was the group law. For elliptic curves, the elements of the group are just the points of the curve. In a hyperelliptic cryptosystem, the elements of the group are points on a 2-dimensional variety associated to the genus-2 curve, called the Jacobian variety. Although there exist polynomial-time methods to represent and compute with them, it took some time before getting a group law that could compete with the elliptic one in terms of speed. Another question that is still not yet fully answered is the computation of the group order, which is important for assessing the security of the associated cryptosystem. This amounts to counting the points of the curve that are defined over the base field or over an extension, and therefore this general question is called point-counting. In the past ten years there have been major improvements on the topic, but there are still cases for which no practical solution is known.

Another recent discovery in public-key cryptography is the fact that having an efficient bilinear map that is hard to invert (in a sense that can be made precise) can lead to powerful cryptographic primitives. The only examples we know of such bilinear maps are associated with algebraic curves, and in particular elliptic curves: this is the so-called Weil pairing (or its variant, the Tate pairing). Initially considered as a threat for elliptic-curve cryptography, they have proven to be quite useful from a constructive point of view, and since the beginning of the decade, hundreds of articles have been published, proposing efficient protocols based on pairings. A long-lasting open question, namely the construction of a practical identity-based encryption scheme, has been solved this way. The first standardization of pairing-based cryptography has recently occurred (see ISO/IEC 14888-3 or IEEE P1363.3), but the recent progress in discrete logarithms in finite fields will probably slow down its large deployment.

Despite the rise of elliptic curve cryptography and the variety of more or less mature alternatives, classical systems (based on factoring or discrete logarithm in finite fields) are still going to be widely used in the next decade, at least, due to resilience: it takes a long time to adopt new standards, and then an even longer time to renew all the software and hardware that is widely deployed.

This context of public-key cryptography motivates us to work on integer factorization, for which we have acquired expertise, both in factoring moderate-sized numbers, using the ECM (Elliptic Curve Method) algorithm, and in factoring large RSA-like numbers, using the number field sieve algorithm. The goal is to follow the transition from RSA to other systems and continuously assess its security to adjust key sizes. We also work on the discrete-logarithm problem in finite fields. This second task is not only necessary for assessing the security of classical public-key algorithms, but is also crucial for the security of pairing-based cryptography.

Another general application for the project is computer algebra systems (CAS), that rely in many places on efficient arithmetic. Nowadays, the objective of a CAS is not only to support an increasing number of features that the user might wish, but also to compute the results fast enough, since in many cases, the CAS are used interactively, and a human is waiting for the computation to complete. To tackle this question, more and more CAS use external libraries, that have been written with speed and reliability as first concern. For instance, most of today's CAS use the GMP library for their computations with big integers. Many of them will also use some external Basic Linear Algebra Subprograms (BLAS) implementation for their needs in numerical linear algebra.

During a typical CAS session, the libraries are called with objects whose sizes vary a lot; therefore being fast on all sizes is important. This encompasses small-sized data, like elements of the finite fields used in cryptographic applications, and larger structures, for which asymptotically fast algorithms are to be used. For instance, the user might want to study an elliptic curve over the rationals, and as a consequence, check its behaviour when reduced modulo many small primes; and then [s]he can search for large torsion points over an extension field, which will involve computing with high-degree polynomials with large integer coefficients.

Writing efficient software for arithmetic as it is used typically in CAS requires the knowledge of many algorithms with their range of applicability, good programming skills in order to spend time only where it should be spent, and finally good knowledge of the target hardware. Indeed, it makes little sense to disregard the specifics of the intended hardware platforms, even more so since in the past years, we have seen a paradigm shift in terms of available hardware: so far, it used to be reasonable to consider that an end-user running a CAS would have access to a single-CPU processor. Nowadays, even a basic laptop computer has a multi-core processor and a powerful graphics card, and a workstation with a reconfigurable coprocessor is no longer science-fiction.

In this context, one of our goals is to investigate and take advantage of these influences and interactions between various available computing resources in order to design better algorithms for basic arithmetic objects. Of course, this is not disconnected from the other goals, since they all rely more or less on integer or polynomial arithmetic.

CARTE Project-Team

3. Research Program

3.1. Computer Virology

From a historical point of view, the first official virus appeared in 1983 on Vax-PDP 11. At the same time, a series of papers was published which always remains a reference in computer virology: Thompson [71], Cohen [39] and Adleman [28]. The literature which explains and discusses practical issues is quite extensive [44], [46]. However, there are only a few theoretical/scientific studies, which attempt to give a model of computer viruses.

A virus is essentially a self-replicating program inside an adversary environment. Self-replication has a solid background based on works on fixed point in λ-calculus and on studies of von Neumann [75]. More precisely we establish in [35] that Kleene's second recursion theorem [59] is the cornerstone from which viruses and infection scenarios can be defined and classified. The bottom line of a virus behavior is

1. a virus infects programs by modifying them,
2. a virus copies itself and can mutate,
3. it spreads throughout a system.

The above scientific foundation justifies our position to use the word virus as a generic word for self-replicating malwares. There is yet a difference. A malware has a payload, and virus may not have one. For example, a worm is an autonous self-replicating malware and so falls into our definition. In fact, the current malware taxonomy (virus, worms, trojans, ...) is unclear and subject to debate.

3.2. Computation over continuous structures

Classical recursion theory deals with computability over discrete structures (natural numbers, finite symbolic words). There is a growing community of researchers working on the extension of this theory to continuous structures arising in mathematics. One goal is to give foundations of numerical analysis, by studying the limitations of machines in terms of computability or complexity, when computing with real numbers. Classical questions are : if a function $f : \mathbb{R} \to \mathbb{R}$ is computable in some sense, are its roots computable? in which time? Another goal is to investigate the possibility of designing new computation paradigms, transcending the usual discrete-time, discrete-space computer model initiated by the Turing machine that is at the base of modern computers.

While the notion of a computable function over discrete data is captured by the model of Turing machines, the situation is more delicate when the data are continuous, and several non-equivalent models exist. In this case, let us mention computable analysis, which relates computability to topology [43], [74]; the Blum-Shub-Smale model (BSS), where the real numbers are treated as elementary entities [34]; the General Purpose Analog Computer (GPAC) introduced by Shannon [69]with continuous time.

3.3. Rewriting

The rewriting paradigm is now widely used for specifying, modelizing, programming and proving. It allows one to easily express deduction systems in a declarative way, and to express complex relations on infinite sets of states in a finite way, provided they are countable. Programming languages and environments with a rewriting based semantics have been developed ; see ASF+SDF [36], MAUDE [38], and TOM [66].

For basic rewriting, many techniques have been developed to prove properties of rewrite systems like confluence, completeness, consistency or various notions of termination. Proof methods have also been proposed for extensions of rewriting such as equational extensions, consisting of rewriting modulo a set of axioms, conditional extensions where rules are applied under certain conditions only, typed extensions, where rules are applied only if there is a type correspondence between the rule and the term to be rewritten, and constrained extensions, where rules are enriched by formulas to be satisfied [30], [42], [70].

An interesting aspect of the rewriting paradigm is that it allows automatable or semi-automatable correctness proofs for systems or programs: the properties of rewriting systems as those cited above are translatable to the deduction systems or programs they formalize and the proof techniques may directly apply to them.

Another interesting aspect is that it allows characteristics or properties of the modelled systems to be expressed as equational theorems, often automatically provable using the rewriting mechanism itself or induction techniques based on completion [41]. Note that the rewriting and the completion mechanisms also enable transformation and simplification of formal systems or programs.

Applications of rewriting-based proofs to computer security are various. Approaches using rule-based specifications have recently been proposed for detection of computer viruses [72], [73]. For several years, in our team, we have also been working in this direction. We already proposed an approach using rewriting techniques to abstract program behaviors for detecting suspicious or malicious programs [31], [32].

CASSIS Project-Team

3. Research Program

3.1. Introduction

Our main goal is to design techniques and to develop tools for the verification of (safety-critical) systems, such as programs or protocols. To this end, we develop a combination of techniques based on automated deduction for program verification, constraint resolution for test generation, and reachability analysis for the verification of infinite-state systems.

3.2. Automated Deduction

The main goal is to prove the validity of assertions obtained from program analysis. To this end, we develop techniques and automated deduction systems based on rewriting and constraint solving. The verification of recursive data structures relies on inductive reasoning or the manipulation of equations and it also exploits some form of reasoning modulo properties of selected operators (such as associativity and/or commutativity).

Rewriting, which allows us to simplify expressions and formulae, is a key ingredient for the effectiveness of many state-of-the-art automated reasoning systems. Furthermore, a well-founded rewriting relation can be also exploited to implement reasoning by induction. This observation forms the basis of our approach to inductive reasoning, with high degree of automation and the possibility to refute false conjectures.

The constraints are the key ingredient to postpone the activity of solving complex symbolic problems until it is really necessary. They also allow us to increase the expressivity of the specification language and to refine theorem-proving strategies. As an example of this, the handling of constraints for unification problems or for the orientation of equalities in the presence of interpreted operators (e.g., commutativity and/or associativity function symbols) will possibly yield shorter automated proofs.

Finally, decision procedures are being considered as a key ingredient for the successful application of automated reasoning systems to verification problems. A decision procedure is an algorithm capable of efficiently deciding whether formulae from certain theories (such as Presburger arithmetic, lists, arrays, and their combination) are valid or not. We develop techniques to build and to combine decision procedures for the domains which are relevant to verification problems. We also perform experimental evaluation of the proposed techniques by combining propositional reasoning (implemented by means of Boolean solvers, e.g., SAT solvers) and decision procedures to get solvers for the problem of Satisfiability Modulo Theories (SMT).

3.3. Synthesizing and Solving Constraints

Applying constraint logic programming technology in the validation and verification area is currently an active way of research. It usually requires the design of specific solvers to deal with the description language's vocabulary. For instance, we are interested in applying a solver for set constraintsto evaluate set-oriented formal specifications. By evaluation, we mean the encoding of the formal model into a constraint system, and the ability for the solver to verify the invariant on the current constraint graph, to propagate preconditions or guards, and to apply a substitution calculus on this graph. The constraint solver is used for animating specifications and automatically generating abstract test cases.

3.4. Rewriting-based Safety Checking

Invariant checking and strenghtening is the dual of reachability analysis, and can thus be used for verifying safety properties of infinite-state systems. In fact, many infinite-state systems are just parameterized systems which become finite state systems when parameters are instantiated. Then, the challenge is to automatically discharge the maximal number of proof obligations coming from the decomposition of the invariance conditions. For parameterized systems, we are interested in a deductive approach where states are defined by first order formulae with equality, and proof obligations are checked by SMT solvers.

PAREO Project-Team

3. Research Program

3.1. Introduction

It is a common claim that rewriting is ubiquitous in computer science and mathematical logic. And indeed the rewriting concept appears from very theoretical settings to very practical implementations. Some extreme examples are the mail system under Unix that uses rules in order to rewrite mail addresses in canonical forms and the transition rules describing the behaviors of tree automata. Rewriting is used in semantics in order to describe the meaning of programming languages [22] as well as in program transformations like, for example, re-engineering of Cobol programs [31]. It is used in order to compute, implicitly or explicitly as in Mathematica or MuPAD, but also to perform deduction when describing by inference rules a logic [18], a theorem prover [20] or a constraint solver [21]. It is of course central in systems making the notion of rule an explicit and first class object, like expert systems, programming languages based on equational logic, algebraic specifications, functional programming and transition systems.

In this context, the study of the theoretical foundations of rewriting have to be continued and effective rewrite based tools should be developed. The extensions of first-order rewriting with higher-order and higher-dimension features are hot topics and these research directions naturally encompass the study of the rewriting calculus, of polygraphs and of their interaction. The usefulness of these concepts becomes more clear when they are implemented and a considerable effort is thus put nowadays in the development of expressive and efficient rewrite based programming languages.

3.2. Rule-based Programming Languages

Programming languages are formalisms used to describe programs, applications, or software which aim to be executed on a given hardware. In principle, any Turing complete language is sufficient to describe the computations we want to perform. However, in practice the choice of the programming language is important because it helps to be effective and to improve the quality of the software. For instance, a web application is rarely developed using a Turing machine or assembly language. By choosing an adequate formalism, it becomes easier to reason about the program, to analyze, certify, transform, optimize, or compile it. The choice of the programming language also has an impact on the quality of the software. By providing high-level constructs as well as static verifications, like typing, we can have an impact on the software design, allowing more expressiveness, more modularity, and a better reuse of code. This also improves the productivity of the programmer, and contributes to reducing the presence of errors.

The quality of a programming language depends on two main factors. First, the *intrinsic design*, which describes the programming model, the data model, the features provided by the language, as well as the semantics of the constructs. The second factor is the programmer and the application which is targeted. A language is not necessarily good for a given application if the concepts of the application domain cannot be easily manipulated. Similarly, it may not be good for a given person if the constructs provided by the language are not correctly understood by the programmer.

In the *Pareo* group we target a population of programmers interested in improving the long-term maintainability and the quality of their software, as well as their efficiency in implementing complex algorithms. Our privileged domain of application is large since it concerns the development of *transformations*. This ranges from the transformation of textual or structured documents such as XML, to the analysis and the transformation of programs and models. This also includes the development of tools such as theorem provers, proof assistants, or model checkers, where the transformations of proofs and the transitions between states play a crucial role. In that context, the *expressiveness* of the programming language is important. Indeed, complex encodings into low level data structures should be avoided, in contrast to high level notions such as abstract types and transformation rules that should be provided.

It is now well established that the notions of *term* and *rewrite rule* are two universal abstractions well suited to model tree based data types and the transformations that can be done upon them. Over the last ten years we have developed a strong experience in designing and programming with rule based languages [23], [14], [12]. We have introduced and studied the notion of *strategy* [13], which is a way to control how the rules should be applied. This provides the separation which is essential to isolate the logic and to make the rules reusable in different contexts.

To improve the quality of programs, it is also essential to have a clear description of their intended behaviors. For that, the *semantics* of the programming language should be formally specified.

There is still a lot of progress to be done in these directions. In particular, rule based programming can be made even more expressive by extending the existing matching algorithms to context-matching or to new data structures such as graphs or polygraphs. New algorithms and implementation techniques have to be found to improve the efficiency and make the rule based programming approach effective on large problems. Separating the rules from the control is very important. This is done by introducing a language for describing strategies. We still have to invent new formalisms and new strategy primitives which are both expressive enough and theoretically well grounded. A challenge is to find a good strategy language we can reason about, to prove termination properties for instance.

On the static analysis side, new formalized typing algorithms are needed to properly integrate rule based programming into already existing host languages such as Java. The notion of traversal strategy merits to be better studied in order to become more flexible and still provide a guarantee that the result of a transformation is correctly typed.

3.3. Rewriting Calculus

The huge diversity of the rewriting concept is obvious and when one wants to focus on the underlying notions, it becomes quickly clear that several technical points should be settled. For example, what kind of objects are rewritten? Terms, graphs, strings, sets, multisets, others? Once we have established this, what is a rewrite rule? What is a left-hand side, a right-hand side, a condition, a context? And then, what is the effect of a rule application? This leads immediately to defining more technical concepts like variables in bound or free situations, substitutions and substitution application, matching, replacement; all notions being specific to the kind of objects that have to be rewritten. Once this is solved one has to understand the meaning of the application of a set of rules on (classes of) objects. And last but not least, depending on the intended use of rewriting, one would like to define an induced relation, or a logic, or a calculus.

In this very general picture, we have introduced a calculus whose main design concept is to make all the basic ingredients of rewriting explicit objects, in particular the notions of rule *application* and *result*. We concentrate on *term* rewriting, we introduce a very general notion of rewrite rule and we make the rule application and result explicit concepts. These are the basic ingredients of the *rewriting*- or ρ-calculus whose originality comes from the fact that terms, rules, rule application and application strategies are all treated at the object level (a rule can be applied on a rule for instance).

The λ-calculus is usually put forward as the abstract computational model underlying functional programming. However, modern functional programming languages have pattern-matching features which cannot be directly expressed in the λ-calculus. To palliate this problem, pattern-calculi [28], [25], [19] have been introduced. The rewriting calculus is also a pattern calculus that combines the expressiveness of pure functional calculi and algebraic term rewriting. This calculus is designed and used for logical and semantical purposes. It could be equipped with powerful type systems and used for expressing the semantics of rule based as well as object oriented languages. It allows one to naturally express exception handling mechanisms and elaborated rewriting strategies. It can be also extended with imperative features and cyclic data structures.

The study of the rewriting calculus turns out to be extremely successful in terms of fundamental results and of applications [16]. Different instances of this calculus together with their corresponding type systems have been proposed and studied. The expressive power of this calculus was illustrated by comparing it with similar

formalisms and in particular by giving a typed encoding of standard strategies used in first-order rewriting and classical rewrite based languages like *ELAN* and *Tom*.

VEGAS Project-Team (section vide)

VERIDIS Project-Team

3. Research Program

3.1. Automated and Interactive Theorem Proving

The VeriDis team unites experts in techniques and tools for interactive and automated verification, and specialists in methods and formalisms designed for developing concurrent and distributed systems and algorithms that are firmly grounded on precise mathematical and semantical abstractions. Our common objective is to advance the state of the art in interactive and automatic deduction techniques, and their combinations, resulting in powerful tools for the computer-aided verification of distributed systems and protocols. Our techniques and tools support sound methods for the development of trustworthy distributed systems that scale to algorithms relevant for practical applications.

VeriDis members from Saarbrücken are developing SPASS [10], one of the leading automated theorem provers for first-order logic based on the superposition calculus [46]. Recent extensions to the system include the integration of dedicated reasoning procedures for specific theories, such as linear arithmetic [56], [45], that are ubiquitous in the verification of systems and algorithms. The group also studies general frameworks for the combination of theories such as the locality principle [57] and automated reasoning mechanisms these induce. Finally, members of the group design effective quantifier elimination methods and decision procedures for algebraic theories, supported by their efficient implementation in the Redlog system [4].

In a complementary approach to automated deduction, VeriDis members from Nancy develop veriT [1], an SMT (Satisfiability Modulo Theories [48]) solver that combines decision procedures for different fragments of first-order logic and that integrates an automatic theorem prover for full first-order logic. The veriT solver is designed to produce detailed proofs; this makes it particularly suitable as a component of a robust cooperation of deduction tools.

We rely on interactive theorem provers for reasoning about specifications at a high level of abstraction. Members of VeriDis have ample experience in the specification and subsequent machine-assisted, interactive verification of algorithms. In particular, we participate in a project at the joint MSR-Inria Centre in Saclay on the development of methods and tools for the formal proof of TLA$^+$ [52] specifications. Our prover relies on a declarative proof language, and we contribute several automatic backends [3].

3.2. Formal Methods for Developing Algorithms and Systems

Powerful theorem provers are not a panacea for system verification: they support sound methodologies for modeling and verifying systems. In this respect, members of VeriDis have gained expertise and recognition in making contributions to formal methods for concurrent and distributed algorithms and systems [2], [9], and in applying them to concrete use cases. In particular, the concept of *refinement* [44], [47], [54] in state-based modeling formalisms is central to our approach. Its basic idea is to present an algorithm or implementation through a series of models, starting from a high-level description that precisely states the problem, and gradually adding details in intermediate models. An important goal in designing such methods is to establish precise proof obligations that can be discharged to a high degree by automatic tools. This requires taking into account specific characteristics of certain classes of systems and tailoring the model to concrete computational models. Our research in this area is supported by carrying out case studies for academic and industrial developments. This activity benefits from and influences the development of our proof tools.

Our vision for the integration of our expertise can be resumed as follows. Based on our experience and related work on specification languages, logical frameworks, and automatic theorem proving tools, we develop an approach that is suited for specification, interactive theorem proving, and for eventual automated analysis and verification, possibly through appropriate translation methods. While specifications are developed by users inside our framework, they are analyzed for errors by our SMT based verification tools. Eventually, properties are proved by a combination of interactive and automatic theorem proving tools.

Today, the formal verification of a new algorithm is typically the subject of a PhD thesis, if it is addressed at all. This situation is not sustainable given the move towards more and more parallelism in mainstream systems: algorithm developers and system designers must be able to productively use verification tools for validating their algorithms and implementations. On a high level, the goal of VeriDis is to make formal verification standard practice for the development of distributed algorithms and systems, just as symbolic model checking has become commonplace in the development of embedded systems and as security analysis for cryptographic protocols is becoming standard practice today. Although the fundamental problems in distributed programming, such as mutual exclusion, leader election, group membership or consensus, are well-known, they pose new challenges in the context of modern system paradigms, including ad-hoc and overlay networks or peer-to-peer systems, and they must be integrated for concrete applications.

CORIDA Team

3. Research Program

3.1. Analysis and control of fluids and of fluid-structure interactions

Participants: Thomas Chambrion, Antoine Henrot, Alexandre Munnier, Lionel Rosier, Jean-François Scheid, Takéo Takahashi, Marius Tucsnak, Jean-Claude Vivalda.

The problems we consider are modeled by the Navier-Stokes, Euler or Korteweg de Vries equations (for the fluid) coupled to the equations governing the motion of the solids. One of the main difficulties of this problem comes from the fact that the domain occupied by the fluid is one of the unknowns of the problem. We have thus to tackle a *free boundary problem*.

The control of fluid flows is a major challenge in many applications: aeronautics, pollution issues, regulation of irrigation channels or of the flow in pipelines, etc. All these problems cannot be easily reduced to finite dimensional models so a methodology of analysis and control based on PDE's is an essential issue. In a first approximation the motion of fluid and of the solids can be decoupled. The most used models for an incompressible fluid are given by the Navier-Stokes or by the Euler equations.

The optimal open loop control approach of these models has been developed from both the theoretical and numerical points of view. Controllability issues for the equations modeling the fluid motion are by now well understood (see, for instance, Imanuvilov [52] and the references therein). The feedback control of fluid motion has also been recently investigated by several research teams (see, for instance Barbu [47] and references therein) but this field still contains an important number of open problems (in particular those concerning observers and implementation issues). One of our aims is to develop efficient tools for computing feedback laws for the control of fluid systems.

In real applications the fluid is often surrounded by or it surrounds an elastic structure. In the above situation one has to study fluid-structure interactions. This subject has been intensively studied during the last years, in particular for its applications in noise reduction problems, in lubrication issues or in aeronautics. In this kind of problems, a PDE's system modeling the fluid in a cavity (Laplace equation, wave equation, Stokes, Navier-Stokes or Euler systems) is coupled to the equations modeling the motion of a part of the boundary. The difficulties of this problem are due to several reasons such as the strong nonlinear coupling and the existence of a free boundary. This partially explains the fact that applied mathematicians have only recently tackled these problems from either the numerical or theoretical point of view. One of the main results obtained in our project concerns the global existence of weak solutions in the case of a two-dimensional Navier–Stokes fluid [59]. Another important result gives the existence and the uniqueness of strong solutions for two or three-dimensional Navier–Stokes fluid [61]. In that case, the solution exists as long as there is no contact between rigid bodies, and for small data in the three-dimensional case.

3.2. Frequency domain methods for the analysis and control of systems governed by PDE's

Participants: Xavier Antoine, Bruno Pinçon, Karim Ramdani.

We use frequency tools to analyze different types of problems. The first one concerns the control, the optimal control and the stabilization of systems governed by PDE's, and their numerical approximations. The second one concerns time-reversal phenomena, while the last one deals with numerical approximation of high-frequency scattering problems.

3.2.1. Control and stabilization for skew-adjoint systems

The first area concerns theoretical and numerical aspects in the control of a class of PDE's. More precisely, in a semigroup setting, the systems we consider have a skew-adjoint generator. Classical examples are the wave, the Bernoulli-Euler or the Schrödinger equations. Our approach is based on an original characterization of exact controllability of second order conservative systems proposed by K. Liu [56]. This characterization can be related to the Hautus criterion in the theory of finite dimensional systems (cf. [51]). It provides for time-dependent problems exact controllability criteria *that do not depend on time, but depend on the frequency variable* conjugated to time. Studying the controllability of a given system amounts then to establishing uniform (with respect to frequency) estimates. In other words, the problem of exact controllability for the wave equation, for instance, comes down to a high-frequency analysis for the Helmholtz operator. This frequency approach has been proposed first by K. Liu for bounded control operators (corresponding to internal control problems), and has been recently extended to the case of unbounded control operators (and thus including boundary control problems) by L. Miller [57]. Using the result of Miller, K. Ramdani, T. Takahashi, M. Tucsnak have obtained in [5] a new spectral formulation of the criterion of Liu [56], which is valid for boundary control problems. This frequency test can be seen as an observability condition for packets of eigenvectors of the operator. This frequency test has been successfully applied in [5] to study the exact controllability of the Schrödinger equation, the plate equation and the wave equation in a square. Let us emphasize here that one further important advantage of this frequency approach lies in the fact that it can also be used for the analysis of space semi-discretized control problems (by finite element or finite differences). The estimates to be proved must then be uniform with respect to *both the frequency and the mesh size*.

In the case of finite dimensional systems one of the main applications of frequency domain methods consists in designing robust controllers, in particular of H^∞ type. Obtaining the similar tools for systems governed by PDE's is one of the major challenges in the theory of infinite dimensional systems. The first difficulty which has to be tackled is that, even for very simple PDE systems, no method giving the parametrisation of all stabilizing controllers is available. One of the possible remedies consists in considering known families of stabilizing feedback laws depending on several parameters and in optimizing the H^∞ norm of an appropriate transfer function with respect to this parameters. Such families of feedback laws yielding computationally tractable optimization problems are now available for systems governed by PDE's in one space dimension.

3.2.2. Time-reversal

The second area in which we make use of frequency tools is the analysis of time-reversal for harmonic acoustic waves. This phenomenon described in Fink [49] is a direct consequence of the reversibility of the wave equation in a non dissipative medium. It can be used to **focus an acoustic wave** on a target through a complex and/or unknown medium. To achieve this, the procedure followed is quite simple. First, time-reversal mirrors are used to generate an incident wave that propagates through the medium. Then, the mirrors measure the acoustic field diffracted by the targets, time-reverse it and back-propagate it in the medium. Iterating the scheme, we observe that the incident wave emitted by the mirrors focuses on the scatterers. An alternative and more original focusing technique is based on the so-called D.O.R.T. method [50]. According to this experimental method, the eigenelements of the time-reversal operator contain important information on the propagation medium and on the scatterers contained in it. More precisely, the number of nonzero eigenvalues is exactly the number of scatterers, while each eigenvector corresponds to an incident wave that selectively focuses on each scatterer.

Time-reversal has many applications covering a wide range of fields, among which we can cite *medicine* (kidney stones destruction or medical imaging), *sub-marine communication* and *non destructive testing*. Let us emphasize that in the case of time-harmonic acoustic waves, time-reversal is equivalent to phase conjugation and involves the Helmholtz operator.

In [2], we proposed the first far field model of time reversal in the time-harmonic case.

3.2.3. Numerical approximation of high-frequency scattering problems

This subject deals mainly with the numerical solution of the Helmholtz or Maxwell equations for open region scattering problems. This kind of situation can be met e.g. in radar systems in electromagnetism or in acoustics for the detection of underwater objects like submarines.

Two particular difficulties are considered in this situation

- the wavelength of the incident signal is small compared to the characteristic size of the scatterer,
- the problem is set in an unbounded domain.

These two problematics limit the application range of most common numerical techniques. The aim of this part is to develop new numerical simulation techniques based on microlocal analysis for modeling the propagation of rays. The importance of microlocal techniques in this situation is that it makes possible a local analysis both in the spatial and frequency domain. Therefore, it can be seen as a kind of asymptotic theory of rays which can be combined with numerical approximation techniques like boundary element methods. The resulting method is called the On-Surface Radiation Condition method.

3.3. Observability, controllability and stabilization in the time domain

Participants: Fatiha Alabau-Boussouira, Xavier Antoine, Thomas Chambrion, Antoine Henrot, Karim Ramdani, Marius Tucsnak, Jean-Claude Vivalda.

Controllability and observability have been set at the center of control theory by the work of R. Kalman in the 1960's and soon they have been generalized to the infinite-dimensional context. The main early contributors have been D.L. Russell, H. Fattorini, T. Seidman, R. Triggiani, W. Littman and J.-L. Lions. The latter gave the field an enormous impact with his book [54], which is still a main source of inspiration for many researchers. Unlike in classical control theory, for infinite-dimensional systems there are many different (and not equivalent) concepts of controllability and observability. The strongest concepts are called exact controllability and exact observability, respectively. In the case of linear systems exact controllability is important because it guarantees stabilizability and the existence of a linear quadratic optimal control. Dually, exact observability guarantees the existence of an exponentially converging state estimator and the existence of a linear quadratic optimal filter. An important feature of infinite dimensional systems is that, unlike in the finite dimensional case, the conditions for exact observability are no longer independent of time. More precisely, for simple systems like a string equation, we have exact observability only for times which are large enough. For systems governed by other PDE's (like dispersive equations) the exact observability in arbitrarily small time has been only recently established by using new frequency domain techniques. A natural question is to estimate the energy required to drive a system in the desired final state when the control time goes to zero. This is a challenging theoretical issue which is critical for perturbation and approximation problems. In the finite dimensional case this issue has been first investigated in Seidman [60]. In the case of systems governed by linear PDE's some similar estimates have been obtained only very recently (see, for instance Miller [57]). One of the open problems of this field is to give sharp estimates of the observability constants when the control time goes to zero.

Even in the finite-dimensional case, despite the fact that the linear theory is well established, many challenging questions are still open, concerning in particular nonlinear control systems.

In some cases it is appropriate to regard external perturbations as unknown inputs; for these systems the synthesis of observers is a challenging issue, since one cannot take into account the term containing the unknown input into the equations of the observer. While the theory of observability for linear systems with unknown inputs is well established, this is far from being the case in the nonlinear case. A related active field of research is the uniform stabilization of systems with time-varying parameters. The goal in this case is to stabilize a control system with a control strategy independent of some signals appearing in the dynamics, i.e., to stabilize simultaneously a family of time-dependent control systems and to characterize families of control systems that can be simultaneously stabilized.

One of the basic questions in finite- and infinite-dimensional control theory is that of motion planning, i.e., the explicit design of a control law capable of driving a system from an initial state to a prescribed final one. Several techniques, whose suitability depends strongly on the application which is considered, have been and are being developed to tackle such a problem, as for instance the continuation method, flatness, tracking or optimal control. Preliminary to any question regarding motion planning or optimal control is the issue of controllability, which is not, in the general nonlinear case, solved by the verification of a simple algebraic criterion. A further motivation to study nonlinear controllability criteria is given by the fact that techniques developed in the domain of (finite-dimensional) geometric control theory have been recently applied successfully to study the controllability of infinite-dimensional control systems, namely the Navier–Stokes equations (see Agrachev and Sarychev [46]).

3.4. Implementation

This is a transverse research axis since all the research directions presented above have to be validated by giving control algorithms which are aimed to be implemented in real control systems. We stress below some of the main points which are common (from the implementation point of view) to the application of the different methods described in the previous sections.

For many infinite dimensional systems the use of co-located actuators and sensors and of simple proportional feed-back laws gives satisfying results. However, for a large class of systems of interest it is not clear that these feedbacks are efficient, or the use of co-located actuators and sensors is not possible. This is why a more general approach for the design of the feedbacks has to be considered. Among the techniques in finite dimensional systems theory those based on the solutions of infinite dimensional Riccati equation seem the most appropriate for a generalization to infinite dimensional systems. The classical approach is to approximate an LQR problem for a given infinite dimensional system by finite dimensional LQR problems. As it has been already pointed out in the literature this approach should be carefully analyzed since, even for some very simple examples, the sequence of feedbacks operators solving the finite dimensional LQR is not convergent. Roughly speaking this means that by refining the mesh we obtain a closed loop system which is not exponentially stable (even if the corresponding infinite dimensional system is theoretically stabilized). In order to overcome this difficulty, several methods have been proposed in the literature : filtering of high frequencies, multigrid methods or the introduction of a numerical viscosity term. We intend to first apply the numerical viscosity method introduced in Tcheougoue Tebou – Zuazua [62], for optimal and robust control problems.

TOSCA Project-Team

3. Research Program

3.1. Research Program

Most often physicists, economists, biologists, engineers need a stochastic model because they cannot describe the physical, economical, biological, etc., experiment under consideration with deterministic systems, either because of its complexity and/or its dimension or because precise measurements are impossible. Then they abandon trying to get the exact description of the state of the system at future times given its initial conditions, and try instead to get a statistical description of the evolution of the system. For example, they desire to compute occurrence probabilities for critical events such as the overstepping of a given thresholds by financial losses or neuronal electrical potentials, or to compute the mean value of the time of occurrence of interesting events such as the fragmentation to a very small size of a large proportion of a given population of particles. By nature such problems lead to complex modelling issues: one has to choose appropriate stochastic models, which require a thorough knowledge of their qualitative properties, and then one has to calibrate them, which requires specific statistical methods to face the lack of data or the inaccuracy of these data. In addition, having chosen a family of models and computed the desired statistics, one has to evaluate the sensitivity of the results to the unavoidable model specifications. The TOSCA team, in collaboration with specialists of the relevant fields, develops theoretical studies of stochastic models, calibration procedures, and sensitivity analysis methods.

In view of the complexity of the experiments, and thus of the stochastic models, one cannot expect to use closed form solutions of simple equations in order to compute the desired statistics. Often one even has no other representation than the probabilistic definition (e.g., this is the case when one is interested in the quantiles of the probability law of the possible losses of financial portfolios). Consequently the practitioners need Monte Carlo methods combined with simulations of stochastic models. As the models cannot be simulated exactly, they also need approximation methods which can be efficiently used on computers. The TOSCA team develops mathematical studies and numerical experiments in order to determine the global accuracy and the global efficiency of such algorithms.

The simulation of stochastic processes is not motivated by stochastic models only. The stochastic differential calculus allows one to represent solutions of certain deterministic partial differential equations in terms of probability distributions of functionals of appropriate stochastic processes. For example, elliptic and parabolic linear equations are related to classical stochastic differential equations, whereas nonlinear equations such as the Burgers and the Navier–Stokes equations are related to McKean stochastic differential equations describing the asymptotic behavior of stochastic particle systems. In view of such probabilistic representations one can get numerical approximations by using discretization methods of the stochastic differential systems under consideration. These methods may be more efficient than deterministic methods when the space dimension of the PDE is large or when the viscosity is small. The TOSCA team develops new probabilistic representations in order to propose probabilistic numerical methods for equations such as conservation law equations, kinetic equations, and nonlinear Fokker–Planck equations.

BIGS Project-Team

3. Research Program

3.1. Online data analysis

Participants: J.-M. Monnez, P. Vallois. Generally speaking, there exists an overwhelming amount of articles dealing with the analysis of high dimensional data. Indeed, this is one of the major challenges in statistics today, motivated by internet or biostatistics applications. Within this global picture, the problem of classification or dimension reduction of online data can be traced back at least to a seminal paper by Mac Queen [53], in which the k-means algorithm is introduced. This popular algorithm, constructed for classification purposes, consists in a stepwise updating of the centers of some classes according to a stream of data entering into the system. The literature on the topic has been growing then rapidly since the beginning of the 90's.

Our point of view on the topic relies on the so-called *French data analysis school*, and more specifically on Factorial Analysis tools. In this context, it was then rapidly seen that stochastic approximation was an essential tool (see Lebart's paper [50]), which allows one to approximate eigenvectors in a stepwise manner. A systematic study of Principal Component and Factorial Analysis has then been lead by Monnez in the series of papers [56], [54], [55], in which many aspects of convergences of online processes are analyzed thanks to the stochastic approximation techniques.

3.2. Local regression techniques

Participants: S. Ferrigno, A. Muller-Gueudin. In the context where a response variable Y is to be related to a set of regressors X, one of the general goals of Statistics is to provide the end user with a model which turns out to be useful in predicting Y for various values of X. Except for the simplest situations, the determination of a good model involves many steps. For example, for the task of predicting the value of Y as a function of the covariate X, statisticians have elaborated models such as the regression model with random regressors:

$$Y = g(X, \theta) + \sigma(X)\epsilon.$$

Many assumptions must be made to reach it as a possible model. Some require much thinking, as for example, those related to the functional form of $g(\cdot, \theta)$. Some are made more casually, as often those related to the functional form of $\sigma(\cdot)$ or those concerning the distribution of the random error term ϵ. Finally, some assumptions are made for commodity. Thus the need for methods that can assess if a model is concordant with the data it is supposed to adjust. The methods fall under the banner of goodness of fit tests. Most existing tests are *directional*, in the sense that they can detect departures from only one or a few aspects of a null model. For example, many tests have been proposed in the literature to assess the validity of an entertained structural part $g(\cdot, \theta)$. Some authors have also proposed tests about the variance term $\sigma(\cdot)$ (cf. [51]). Procedures testing the normality of the ϵ_i are given, but for other assumptions much less work has been done. Therefore the need of a global test which can evaluate the validity of a global structure emerges quite naturally.

With these preliminaries in mind, let us observe that one quantity which embodies all the information about the joint behavior of (X, Y) is the cumulative conditional distribution function, defined by

$$F(y|x) = P(Y \leq y|X = x).$$

The (nonparametric) estimation of this function is thus of primary importance. To this aim, notice that modern estimators are usually based on the local polynomial approach, which has been recognized as superior to classical estimates based on the Nadaraya-Watson approach, and are as good as the recent versions based on spline and other methods. In some recent works [41], [42], we address the following questions:

- Construction of a global test by means of Cramér-von Mises statistic.
- Optimal bandwidth of the kernel used for approximation purposes.

We also obtain sharp estimates on the conditional distribution function in [43].

3.3. Stochastic modeling for complex and biological systems

Participants: R. Azaïs, T. Bastogne, C. Lacaux, A. Muller-Gueudin, S. Tindel, P. Vallois, S. Wantz-Mézières

In most biological contexts, mathematics turn out to be useful in producing accurate models with dual objectives: they should be simple enough and meaningful for the biologist on the one hand, and they should provide some insight on the biological phenomenon at stake on the other hand. We have focused on this kind of issue in various contexts that we shall summarize below.

Photodynamic Therapy: Photodynamic therapy induces a huge demand of interconnected mathematical systems, among which we have studied recently the following ones:

• The tumor growth model is of crucial importance in order to understand the behavior of the whole therapy. We have considered the tumor growth as a stochastic equation, for which we have handled the problem uncertainties on the measure times [27] as well as mixed effects for parameter estimation.

• Another important aspect to quantify for photodynamic therapy calibration is the response to radiotherapy treatments. There are several valid mathematical ways to describe this process, among which we distinguish the so-called hit model. This model assumes that whenever a group of sensitive targets (chromosomes, membrane) in the cell are reached by a sufficient number of radiations, then the cell is inactivated and dies. We have elaborated on this scheme in order to take into account two additional facts: (i) The reduction of the cell situation to a two-state model might be an oversimplification. (ii) Several doses of radiations are inoculated as time passes. These observations have lead us to introduce a new model based on multi-state Markov chains arguments (Keinj & al, 2012), in which cell proliferation can be incorporated.

Bacteriophage therapy: Let us mention a starting collaboration between BIGS and the Genetics and Microbiology department at the Universitat Autònoma de Barcelona, on the modeling of bacteriophage therapies. The main objective here is to describe how a certain family of benign viruses is able to weaken a bacterium induced disease, which naturally leads to the introduction of a noisy predator-prey system of equations. It should be mentioned that some similar problems have been treated (in a rather informal way, invoking a linearization procedure) by Carletti in [34]. These tools cannot be applied directly to our system, and our methods are based on concentration and large deviations techniques (on which we already had an expertise [57], [60]) in order to combine convergence to equilibrium for the deterministic system and deviations of the stochastic system. Notice that A. Muller-Gueudin is also working with A. Debussche and O. Radulescu on a related topic [37], namely the convergence of a model of cellular biochemical reactions.

Gaussian signals: Nature provides us with many examples of systems such that the observed signal has a given Hölder regularity, which does not correspond to the one we might expect from a system driven by ordinary Brownian motion. This situation is commonly handled by noisy equations driven by Gaussian processes such as fractional Brownian motion or (in higher dimensions of the parameter) fractional fields.

The basic aspects of differential equations driven by a fractional Brownian motion (fBm) and other Gaussian processes are now well understood, mainly thanks to the so-called *rough paths* tools [52], but also invoking the Russo-Vallois integration techniques [59]. The specific issue of Volterra equations driven by fBm, which is central for the subdiffusion within proteins problem, is addressed in [38].

Fractional fields are very often used to model irregular phenomena which exhibit a scale invariance property, fractional Brownian motion being the historical fractional model. Nevertheless, its isotropy property is a serious drawback for instance in hydrology or in medecine (see [33]). Moreover, the fractional Brownian motion cannot be used to model some phenomena for which the regularity varies with time. Hence, many generalizations (Gaussian or not) of this model have been recently proposed, see for instance [28] for some Gaussian locally self-similar fields, [46] for some non-Gaussian models, [31] for anisotropic models.

Our team has thus contributed [36], [47], [46], [48], [58] and still contributes [30], [32], [31], [49], [44] to this theoretical study: Hölder continuity, fractal dimensions, existence and uniqueness results for differential equations, study of the laws to quote a few examples. As we shall see below, this line of investigation also has some impact in terms of applications: we shall discuss how we plan to apply our results to osteoporosis on the one hand and to fluctuations within protein molecules on the other hand.

3.4. Parameter identifiability and estimation

Participants: R. Azaïs, T. Bastogne, S. Tindel, P. Vallois, S. Wantz-Mézières

When one desires to confront theoretical probabilistic models with real data, statistical tools are obviously crucial. We have focused on two of them: parameter identifiability and parameter estimation.

Parameter identifiability [62] deals with the possibility to give a unique value to each parameter of a mathematical model structure in inverse problems. There are many methods for testing models for identifiability: Laplace transform, similarity transform, Taylor series, local state isomorphism or elimination theory. Most of the current approaches are devoted to *a priori* identifiability and are based on algebraic techniques. We are particularly concerned with *a posteriori* identifiability, *i.e.*, after experiments or in a constrained experimental framework and the link with experimental design techniques. Our approach is based on statistical techniques through the use of variance-based methods. These techniques are strongly connected with global sensitivity approaches and Monte Carlo methods.

The parameter estimation for a family of probability laws has a very long story in statistics, and we refer to [29] for an elegant overview of the topic. Moving to the references more closely related to our specific projects, let us recall first that the mathematical description of photodynamic therapy can be split up into three parametric models : the uptake model (pharmacokinetics of the photosensitizing drug into cancer cells), the photoreaction model and the tumor growth model. Several papers have been reported for the application of system identification techniques to pharmacokinetics modeling problems. But two issues were ignored in these previous works: presence of timing noise and identification from longitudinal data. In [27], we have proposed a bounded-error estimation algorithm based on interval analysis to solve the parameter estimation problem while taking into consideration uncertainty on observation time instants. Statistical inference from longitudinal data based on mixed effects models can be performed by the *Monolix* software (http://www.lixoft.eu/products/monolix/product-monolix-overview/) developed by the Monolix group chaired by Marc Lavielle and France Mentré, and supported by Inria. In the recent past, we have used this tool for tumor growth modeling. According to what we know so far, no parameter estimation study has been reported about the photoreaction model in photodynamic therapy. A photoreaction model, composed of six stochastic differential equations, is proposed in [39]. The main open problem is to access to data. We currently build on an experimental platform which aims at overcoming this technical issue. Moreover, an identifiability study coupled to a global sensitivity analysis of the photoreaction model are currently in progress. Tumor growth is generally described by population dynamics models or by cell cycle models. Faced with this wide variety of descriptions, one of the main open problems is to identify the suitable model structure. As mentioned above, we currently investigate alternative representations based on branching processes and Markov chains, with a model selection procedure in mind.

A few words should be said about the existing literature on statistical inference for diffusion or related processes, a topic which will be at the heart of three of our projects (namely photodynamic and bacteriophage therapies, as well as fluctuations within molecules). The monograph [45] is a good reference on the basic estimation techniques for diffusion processes. The problem of estimating diffusions observed at discrete times,

of crucial importance for applications, has been addressed mainly since the mid 90s. The maximum likelihood techniques, which are also classical for parameter estimation, are well represented by the contributions [40].

Some attention has been paid recently to the estimation of the coefficients of fractional or multifractional Brownian motion according to a set of observations. Let us quote for instance the nice surveys [26], [35]. On the other hand, the inference problem for diffusions driven by a fractional Brownian motion is still in its infancy. A good reference on the question is [61], dealing with some very particular families of equations, which do not cover the cases of interest for us.

MASAIE Project-Team

3. Research Program

3.1. Description

Our conceptual framework is that of Control Theory: the system is described by state variables with inputs (actions on the system) and outputs (the available measurements). Our system is either an epidemiological or immunological system or a harvested fish population. The control theory approach begins with the mathematical modeling of the system. When a “satisfying" model is obtained, this model is studied to understand the system. By “satisfying", an ambiguous word, we mean validation of the model. This depends on the objectives of the design of the model: explicative model, predictive model, comprehension model, checking hypotheses model. Moreover the process of modeling is not sequential. During elaboration of the model, a mathematical analysis is often done in parallel to describe the behavior of the proposed model. By behavior we intend not only asymptotic behavior but also such properties as observability, identifiability, robustness...

3.2. Structure and modeling

Problems in epidemiology, immunology and virology can be expressed as standard problems in control theory. But interesting new questions do arise. The control theory paradigm, input-output systems built out of simpler components that are interconnected, appears naturally in this context. Decomposing the system into several sub-systems, each of which endowed with certain qualitative properties, allows the behavior of the complete system to be deduced from the behavior of its parts. This paradigm, the toolbox of feedback interconnection of systems, has been used in the so-called theory of large-scale dynamic systems in control theory [23]. Reasons for decomposing are multiple. One reason is conceptual. For example connection of the immune system and the parasitic systems is a natural biological decomposition. Others reasons are for the sake of reducing algorithmic complexities or introducing intended behavior.... In this case subsystems may not have biological interpretation. For example a chain of compartments can be introduced to simulate a continuous delay [19], [20]. Analysis of the structure of epidemiological and immunological systems is vital because of the paucity of data and the dependence of behavior on biological hypotheses. The issue is to identify those parts of models that have most effects on dynamics. The concepts and techniques of interconnection of systems (large-scale systems) is useful in this regard.

In mathematical modeling in epidemiology and immunology, as in most other areas of mathematical modeling, there is always a trade-off between simple models, that omit details and are designed to highlight general qualitative behavior, and detailed models, usually designed for specific situations, including short-terms quantitative predictions. Detailed models are generally difficult to study analytically and hence their usefulness for theoretical purposes is limited, although their strategic value may be high. Simple models can be considered as building blocks of models that include detailed structure. The control theory tools of large-scale systems and interconnections of systems are a mean to conciliate the two approaches, simple models versus detailed systems.

3.3. Dynamic Problems

Many dynamical questions addressed by systems theory are precisely what biologist are asking. One fundamental problem is the problem of equilibria and their stability. To quote J.A. Jacquez

> A major project in deterministic modeling of heterogeneous populations is to find conditions for local and global stability and to work out the relations among these stability conditions, the threshold for epidemic take-off, and endemicity, and the basic reproduction number.

The basic reproduction number $\mathcal{R}_0$ is an important quantity in the study in epidemics. It is defined as the average number of secondary infections produced when one infected individual is introduced into a host population where everyone is susceptible. The basic reproduction number $\mathcal{R}_0$ is often considered as the threshold quantity that determines when an infection can invade and persist in a new host population. To the problem of stability is related the problem of robustness, a concept from control theory. In other words how near is the system to an unstable one? Robustness is also in relation with uncertainty of the systems. This is a key point in epidemiological and immunological systems, since there are many sources of uncertainties in these models. The model is uncertain (parameters, functions, structure in some cases), the inputs also are uncertain and the outputs highly variable. That robustness is a fundamental issue and can be seen by means of an example: if policies in public health are to be taken from modeling, they must be based on robust reasons!

3.4. Observers

The concept of observer originates in control theory. This is particularly pertinent for epidemiological systems. To an input-output system, is associated the problem of reconstruction of the state. Indeed for a given system, not all the states are known or measured, this is particularly true for biological systems. This fact is due to a lot of reasons: this is not feasible without destroying the system, this is too expensive, there are no available sensors, measures are too noisy...The problem of knowledge of the state at present time is then posed. An observer is another system, whose inputs are the inputs and the outputs of the original system and whose output gives an estimation of the state of the original system at present time. Usually the estimation is required to be exponential. In other words an observer, using the signal information of the original system, reconstructs dynamically the state. More precisely, consider an input-output nonlinear system described by

$$\begin{cases} \dot{x} = f(x,u) \\ y = h(x), \end{cases} \tag{1}$$

where $x(t) \in \mathbb{R}^n$ is the state of the system at time t, $u(t) \in U \subset \mathbb{R}^m$ is the input and $y(t) \in \mathbb{R}^q$ is the measurable output of the system.
An observer for the system (1) is a dynamical system

$$\dot{\widehat{x}}(t) = g(\widehat{x}(t), y(t), u(t)), \tag{2}$$

where the map g has to be constructed such that: the solutions $x(t)$ and $\widehat{x}(t)$ of (1) and (2) satisfy for any initial conditions $x(0)$ and $\widehat{x}(0)$

$$\|x(t) - \widehat{x}(t)\| \leq c \, \|x(0) - \widehat{x}(0)\| \, e^{-at} \ , \ \forall t > 0$$

or at least $\|x(t) - \widehat{x}(t)\|$ converges to zero as time goes to infinity.
The problem of observers is completely solved for linear time-invariant systems (LTI). This is a difficult problem for nonlinear systems and is currently an active subject of research. The problem of observation and observers (software sensors) is central in nonlinear control theory. Considerable progress has been made in the last decade, especially by the “French school", which has given important contributions (J.P. Gauthier, H. Hammouri, E. Busvelle, M. Fliess, L. Praly, J.L. Gouzé, O. Bernard, G. Sallet) and is still very active in this area. Now the problem is to identify relevant classes of systems for which reasonable and computable observers can be designed. The concept of observer has been ignored by the modeler community in epidemiology, immunology and virology. To our knowledge one of the first use of an observer in virology was the work of Velasco-Hernandez J., Garcia J. and Kirschner D. [24] in modeling the chemotherapy of HIV, but this observer, based on classical linearization method, is a local observer and does not allow to deal with the nonlinearities.

3.5. Delays

Another crucial issue for biological systems is the question of delays. Delays, in control theory, are traditionally discrete (more exactly, the delays are lags) whereas in biology they usually are continuous and distributed. For example, the entry of a parasite into a cell initiates a cascade of events that ultimately leads to the production of new parasites. Even in a homogeneous population of cells, it is unreasonable to expect that the time to complete all these processes is the same for every cell. If we furthermore consider differences in cell activation state, metabolism, position in the cell cycle, pre-existing stores of nucleotides and other precursors needed for the reproduction of parasites, along with genetic variations in the parasite population, such variations in infection delay times become a near certainty. The rationale for studying continuous delays is supported by such considerations. In the literature on dynamical systems, we find a wealth of theorems dealing with delay differential equations. However they are difficult to apply. Control theory approache (interconnections of systems) is a mean to study the influence of continuous delays on the stability of such systems. We have obtained some results in this direction [6].

NEUROSYS Team

3. Research Program

3.1. Main Objectives

The main challenge in computational neuroscience is the high complexity of neural systems. The brain is a complex system and exhibits a hierarchy of interacting subunits. On a specific hierarchical level, such subunits evolve on a certain temporal and spatial scale. The interactions of small units on a low hierarchical level build up larger units on a higher hierarchical level evolving on a slower time scale and larger spatial scale. By virtue of the different dynamics on each hierarchical level, until today the corresponding mathematical models and data analysis techniques on each level are still distinct. Only few analysis and modeling frameworks are known which link successfully at least two hierarchical levels.
Once having extracted models for different description levels, typically they are applied to obtain simulated activity which is supposed to reconstruct features in experimental data. Although this approach appears straight-forward, it implies various difficulties. Usually the models involve a large set of unknown parameters which determine the dynamical properties of the models. To optimally reconstruct experimental features, it is necessary to formulate an inverse problem to extract optimally such model parameters from the experimental data. Typically this is a rather difficult problem due to the low signal-to-noise ratio in experimental brain signals. Moreover, the identification of signal features to be reconstructed by the model is not obvious in most applications. Consequently an extended analysis of the experimental data is necessary to identify the interesting data features. It is important to combine such a data analysis step with the parameter extraction procedure to achieve optimal results. Such a procedure depends on the properties of the experimental data and hence has to be developped for each application separately.

3.2. Challenges

Eventually the implementation of the models and analysis techniques achieved promises to be able to construct novel data monitor. This construction involves additional challenges and stipulates the contact to realistic environments. By virtue of the specific applications of the research, the close contact to hospitals and medical enterprises shall be established in a longer term in order to (i) gain deeper insight into the specific application of the devices and (ii) build specific devices in accordance to the actual need. Collaborations with local and national hospitals and the pharmaceutical industry already exist.

3.3. Research Directions

- From the microscopic to the mesoscopic scale:
  One research direction focusses on the *relation of single neuron activity* on the microscopic scale *to the activity of neuronal populations*. To this end, the team investigates the stochastic dynamics of single neurons subject to external random inputs and involving random microscopic properties, such as random synaptic strengths and probability distributions of spatial locations of membrane ion channels. Such an approach yields a stochastic model of single neurons and allows the derivation of a stochastic neural population model.
  This bridge between the microscopic and mesoscopic scale may be performed via two pathways. The analytical and numerical treatment of the microscopic model may be called a *bottom-up approach*, since it leads to a population activity model based on microscopic activity. This approach allows to compare theoretical neural population activity to experimentally obtained population activity. The *top-down approach* aims at extracting signal features from experimental data gained from neural populations which give insight into the dynamics of neural populations and the underlying microscopic activity. The work on both approaches represents a well-balanced investigation of the neural system based on the systems properties.

- From the mesoscopic to the macroscopic scale:
  The other research direction aims to link neural population dynamics to macroscopic activity and behaviour or, more generally, to phenomenological features. This link is more indirect but a very powerful approach to understand the brain, e.g., in the context of medical applications. Since real neural systems, such as in mammals, exhibit an interconnected network of neural populations, the team studies analytically and numerically the network dynamics of neural populations to gain deeper insight into possible phenomena, such as traveling waves or enhancement and diminution of certain neural rhythms. Electroencephalography (EEG) is a wonderful brain imaging technique to study the overall brain activity in real time noninvasively. However it is necessary to develop robust techniques based on stable features by investigating the time and frequency domains of brain signals. Two types of information are typically used in EEG signals: (i) transient events such as evoked potentials, spindles and K-complexes and (ii) the power in specific frequency bands.

SHACRA Project-Team

3. Research Program

3.1. Real-Time Biophysical Models

The principal objective of this scientific challenge is the modeling of the operative field, *i.e.* the anatomy and physiology of the patient that will be directly or indirectly targeted by a medical intervention. This requires to describe various phenomena such as soft-tissue deformation, fluid dynamics, electrical diffusion, or heat transfer. These models will help simulate the reaction of the patient's anatomy to the procedure, but also represent the behavior of complex organs such as the brain, the liver or the heart. A common requirement across these developments is the need for fast, possibly real-time, computation.

3.1.1. Real-time biomechanical modeling of solid structures

Soft tissue modeling holds a very important place in medical simulation. A large part of the realism of a simulation, in particular for surgery or laparoscopy simulation, relies upon the ability to describe soft tissue response during the simulated intervention. Several approaches have been proposed over the past ten years to model soft-tissue deformation in real-time (mainly for solid organs), usually based on elasticity theory and a finite element approach to solve the equations. We were among the first to propose an approach [3] using different computational strategies. Although significant improvements were obtained later on (for instance with the use of co-rotational methods to handle geometrical non-linearities) these works remain of limited clinical use as they essentially rely on linearized constitutive laws, and are rarely validated. An important part of our research remains dedicated to the development of new, more accurate models that are compatible with real-time computation. Such advanced models will not only permit to increase the realism of future training systems, but they will act as a bridge toward the development of patient-specific preoperative planning as well as augmented reality tools for the operating room.

3.1.2. Real-time biomechanical modeling of hollow structures

A large number of anatomical structures in the human body are vascularized (brain, liver, heart, kidneys, etc.) and recent interventions (such as interventional radiology procedures) rely on the vascular network as a therapeutical pathway. It is therefore essential to model the shape and deformable behavior of blood vessels. This can be done at two levels, depending of the objective. The global deformation of a vascular network can be represented using the vascular skeleton as a deformable (tree) structure, while local deformations need to be described using models of deformable surfaces. Other structures such as aneurysms, the colon or stomach can also benefit from being modeled as deformable surface, and we can rely on shell or thin plate theory to reach this objective.

3.1.3. Coupled physical models

Beyond biomechanical modeling of soft tissues, other physical phenomena have to be taken into account. In the context of percutaneous tumor ablation, both thermal and mechanical behaviors have to be modelled. Focusing especially on the simulation of cryoablation (freezing the pathological tissue), models for heat transfer have been implemented to simulate the evolution of temperature within living tissues. Pre-operative planning of the iceball can thus be envisaged. Moreover, this demonstrates the multi-physics aspect of SOFA.

3.1.4. Real-time electrophysiology

Electrophysiology plays an important role in the physiology of the human body, for instance by inducing muscles motion, and obviously through the nervous system. Also, many clinical procedures rely on electrical stimulation, such as defibrillation, neuromuscular or deep brain stimulation for instance. Yet, the modeling and the simulation of this phenomenon is still in its early stages. Our primary objective is to focus on cardiac electrophysiology, which plays a critical role in the understanding of heart mechanisms, and also in the planning of certain cardiac procedures. We propose to develop models and computational strategies aimed at real-time simulation, and to also provide personalization tools (parameter estimation) allowing to run patient-specific simulations from clinical data.

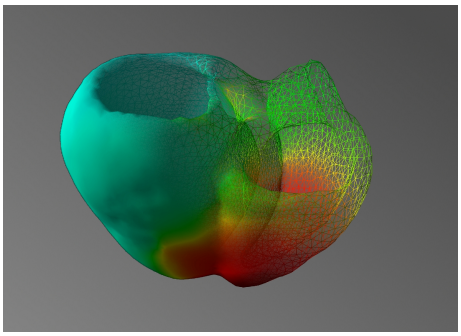

Figure 3. Patient-specific electrophysiology model of the human heart running in real-time

3.2. Numerical Methods for Complex Interactions

3.2.1. Dynamic topological changes

As mentioned previously, assisting the surgeon by providing either pre-operative planning or per-operative guidance assumes to increase the level of complexity and accuracy of our models, thus making the simulation more computationally-demanding. Innovative numerical methods must therefore be investigated. For instance, efforts are made to couple SOFA with CGoGN. CGoGN is a library based on combinatorial maps theory, specialized for representing and manipulating meshes. It is able to represent consistently objects of different dimensions composed of arbitrary cells (polygonal faces, polyhedral volumes). It provides an efficient way to explore the cells and their neighborhood; it allows to store data with the cells (both at execution time and compile time) and to efficiently modify the connectivity of the mesh even in highly dynamic cases. Adaptive meshing and efficient topological algorithm are thus available in SOFA.

3.2.2. Constraint models and boundary conditions

To simulate soft-tissue deformations accurately, the modeling technique must account for the intrinsic behavior of the modeled organ as well as for its biomechanical interactions with surrounding tissues or medical devices. While the biomechanical behavior of important organs (such as the brain or liver) has been studied extensively in the past, only few works exist dealing with the mechanical interactions between the anatomical structures. For tissue–tool interactions, most techniques rely on simple contact models, whereas advanced phenomena such as friction are rarely taken into account. While simplifications can produce plausible results in the case of interaction between the manipulator of a laparoscopic instrument and the surface of an organ, it is generally an insufficient approximation. As we move towards the simulations for planning or rehearsal, accurate modeling of contacts is playing an increasingly important role. For example, we have shown in [36] and [37] that complex interactions between a coil and an aneurysm, or alternatively between a flexible needle and a soft-tissue can be computed in real-time. In laparoscopic surgery, the main challenge is represented by modeling of interactions between anatomical structures rather than only between the instruments and the surface of the organ. Consequently, our objective was to model accurately the contacts with friction and other type on non-smooth interactions in a heterogeneous environment and to allow for stable haptic rendering. When different time integration strategies are used, another challenge is to compute the contact forces in such a way that integrity and stability of the overall simulation are maintained. Our objective was to propose a unified definition of such various boundary conditions and develop new numerical methods for simulations of heterogeneous objects.

3.3. Image-Driven Simulation: towards pre-operative planning and per-operative guidance

Image-guided therapy is a recent area of research that has the potential to bridge the gap between medical imaging and clinical routine by adapting pre-operative data to the time of the procedure. Several challenges are typically related to image-guided therapy, such as multi-modality image registration, which serves to align pre-operative images onto the patient. As most procedures deal with soft-tissues, elastic registration techniques are necessary to perform this step. Novel registration techniques began to account for soft tissue deformation using physically-based methods. Yet, several limitations still hinder the use of image-guided therapy in clinical routine. First, as registration methods become more complex, their computation time increases, thus lacking responsiveness. Second, as we have seen previously, many factors influence the deformation of soft-tissues, from patient-specific material properties to boundary conditions with surrounding anatomy. Another very similar, and related, problem is augmented reality, i.e. the real-time superposition of a virtual model onto the reality. In a clinical context, this can be very useful to help "see through" the anatomy. In this case, however, real-time registration of the virtual information onto the patient is mandatory. Our objective in this area is to combine our expertise in real-time soft-tissue modeling, complex interactions with image data to provide accurate and real-time registration, deformation, and tracking of virtual anatomical structures onto the patient.

The predictive capabilities of computer simulations may also be used to improve minimally invasive surgical procedures. While simulation results are sensitive to model parameters, initial and boundary conditions, we aim at combining computer-vision algorithms and simulation algorithms in order to produce dynamic data-driven simulation in clinical applications. The main idea is to use computer-vision algorithms from pre-operative diagnoses or per-operative video streams in order to extract meaningful data to feed the simulation engine and thus to increase the accuracy of the simulation. Clinical outcomes are expected in interventional radiology where the guidance is based on fluoroscopic imaging modality inducing high absorbed dose of X-rays for the patient and the clinical staff. In that context, using the prediction capabilities of the simulation may decrease the acquisition frequency of images, leading to a lower exposure of X-rays. Our objective in this area is to combine our expertise in patient-specific modeling and constraint models to achieve the dynamic coupling between images, pre-operative data and computer simulation.

TONUS Team

3. Research Program

3.1. Kinetic models for plasmas

The fundamental model for plasma physics is the coupled Vlasov-Maxwell kinetic model: the Vlasov equation describes the distribution function of particles (ions and electrons), while the Maxwell equations describe the electromagnetic field. In some applications, it may be necessary to take into account relativistic particles, which lead to consider the relativistic Vlasov equation, but generally, tokamak plasmas are supposed to be non relativistic. The particles distribution function depends on seven variables (three for space, three for velocity and one for time), which yields a huge amount of computations.

To these equations we must add several types of source terms and boundary conditions for representing the walls of the tokamak, the applied electromagnetic field that confines the plasma, fuel injection, collision effects, etc.

Tokamak plasmas possess particular features, which require developing specialized theoretical and numerical tools.

Because the magnetic field is strong, the particle trajectories have a very fast rotation around the magnetic field lines. A full resolution would require prohibitive amount of calculations. It is then necessary to develop models where the cyclotron frequency tends to infinity in order to obtain tractable calculations. The resulting model is called a gyrokinetic model. It allows us to reduce the dimensionality of the problem. Such models are implemented in GYSELA and Selalib. Those models require averaging of the acting fields during a rotation period along the trajectories of the particles. This averaging is called the gyroaverage and requires specific discretizations.

The tokamak and its magnetics fields present a very particular geometry. Some authors have proposed to return to the intrinsic geometrical versions of the Vlasov-Maxwell system in order to build better gyrokinetic models and adapted numerical schemes. This implies the use of sophisticated tools of differential geometry: differential forms, symplectic manifolds, and hamiltonian geometry.

In addition to theoretical modeling tools, it is necessary to develop numerical schemes adapted to kinetic and gyrokinetic models. Three kinds of methods are studied in TONUS: Particle-In-Cell (PIC) methods, semi-Lagrangian and fully Eulerian approaches.

3.1.1. Gyrokinetic models: theory and approximation

In most phenomena where oscillations are present, we can establish a three-model hierarchy: (i) the model parameterized by the oscillation period, (ii) the limit model and (iii) the Two-Scale model, possibly with its corrector. In a context where one wishes to simulate such a phenomenon where the oscillation period is small and where the oscillation amplitude is not small, it is important to have numerical methods based on an approximation of the Two-Scale model. If the oscillation period varies significantly over the domain of simulation, it is important to have numerical methods that approximate properly and effectively the model parameterized by the oscillation period and the Two-Scale model. Implemented Two-Scale Numerical Methods (for instance by Frénod et al. [36]) are based on the numerical approximation of the Two-Scale model. These are called of order 0. A Two-Scale Numerical Method is called of order 1 if it incorporates information from the corrector and from the equation to which this corrector is a solution. If the oscillation period varies between very small values and values of order 1, it is necessary to have new types of numerical schemes (Two-Scale Asymptotic Preserving Schemes of order 1 or TSAPS) with the property being able to preserve the asymptotics between the model parameterized by the oscillation period and the Two-Scale model with its corrector. A first work in this direction has been initiated by Crouseilles et al. [32].

3.1.2. Semi-Lagrangian schemes

The Strasbourg team has a long and recognized experience in numerical methods of Vlasov-type equations. We are specialized in both particle and phase space solvers for the Vlasov equation: Particle-in-Cell (PIC) methods and semi-Lagrangian methods. We also have a longstanding collaboration with the CEA of Cadarache for the development of the GYSELA software for gyrokinetic tokamak plasmas.

The Vlasov and the gyrokinetic models are partial differential equations that express the transport of the distribution function in the phase space. In the original Vlasov case, the phase space is the six-dimension position-velocity space. For the gyrokinetic model, the phase space is five-dimensional because we consider only the parallel velocity in the direction of the magnetic field and the gyrokinetic angular velocity instead of three velocity components.

A few years ago, Eric Sonnendrücker and his collaborators introduce a new family of methods for solving transport equations in the phase space. This family of methods are the semi-Lagrangian methods. The principle of these methods is to solve the equation on a grid of the phase space. The grid points are transported with the flow of the transport equation for a time step and interpolated back periodically onto the initial grid. The method is then a mix of particle Lagrangian methods and eulerian methods. The characteristics can be solved forward or backward in time leading to the Forward Semi-Lagrangian (FSL) or Backward Semi-Lagrangian (BSL) schemes. Conservative schemes based on this idea can be developed and are called Conservative Semi-Lagrangian (CSL).

GYSELA is a 5D full gyrokinetic code based on a classical backward semi-Lagrangian scheme (BSL) [43] for the simulation of core turbulence that has been developed at CEA Cadarache in collaboration with our team [37]. Although GYSELA was carefully developed to be conservative at lowest order, it is not exactly conservative, which might be an issue when the simulation is under-resolved, which always happens in turbulence simulations due to the formation of vortices which roll up.

3.1.3. PIC methods

Historically PIC methods have been very popular for solving the Vlasov equations. They allow solving the equations in the phase space at a relatively low cost. The main disadvantage of the method is that, due to its random aspect, it produces an important numerical noise that has to be controlled in some way, for instance by regularizations of the particles, or by divergence correction techniques in the Maxwell solver. We have a longstanding experience in PIC methods and we started implement them in SeLaLib. An important aspect is to adapt the method to new multicore computers. See the work by Crestetto and Helluy [31].

3.2. Reduced kinetic models for plasmas

As already said, kinetic plasmas computer simulations are very intensive, because of the gyrokinetic turbulence. In some situations, it is possible to make assumptions on the shape of the distribution function that simplify the model. We obtain in this way a family of fluid or reduced models.

Assuming that the distribution function has a Maxwellian shape, for instance, we obtain the MagnetoHydroDynamic (MHD) model. It is physically valid only in some parts of the tokamak (at the edges for instance). The fluid model is generally obtained from the hypothesis that the collisions between particles are strong. At Inria, fine collision models are mainly investigated in the KALIFFE team. In our approach we do not assume that the collisions are strong, but rather try to adapt the representation of the distribution function according to its shape, keeping the kinetic effects. The reduction is not necessarily a consequence of collisional effects. Indeed, even without collisions, the plasma may still relax to an equilibrium state over sufficiently long time scales (Landau damping effect). Recently, a team at the Plasma Physics Institut (IPP) in Garching has carried out a statistical analysis of the 5D distribution functions obtained from gyrokinetic tokamak simulations [38]. They discovered that the fluctuations are much higher in the space directions than in the velocity directions (see Figure 1).

This indicates that the approximation of the distribution function could require fewer data while still achieving a good representation, even in the collisionless regime.

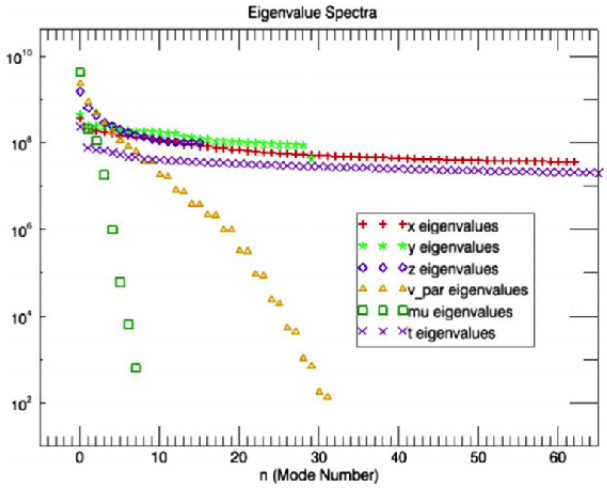

Figure 1. Space and velocity fluctuations spectra (from [38])

Our approach is different from the fluid approximation. In what follows we call this the "reduced model" approach. A reduced model is a model where the explicit dependence on the velocity variable is removed. In a more mathematical way, we consider that in some regions of the plasma, it is possible to exhibit a (preferably small) set of parameters α that allows us to describe the main properties of the plasma with a generalized "Maxwellian" M. Then

$$f(x, v, t) = M(\alpha(x, t), v).$$

In this case it is sufficient to solve for $\alpha(x, t)$. Generally, the vector α is solution of a first order hyperbolic system.

Several approaches are possible: waterbag approximations, velocity space transforms, *etc.*

3.2.1. Velocity space transformations

An experiment made in the 60's [41] exhibits in a spectacular way the reversible nature of the Vlasov equations. When two perturbations are applied to a plasma at different times, at first the plasma seems to damp and reach an equilibrium. But the information of the perturbations is still here and "hidden" in the high frequency microscopic oscillations of the distribution function. At a later time a resonance occurs and the plasma produces an echo. The time at which the echo occurs can be computed (see Villani [0], page 74). The fine mathematical study of this phenomenon allowed C. Villani and C. Mouhot to prove their famous result on the rigorous nonlinear Landau damping [42].

More practically, this experiment and its theoretical framework show that it is interesting to represent the distribution function by an expansion on an orthonormal basis of oscillating functions in the velocity variables. This representation allows a better control of the energy transfer between the low frequencies and the high frequencies in the velocity direction, and thus provides more relevant numerical methods. This kind of approach is studied for instance by Eliasson in [34] with the Fourier expansion.

[0] Landau damping. CEMRACS 2010 lectures. http://smai.emath.fr/cemracs/cemracs10/PROJ/Villani-lectures.pdf

In long time scales, filamentation phenomena result in high frequency oscillations in velocity space that numerical schemes cannot resolve. For stability purposes, most numerical schemes contain dissipation mechanisms that may affect the precision of the finest oscillations that could be resolved.

3.2.2. Adaptive modeling

Another trend in scientific computing is to optimize the computation time through adaptive modeling. This approach consists in applying the more efficient model locally, in the computational domain, according to an error indicator. In tokamak simulations, this kind of approach could be very efficient, if we are able to choose locally the best intermediate kinetic-fluid model as the computation runs. This field of research is very promising. It requires developing a clever hierarchy of models, rigorous error indicators, versatile software architecture, and algorithms adapted to new multicore computers.

3.2.3. Numerical schemes

As previously indicated, an efficient method for solving the reduced models is the Discontinuous Galerkin (DG) approach. It is possible to make it of arbitrary order. It requires limiters when it is applied to nonlinear PDEs occurring for instance in fluid mechanics. But the reduced models that we intent to write are essentially linear. The nonlinearity is concentrated in a few coupling source terms.

In addition, this method, when written in special set of variables, called the entropy variables, has nice properties concerning the entropy dissipation of the model. It opens the door to constructing numerical schemes with good conservation properties and no entropy dissipation, as already used for other systems of PDEs [44], [30], [40], [39].

3.3. Electromagnetic solvers

A precise resolution of the electromagnetic fields is essential for proper plasma simulation. Thus it is important to use efficient solvers for the Maxwell systems and its asymptotics: Poisson equation and magnetostatics.

The proper coupling of the electromagnetic solver with the Vlasov solver is also crucial for ensuring conservation properties and stability of the simulation.

Finally plasma physics implies very different time scales. It is thus very important to develop implicit Maxwell solvers and Asymptotic Preserving (AP) schemes in order to obtain good behavior on long time scales.

3.3.1. Coupling

The coupling of the Maxwell equations to the Vlasov solver requires some precautions. The most important is to control the charge conservation errors, which are related to the divergence conditions on the electric and magnetic fields. We will generally use divergence correction tools for hyperbolic systems presented for instance in [29] (and included references).

3.3.2. Implicit solvers

As already pointed out, in a tokamak, the plasma presents several different space and time scales. It is not possible in practice to solve the initial Vlasov-Maxwell model. It is first necessary to establish asymptotic models by letting some parameters (such as the Larmor frequency or the speed of light) tend to infinity. This is the case for the electromagnetic solver and this requires implementing implicit time solvers in order to efficiently capture the stationary state, the solution of the magnetic induction equation or the Poisson equation.

ALGORILLE Project-Team

3. Research Program

3.1. Structuring Applications

Computing on different scales is a challenge under constant development that, almost by definition, will always try to reach the edge of what is possible at any given moment in time: in terms of the scale of the applications under consideration, in terms of the efficiency of implementations and in what concerns the optimized utilization of the resources that modern platforms provide or require. The complexity of all these aspects is currently increasing rapidly.

3.1.1. Diversity of platforms

Design of processing hardware is diverging in many different directions. Nowadays we have SIMD registers inside processors, on-chip or off-chip accelerators (many-core boards, GPU, FPGA, vector-units), multi-cores and hyperthreading, multi-socket architectures, clusters, grids, clouds... The classical monolithic architecture of one-algorithm/one-implementation that solves a problem is obsolete in many cases. Algorithms (and the software that implements them) must deal with this variety of execution platforms robustly.

As we know, the "*free lunch*" for sequential algorithms provided by the increase of processor frequencies is over, we have to go parallel. But the "*free lunch*" is also over for many automatic or implicit adaptation strategies between codes and platforms: e.g the best cache strategies can't help applications that access memory randomly, or algorithms written for "simple" CPU (von Neumann model) have to be adapted substantially to run efficiently on vector units.

3.1.2. The communication bottleneck

Communication and processing capacities evolve at a different pace, thus the *communication bottleneck* is always narrowing. An efficient data management is becoming more and more crucial.

Not many implicit data models have yet found their place in the HPC domain, because of a simple observation: latency issues easily kill the performance of such tools. In the best case, they will be able to hide latency by doing some intelligent caching and delayed updating. But they can never hide the bottleneck for bandwidth. An efficient solution to this problem is the use of asynchronism in the algorithms. However, until now its application has been limited to iterative processes with specific constraints over the computational scheme.

HPC was previously able to cope with the communication bottleneck by using an explicit model of communication, namely MPI. It has the advantage of imposing explicit points in code where some guarantees about the state of data can be given. It has the clear disadvantage that coherence of data between different participants is difficult to manage and is completely left to the programmer.

Here, our approach is and will be to timely request explicit actions (like MPI) that mark the availability of (or need for) data. Such explicit actions ease the coordination between tasks (coherence management) and allow the platform underneath the program to perform a pro-active resource management.

3.1.3. Models of interdependence and consistency

Interdependence of data between different tasks of an application and components of hardware will be crucial to ensure that developments will possibly scale on the ever diverging architectures. We have up to now presented such models (PRO, DHO, ORWL) and their implementations, and proved their validity for the context of SPMD-type algorithms.

Over the next years we will have to enlarge the spectrum of their application. On the algorithm side we will have to move to heterogeneous computations combining different types of tasks in one application. Concerning the architectures, we will have to take into account the fact of increased heterogeneity, processors of different speeds, multi-cores, accelerators (FPU, GPU, vector units), communication links of different bandwidth and latency, memory and generally storage capacity of different size, speed and access characteristics. First implementations using ORWL in that context look particularly promising.

The models themselves will have to evolve to be better suited for more types of applications, such that they allow for a more fine-grained partial locking and access of objects. They should handle *e.g.* collaborative editing or the modification of just some fields in a data structure. This work has already started with DHO which allows the locking of *data ranges* inside an object. But a more structured approach would certainly be necessary here to be usable more comfortably in most applications.

3.1.4. Frequent I/O

A complete parallel application includes I/O of massive data, at an increasing frequency. In addition to applicative input and output data flows, I/O are used for checkpointing or to store traces of execution. These then can be used to restart in case of failure (hardware or software) or for a post-mortem analysis of a chain of computations that led to catastrophic actions (for example in finance or in industrial system control). The difficulty of frequent I/O is more pronounced on hierarchical parallel architectures that include accelerators with local memory.

I/O have to be included in the design of parallel programming models and tools. The ORWL library (Ordered Read-Write Lock) should be enriched with such tools and functionalities, in order to ease the modeling and development of parallel applications that include data IO, and to exploit most of the performance potential of parallel and distributed architectures.

3.1.5. Algorithmic paradigms

Concerning asynchronous algorithms, we have studied different variants of asynchronous models and developed several versions of implementations, allowing us to precisely study the impact of our design choices. However, we are still convinced that improvements are possible in order to extend the applicability of asynchronism, especially concerning the control of its behavior and the termination detection (global convergence of iterative algorithms). We have proposed some generic and non-intrusive way of implementing such a procedure in any parallel iterative algorithm.

3.1.6. Cost models and accelerators

We have already designed some models that relate computation power and energy consumption. Our present works in this topic concern the design and implementation of an auto-tuning system that controls the application according to user defined optimization criteria (computation and/or energy performance). This implies the insertion of multi-schemes and/or multi-kernels into the application such that it will be able to adapt its behavior to the requirements.

3.1.7. Design of dynamical systems for computational tasks

In the context of a collaboration with Nazim Fatès over dynamical systems, and especially cellular automata, we address a new way to study dynamical systems, that is more development oriented than analysis oriented. In fact, until now, most of the studies related to dynamical systems consisted in analyzing the dynamical properties (convergence, fixed points, cycles, initialization,...) of some given systems, and in describing the emergence of complex behaviors. Here, we focus on the dual approach that consists in designing dynamical systems in order to fulfill some given tasks. In this approach, we consider both theoretical and practical aspects.

3.2. Transparent Resource Management for Clouds

Given the extremely large offer of resources by public or private clouds, users need software assistance to make provisioning decisions. Our goal is to design a **cloud resource broker** which handles the workload of a user or

of a community of users as a multi-criteria optimization problem. The notions of resource usage, scheduling, provisioning and task management have been adapted to this new context. For example, to minimize the makespan of a DAG of tasks, usually a fixed number of resources is assumed. On IaaS clouds, the amount of resources can be provisioned at any time, and hence the scheduling problem must be redefined using one new prevalent optimization criterion: the financial cost of the computation.

3.2.1. Provisioning strategies

The provisioning strategies are hence central to the broker. They are designed after heuristics which aim to fit execution constraints and satisfy user preferences. For instance, lowering the costs can be achieved with strategies aiming at reusing already leased resources, or switch to less powerful and cheaper resources. However, some economic models proposed by cloud providers involve a complex cost-benefit analysis which we plan to address. Moreover, these economic models incur additional costs, *e.g.* for data storage or transfer, which have to be taken into account to design a comprehensive broker.

3.2.2. User workload analysis

Another possible extension of the capability of such a broker is the analysis of user workloads. Characterizing the workload might help to anticipate the behavior of each alternative provisioning strategy. The objective is to allow the user to select the suitable provisioning solution thanks to concrete information, such as completion time and financial cost.

3.2.3. Simulation of cloud platforms

Providing concrete information about provisioning solutions can also be achieved through simulation. Although predicting the behavior of applicative cases in real grid environment is made very difficult by the shared (*e.g.* multi-tenant), heterogeneous and dynamic nature of the resources, cloud resources (i.e. VMs) are perceived as reserved and homogeneous and stable by the end-user. Therefore, proposing an accurate prediction of the different strategies through an accurate simulation process would be a strong decision support for the user.

3.3. Experimental Methodologies for the Evaluation of Distributed Systems

Distributed systems are very challenging to study, test, and evaluate. Computer scientists traditionally prefer to study their systems *a priori* by reasoning theoretically on the constituents and their interactions. But the complexity of large-scale distributed systems makes this methodology near to impossible, explaining that most of the studies are done *a posteriori* through experiments.

In AlGorille, we strive at designing a comprehensive set of solutions for experimentation on distributed systems by working on several methodologies (formal assessment, simulation, use of experimental facilities, emulation) and by leveraging the convergence opportunities between methodologies (co-development, shared interfaces, validation combining several methodologies).

3.3.1. Simulation and Dynamic Verification

Our team plays a key role in the SimGrid project, a mature simulation toolkit widely used in the distributed computing community. Since more than ten years, we work on the validity, scalability and robustness of our tool.

Our current medium term goal is to extend the tool applicability to **Clouds and Exascale systems**. In the last years, we therefore worked toward disk and memory models in addition to the previously existing network and CPU models. The tool's scalability and efficiency also constitutes a permanent concern to us. **Interfaces** constitute another important work axis, with the addition of specific APIs on top of our simulation kernel. They provide the "syntactic sugar" needed to express algorithms of these communities. For example, virtual machines are handled explicitly in the interface provided for Cloud studies. Similarly, we pursue our work on an implementation of the full MPI standard allowing to study real applications using that interface. This work may also be extended in the future to other interfaces such as OpenMP or OpenCL.

We integrated a model checking kernel in SimGrid to enable **formal correctness studies** in addition to the practical performance studies enabled by simulation. Being able to study these two fundamental aspects of distributed applications within the same tool constitutes a major advantage for our users. In the future, we will enforce this capacity for the study of correctness and performance such that we hope to tackle their usage on real applications.

3.3.2. Experimentation on testbeds and production facilities, emulation

Our work in this research axis is meant to bring major contributions to the **industrialization of experimentation** on parallel and distributed systems. It is structured through multiple layers that range from the design of a testbed supporting high-quality experimentation, to the study of how stringent experimental methodology could be applied to our field, as depicted in Figure 2 .

During the last years, we have played a **key role in the design and development of Grid'5000** by leading the design and technical developments, and by managing several engineers working on the platform. We pursue our involvement in the design of the testbed with a focus on ensuring that the testbed provides all the features needed for high-quality experimentation. We also collaborate with other testbeds sharing similar goals in order to exchange ideas and views. We now work on **basic services supporting experimentation** such as resources verification, management of experimental environments, control of nodes, management of data, etc. Appropriate collaborations will ensure that existing solutions are adopted to the platform and improved as much as possible.

One key service for experimentation is the ability to alter experimental conditions using emulation. We work on the **Distem emulator**, focusing on its validation and on adding features (such as the ability to emulate faults, varying availability, churn, load injection, etc) and investigate if altering memory and disk performance is possible. Other goals are to scale the tool up to 20000 virtual nodes while improving the tool usability and documentation.

We work on **orchestration of experiments** in order to combine all the basic services mentioned previously in an efficient and scalable manner, with the design of a workflow-based experiment control engine named **XPFlow**.

3.3.3. Convergence and co-design of experimental methodologies

We see the experimental methodologies we work on as steps of a common experimental staircase: ideally, **one could and should leverage the various methodologies to address different facets of the same problem**. To facilitate that, we must co-design common or compatible formalisms, semantics and data formats.

Other experimental sciences such as biology and physics have paved the way in terms of scientific methodology. We **should learn from other experimental sciences, adopt good practices and adapt them** to Computer Science's specificities.

But Computer Science also has specific features that make it the ideal field to **create a truly Open Science**: provide infrastructure and tools for publishing and reproducing experiments and results, linked with our own methodologies and tools.

Finally, one important part of our work is to maintain a deep understanding of systems and their environments, in order to properly model them and experiment on them. Similarly, we need to understand the emerging scientific challenges in our field in order to improve adequately our experimental tools.

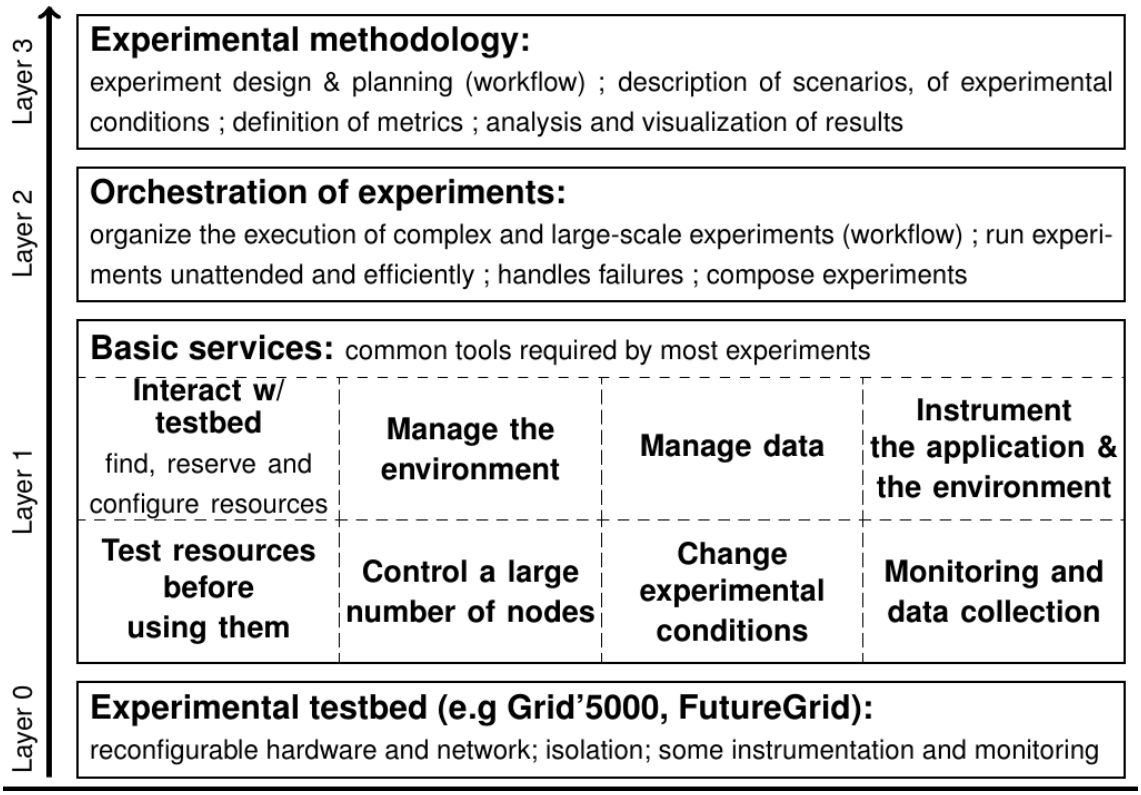

Figure 2. General structure of our project: We plan to address all layers of the experimentation stack.

COAST Team

3. Research Program

3.1. Introduction

Our scientific foundations are grounded on distributed collaborative systems supported by sophisticated data sharing mechanisms and on service oriented computing with an emphasis on orchestration and on non functional properties.

Distributed collaborative systems enable distributed group work supported by computer technologies. Designing such systems require an expertise in Distributed Systems and in Computer-supported collaborative activities research area. Besides theoretical and technical aspects of distributed systems, design of distributed collaborative systems must take into account the human factor to offer solutions suitable for users and groups. The COAST team vision is to move away from a centralized authority based collaboration towards a decentralized collaboration where users have full control over their data that they can store locally and decide with whom to share them. The Coast team investigates the issues related to the management of distributed shared data and coordination between users and groups.

Service oriented Computing [27] is an established domain on which the ECOO, SCORE and now the Coast team have been contributing for a long time. It refers to the general discipline that studies the development of computer applications on the web. A service is an independent software program with a specific functional context and capabilities published as a service contract (or more traditionally an API). A service composition aggregates a set of services and coordinates their interactions. The scale, the autonomy of services, the heterogeneity and some design principles underlying Service Oriented Computing open new research questions that are at the basis of our research. They span the disciplines of distributed computing, software engineering and computer supported collaborative work (CSCW). Our approach to contribute to the general vision of Service Oriented Computing and more generally to the emerging discipline of Service Science has been and is still to focus on the question of the efficient and flexible construction of reliable and secure high level services through the coordination/orchestration/composition of other services provided by distributed organizations or people.

3.2. Consistency Models for Distributed Collaborative Systems

Collaborative systems are distributed systems that allow users to share data. One important issue is to manage consistency of shared data according to concurrent access. Traditional consistency criteria such as locking, serializability, linearizability are not adequate for collaborative systems.

Causality, Convergence and Intention preservation (CCI) [30] are more suitable for developing middleware for collaborative applications.

We develop algorithms for ensuring CCI properties on collaborative distributed systems. Constraints on the algorithms are different according to the type of distributed system and type of data. The distributed system can be centralized, decentralized or peer-to-peer. The type of data can include strings, growable arrays, ordered trees, semantic graphs and multimedia data.

3.3. Optimistic Replication

Replication of data among different nodes of a network allows improving reliability, fault-tolerance, and availability. When data are mutable, consistency among the different replicas must be ensured. Pessimistic replication is based on the principle of single-copy consistency while optimistic replication allows the replicas to diverge during a short time period. The consistency model for optimistic replication [29] is called eventual consistency, meaning that replicas are guaranteed to converge to the same value when the system is idle.

Our research focuses on the two most promising families of optimistic replication algorithms for ensuring CCI:

- the operational transformation (OT) algorithms [25]
- the algorithms based on commutative replicated data types (CRDT) [28].

Operational transformation algorithms are based on the application of a transformation function when a remote modification is integrated into the local document. Integration algorithms are generic, being parametrized by operational transformation functions which depend on replicated document types. The advantage of these algorithms is their genericity. These algorithms can be applied to any data type and they can merge heterogeneous data in a uniform manner.

Commutative replicated data types is a new class of algorithms initiated by WOOT [26] a first algorithm designed WithOut Operational Transformations. They ensure consistency of highly dynamic content on peer-to-peer networks. Unlike traditional optimistic replication algorithms, they can ensure consistency without concurrency control. CRDT algorithms rely on natively commutative operations defined on abstract data types such as lists or ordered trees. Thus, they do not require a merge algorithm or an integration procedure.

3.4. Process Orchestration and Management

Process Orchestration and Management is considered as a core discipline behind Service Management and Computing. It includes the analysis, the modelling, the execution, the monitoring and the continuous improvement of enterprise processes and is for us a central domain of studies.

Much efforts has been devoted in the past years to establish standard business process models founded on well grounded theories (e.g. Petri Nets) that meet the needs of both business analysts but also of software engineers and software integrators. This has lead to heated debate as both points of view are very difficult to reconciliate between the analyst side and the IT side. On one side, the business people in general require models that are easy to use and understand and that can be quickly adapted to exceptional situations. On the other side, IT people need models with an operational semantic in order to be able transform them into executable artefacts. Part of our work has been an attempt to reconcile these point of views. It has lead to the development of Bonita product and more recently on our work in crisis management where the same people are designing, executing and monitoring the process as it executes. But more generally, and at a larger scale, we have been considering the problem of process spanning the barriers of organisations. This leads us to consider the more general problem of service composition as a way to coordinate inter organisational construction of applications providing value based on the composition of lower level services [24].

3.5. Service Composition

More and more, we are considering processes as pieces of software whose execution traverse the boundaries of organisations. This is especially true with service oriented computing where processes compose services produced by many organisations. We tackle this problem from very different perspectives, trying to find the best compromise between the need for privacy of internal processes from organisations and the necessity to publicize large part of them, proposing to distribute the execution and the orchestration of processes among the organisations themselves, and attempting to ensure non functional properties in this distributed setting [23].

Non functional aspects of service composition relate to all the properties and service agreements that one want to ensure and that are orthogonal to the actual business but that are important when a service is selected and integrated in a composition. This includes transactional context, security, privacy, and quality of service in general. Defining and orchestrating services on a large scale while providing the stakeholders with some strong guarantees on their execution is a first class problem for us. For a long time, we have proposed models and solutions to ensure that some properties (e.g. transactional properties) were guaranteed on process execution, either through design or through the definition of some protocols. Our work has also been extended to the problems of security, privacy and service level agreement among partners. These questions are still central in our work. Then, one major problem of current approaches is to monitor the execution

of the compositions, integrating the distributed dimension. This problem can be tackled using event-based algorithms and techniques. Using our event oriented composition framework DISC, we have obtained new results dedicated to the runtime verification of violations in service choreographies.

MADYNES Project-Team

3. Research Program

3.1. Evolutionary needs in network and service management

The foundation of the MADYNES research activity is the ever increasing need for automated monitoring and control within networked environments. This need is mainly due to the increasing dependency of both people and goods towards communication infrastructures as well as the growing demand towards services of higher quality. Because of its strategic importance and crucial requirements for interoperability, the management models were constructed in the context of strong standardization activities by many different organizations over the last 15 years. This has led to the design of most of the paradigms used in today's deployed approaches. These paradigms are the Manager/Agent interaction model, the Information Model paradigm and its container, together with a naming infrastructure called the Management Information Base. In addition to this structure, five functional areas known under Fault, Configuration, Accounting, Performance and Security are associated to these standards.

While these models were well suited for the specific application domains for which they were designed (telecommunication networks or dedicated protocol stacks), they all show the same limits. Especially they are unable:

1. to deal with any form of dynamicity in the managed environment,
2. to master the complexity, the operating mode and the heterogeneity of the emerging services,
3. to scale to new networks and service environments.

These three limits are observed in all five functional areas of the management domain (fault, configuration, accounting, performance and security) and represent the major challenges when it comes to enable effective automated management and control of devices, networks and services in the next decade.

MADYNES addresses these challenges by focusing on the design of management models that rely on inherently dynamic and evolving environments. The project is centered around two core activities. These activities are, as mentioned in the previous section, the design of an autonomous management framework and its application to three of the standard functional areas namely security, configuration and performance.

3.2. Autonomous management

3.2.1. Models and methods for a self-management plane

Self organization and automation are fundamental requirements within the management plane in today's dynamic environments. It is necessary to automate the management processes and enable management frameworks to operate in time sensitive evolving networks and service environments. The automation of the organization of devices, software components, networks and services is investigated in many research projects and has already led to several solution proposals. While these proposals are successful at several layers, like IP auto-configuration or service discovery and binding facilities, they did not enhance the management plane at all. For example, while self-configuration of IP devices is commonplace, no solution exists that provides strong support to the management plane to configure itself (e.g. finding the manager to which an agent has to send traps or organizing the access control based on locality or any other context information). So, this area represents a major challenge in extending current management approaches so that they become self-organized.

Our approach is bottom-up and consists in identifying those parameters and framework elements (manager data, information model sharing, agent parameters, protocol settings, ...) that need dynamic configuration and self-organization (like the address of a trap sink). For these parameters and their instantiation in various management frameworks (SNMP, Netconf, WBEM, ...), we investigate and elaborate novel approaches enabling fully automated setup and operation in the management plane.

3.2.2. Design and evaluation of P2P-based management architectures

Over the last years, several models have emerged and gained wide acceptance in the networking and service world. Among them, the overlay networks together with the P2P paradigms appear to be very promising. Since they rely mainly on fully decentralized models, they offer excellent fault tolerance and have a real potential to achieve high scalability. Mainly deployed in the content delivery and the cooperation and distributed computation disciplines, they seem to offer all features required by a management framework that needs to operate in a dynamic world. This potential however needs an in depth investigation because these models have also many characteristics that are unusual in management (e.g. a fast and uncontrolled evolution of the topology or the existence of a distributed trust relationship framework rather than a standard centralized security framework).

Our approach envisions how a complete redesign of a management framework is done given the characteristics of the underlying P2P and overlay services. Among the topics of interest we study the concept of management information and operations routing within a management overlay as well as the distribution of management functions in a multi-manager/agent P2P environment. The functional areas targeted in our approach by the P2P model are network and service configuration and distributed monitoring. The models are to be evaluated against highly dynamic frameworks such as ad-hoc environments (network or application level) and mobile devices.

3.2.3. Integration of management information

Representation, specification and integration of management information models form a foundation for network and service management and remains an open research domain. The design and specification of new models is mainly driven by the appearance of new protocols, services and usage patterns. These need to be managed and exposed through well designed management information models. Integration activities are driven by the multiplication of various management approaches. To enable automated management, these approaches need to inter-operate which is not the case today.

The MADYNES approach to this problem of modeling and representation of management information aims at:

1. enabling application developers to establish their management interface in the same workspace, with the same notations and concepts as the ones used to develop their application,
2. fostering the use of standard models (at least the structure and semantics of well defined models),
3. designing a naming structure that allows the routing of management information in an overlay management plane, and
4. evaluating new approaches for management information integration especially based on management ontologies and semantic information models.

3.2.4. Modeling and benchmarking of dynamic networks

The impact of a management approach on the efficiency of the managed service is highly dependent on three factors:

- the distribution of the considered service and their associated management tasks,
- the management patterns used (e.g. monitoring frequency, granularity of the management information considered),
- the cost in terms of resources these considered functions have on the managed element (e.g. method call overhead, management memory footprint).

MADYNES addresses this problem from multiple viewpoints: communication patterns, processing and memory resources consumption. Our goal is to provide management patterns combining optimized management technologies so as to optimize the resources consumed by the management activity imposed by the operating environment while ensuring its efficiency in large dynamic networks.

3.3. Functional areas

3.3.1. Security management

Securing the management plane is vital. While several proposals are already integrated in the existing management frameworks, they are rarely used. This is due to the fact that these approaches are completely detached from the enterprise security framework. As a consequence, the management framework is "managed" separately with different models; this represents a huge overhead. Moreover the current approaches to security in the management plane are not inter-operable at all, multiplying the operational costs in a heterogeneous management framework.

The primary goal of the research in this activity is the design and the validation of a security framework for the management plane that will be open and capable to integrate the security services provided in today's management architectures. Management security interoperability is of major importance in this activity.

Our activity in this area aims at designing a generic security model in the context of multi-party / multi-technology management interactions. Therefore, we develop research on the following directions:

1. Abstraction of the various access control mechanisms that exist in today's management frameworks. We are particularly interested in extending these models so that they support event-driven management, which is not the case for most of them today.
2. Extension of policy and trust models to ease and to ensure coordination among managers towards one agent or a subset of the management tree. Provisional policies are of great interest to us in this context.
3. Evaluation of the adequacy of key distribution architectures to the needs of the management plane as well as selecting reputation models to be used in the management of highly dynamic environments (e.g. multicast groups, ad-hoc networks).

A strong requirement towards the future generic model is that it needs to be instantiated (with potential restrictions) into standard management platforms like SNMP, WBEM or Netconf and to allow interoperability in environments where these approaches coexist and even cooperate. A typical example of this is the security of an integration agent which is located in two management worlds.

Since 2006 we have also started an activity on security assessment. The objective is to investigate new methods and models for validating the security of large scale dynamic networks and services. The first targeted service is VoIP.

3.3.2. Configuration: automation of service configuration and provisioning

Configuration covers many processes which are all important to enable dynamic networks. Within our research activity, we focus on the operation of tuning the parameters of a service in an automated way. This is done together with the activation topics of configuration management and the monitoring information collected from the underlying infrastructure. Some approaches exist today to automate part of the configuration process (download of a configuration file at boot time within a router, on demand code deployment in service platforms). While these approaches are interesting they all suffer from the same limits, namely:

1. they rely on specific service life cycle models,
2. they use proprietary interfaces and protocols.

These two basic limits have high impacts on service dynamics in a heterogeneous environment.

We follow two research directions in the topic of configuration management. The first one aims at establishing an abstract life-cycle model for either a service, a device or a network configuration and to associate with this model a generic command and programming interface. This is done in a way similar to what is proposed in the area of call control in initiatives such as Parlay or OSA.

In addition to the investigation of the life-cycle model, we work on technology support for distributing and exchanging configuration management information. Especially, we investigate policy-driven approaches for representing configurations and constraints while we study XML-based protocols for coordinating distribution and synchronization. Off and online validation of configuration data is also part of this effort.

3.3.3. Performance and availability monitoring

Performance management is one of the most important and deployed management function. It is crucial for any service which is bound to an agreement about the expected delivery level. Performance management needs models, metrics, associated instrumentation, data collection and aggregation infrastructures and advanced data analysis algorithms.

Today, a programmable approach for end-to-end service performance measurement in a client server environment exists. This approach, called Application Response Measurement (ARM) defines a model including an abstract definition of a unit of work and related performance records; it offers an API to application developers which allows easy integration of measurement within their distributed application. While this approach is interesting, it is only a first step toward the automation of performance management.

We are investigating two specific aspects. First we are working on the coupling and possible automation of performance measurement models with the upper service level agreement and specification levels. Second we are working on the mapping of these high level requirements to the lower level of instrumentation and actual data collection processes available in the network. More specifically we are interested in providing automated mapping of service level parameters to monitoring and measurement capabilities. We also envision automated deployment and/or activation of performance measurement sensors based on the mapped parameters. This activity also incorporates self-instrumentation (and when possible on the fly instrumentation) of software components for performance monitoring purpose.

ALICE Project-Team

3. Research Program

3.1. Introduction

Computer Graphics is a quickly evolving domain of research. These last few years, both acquisition techniques (e.g., range laser scanners) and computer graphics hardware (the so-called GPU's, for Graphics Processing Units) have made considerable advances. However, despite these advances, fundamental problems still remain open. For instance, a scanned mesh composed of hundred million triangles cannot be used directly in real-time visualization or complex numerical simulation. To design efficient solutions for these difficult problems, ALICE studies two fundamental issues in Computer Graphics:

- the representation of the objects, i.e., their geometry and physical properties;
- the interaction between these objects and light.

Historically, these two issues have been studied by independent research communities. However, we think that they share a common theoretical basis. For instance, multi-resolution and wavelets were mathematical tools used by both communities [28]. We develop a new approach, which consists in studying the geometry and lighting from the *numerical analysis* point of view. In our approach, geometry processing and light simulation are systematically restated as a (possibly non-linear and/or constrained) functional optimization problem. This type of formulation leads to algorithms that are more efficient. Our long-term research goal is to find a formulation that permits a unified treatment of geometry and illumination over this geometry.

3.2. Geometry Processing for Engineering

Keywords: Mesh processing, parameterization, splines

Geometry processing recently emerged (in the middle of the 90's) as a promising strategy to solve the geometric modeling problems encountered when manipulating meshes composed of hundred millions of elements. Since a mesh may be considered to be a *sampling* of a surface - in other words a *signal* - the *digital signal processing* formalism was a natural theoretic background for this subdomain (see e.g., [29]). Researchers of this domain then studied different aspects of this formalism applied to geometric modeling.

Although many advances have been made in the geometry processing area, important problems still remain open. Even if shape acquisition and filtering is much easier than 30 years ago, a scanned mesh composed of hundred million triangles cannot be used directly in real-time visualization or complex numerical simulation. For this reason, automatic methods to convert those large meshes into higher level representations are necessary. However, these automatic methods do not exist yet. For instance, the pioneer Henri Gouraud often mentions in his talks that the *data acquisition* problem is still open. Malcolm Sabin, another pioneer of the "Computer Aided Geometric Design" and "Subdivision" approaches, mentioned during several conferences of the domain that constructing the optimum control-mesh of a subdivision surface so as to approximate a given surface is still an open problem. More generally, converting a mesh model into a higher level representation, consisting of a set of equations, is a difficult problem for which no satisfying solutions have been proposed. This is one of the long-term goals of international initiatives, such as the AIMShape European network of excellence.

Motivated by gridding application for finite elements modeling for oil and gas exploration, in the frame of the Gocad project, we started studying geometry processing in the late 90's and contributed to this area at the early stages of its development. We developed the LSCM method (Least Squares Conformal Maps) in cooperation with Alias Wavefront [24]. This method has become the de-facto standard in automatic unwrapping, and was adopted by several 3D modeling packages (including Maya and Blender). We experimented various applications of the method, including normal mapping, mesh completion and light simulation [2].

However, classical mesh parameterization requires to partition the considered object into a set of topological disks. For this reason, we designed a new method (Periodic Global Parameterization) that generates a continuous set of coordinates over the object [5]. We also showed the applicability of this method, by proposing the first algorithm that converts a scanned mesh into a Spline surface automatically [4].

We are still not fully satisfied with these results, since the method remains quite complicated. We think that a deeper understanding of the underlying theory is likely to lead to both efficient and simple methods. For this reason, in 2012 we studied several ways of discretizing partial differential equations on meshes, including Finite Element Modeling and Discrete Exterior Calculus. In 2013, we also explored Spectral Geometry Processing and Sampling Theory (more on this below).

3.3. Computer Graphics

Keywords: texture synthesis, shape synthesis, texture mapping, visibility

Content creation is one of the major challenges in Computer Graphics. Modeling shapes and surface appearances which are visually appealing and at the same time enforce precise design constraints is a task only accessible to highly skilled and trained designers.

In this context the team focuses on methods for by-example content creation. Given an input example and a set of constraints, we design algorithms that can automatically generate a new shape (geometry+texture). We formulate the problem of content synthesis as the joint optimization of several objectives: Preserving the local appearance of the example, enforcing global objectives (size, symmetries, mechanical properties), reaching user defined constraints (locally specified geometry, contacts). This results in a wide range of optimization problems, from statistical approaches (Markov Random fields), to combinatorial and linear optimization techniques.

As as complement to the design of techniques for automatic content creation, we also work on the representation of the content, so as to allow for its efficient manipulation. In this context we develop data-structures and algorithms targeted at massively parallel architectures, such as GPUs. These are critical to reach the interactive rates expected from a content creation technique. We also propose novel ways to store and access content stored along surfaces [6] or in volumes [1] [23].

The team also continues research in core topics of computer graphics at the heart of realistic rendering and realistic light simulation techniques; for example, mapping textures on surfaces, or devising visibility relationships between 3D objects populating space.

MAGRIT Project-Team

3. Research Program

3.1. Matching and 3D tracking

One of the most basic problems currently limiting AR applications is the registration problem. The objects in the real and virtual worlds must be properly aligned with respect to each other, or the illusion that the two worlds coexist will be compromised.

As a large number of potential AR applications are interactive, real time pose computation is required. Although the registration problem has received a lot of attention in the computer vision community, the problem of real-time registration is still far from being a solved problem, especially for unstructured environments. Ideally, an AR system should work in all environments, without the need to prepare the scene ahead of time, independently of the variations in experimental conditions (lighting, weather condition,...) which may exist between the application and the time the model of the scene was acquired.

For several years, the MAGRIT project has been aiming at developing on-line and marker-less methods for camera pose computation. The main difficulty with on-line tracking is to ensure robustness of the process over time. For off-line processes, robustness is achieved by using spatial and temporal coherence of the considered sequence through move-matching techniques. To get robustness for open-loop systems, we have investigated various methods, ranging from statistical methods to the use of hybrid camera/sensor systems. Many of these methods are dedicated to piecewise-planar scenes and combine the advantage of move-matching methods and model-based methods. In order to reduce statistical fluctuations in viewpoint computation, which lead to unpleasant jittering or sliding effects, we have also developed model selection techniques which allow us to noticeably improve the visual impression and to reduce drift over time. Another line of research which has been considered in the team to improve the reliability and the robustness of pose algorithms is to combine the camera with another form of sensor in order to compensate for the shortcomings of each technology [1].

The success of pose computation over time largely depends on the quality of the matching stage at the initialization stage. Indeed, the current image may be very different from the appearances described in the model both on the geometrical and the photometric sides. Research is thus conducted in the team on the use of probabilistic methods to establish robust correspondences of features. The use of *a contrario* has been investigated to achieve this aim [6]. We especially addressed the complex case of matching in scenes with repeated patterns which are common in urban scenes. We are also investigating the problem of matching images taken from very different viewpoints which is central for the re-localization issue in AR. Within the context of a scene model acquired with structure from motion techniques, we are currently investigating the use of viewpoint simulation in order to allow successful pose computation even if the considered image is far from the positions used to build the model.

Recently, the issue of tracking deformable objects has gained importance in the team. This topic is mainly addressed in the context of medical applications through the design of bio-mechanical models guided by visual features [2]. We have successfully investigated the use of such models in laparoscopy, with a vascularized model of the liver and with an hyper-elastic model for tongue tracking in US images. However, these results have been obtained so far in relatively controlled environments, with non pathological cases. When clinical routine applications are to be considered, many parameters and considerations need to be taken into account. Among the problems that need to be solved are the model representation, the specification of the range of physical parameters and the need to enforce the robustness of the tracking with respect to outliers, which are common in the interventional context...

3.2. Image-based Modeling

Modeling the scene is a fundamental issue in AR for many reasons. First, pose computation algorithms often use a model of the scene or at least some 3D knowledge on the scene. Second, effective AR systems require a model of the scene to support interactions between the virtual and the real objects such as occlusions, lighting reflexions, contacts...in real time. Unlike pose computation which has to be computed in a sequential way, scene modeling can be considered as an off-line or an on-line problem depending on the requirements of the targeted application. Interactive in-situ modeling techniques have thus been developed with the aim to enable the user to define what is relevant at the time the model is being built during the application. On the other hand, we also proposed off-line multimodal techniques, mainly dedicated to AR medical applications, with the aim to obtain realistic and possibly dynamic models of organs suitable for real time simulation.

In-situ modeling
In situ modeling allows a user to directly build a 3D model of his/her surrounding environment and verify the geometry against the physical world in real time. This is of particular interest in using AR in unprepared environments or building scenes that have an ephemeral existences (e.g. a film set) or cannot be accessed frequently (e.g. a nuclear power plant). We have especially investigated two systems, one based on the image content only and the other based on multiple data coming from different sensors (camera, inertial measurement unit, laser rangefinder). Both systems use the camera-mouse principle [5] (i.e. interactions are performed by aiming at the scene through a video camera) and both systems have been designed to acquire polygonal textured models, which are particularly useful for camera tracking and object insertion in AR.

Multimodal modeling for real time simulation
With respect to classical AR applications, AR in medical context differs in the nature and the size of the data which are available: a large amount of multimodal data is acquired on the patient or possibly on the operating room through sensing technologies or various image acquisitions. The challenge is to analyze these data, to extract interesting features, to fuse and to visualize this information in a proper way. Within the MAGRIT team, we address several key problems related to medical augmented environments. Being able to acquire multimodal data which are temporally synchronized and spatially registered is the first difficulty we face when considering medical AR. Another key requirement of AR medical systems is the availability of 3D (+t) models of the organ/patient built from images, to be overlaid onto the users's view of the environment.

Methods for multimodal modeling are strongly dependent on the image modalities and the organ specificities. We thus only address a restricted number of medical applications –interventional neuro-radiology, laparoscopic surgery, Augmented Head project– for which we have a strong expertise and close relationships with motivated clinicians. In these applications, our aim is to produce realistic models and then realistic simulations of the patient to be used for surgeon's training or patient's re-education/learning.

One of our main applications is about neuroradiology. For the last 20 years, we have been working in close collaboration with the neuroradiology laboratory (CHU-University Hospital of Nancy) and GE Healthcare. As several imaging modalities are now available in a intraoperative context (2D and 3D angiography, MRI, ...), our aim is to develop a multi-modality framework to help therapeutic decision and treatment.

We have mainly been interested in the effective use of a multimodality framework in the treatment of arteriovenous malformations (AVM) and aneurysms in the context of interventional neuroradiology. The goal of interventional gestures is to guide endoscopic tools towards the pathology with the aim to perform embolization of the AVM or to fill the aneurysmal cavity by placing coils. We have proposed and developed multimodality and augmented reality tools which make various image modalities (2D and 3D angiography, fluoroscopic images, MRI, ...) cooperate in order to help physicians in clinical routine. One of the successes of this collaboration is the implementation of the concept of *augmented fluoroscopy*, which helps the surgeon to guide endoscopic tools towards the pathology. Lately, in cooperation with the EPC SHACRA, we have proposed new methods for implicit modeling of the aneurysms with the aim of obtaining near real time simulation of the coil deployment in the aneurysm [8]. These works open the way towards near real time patient-based simulations of interventional gestures both for training or for planning.

3.3. Parameter estimation

Many problems in computer vision or image analysis can be formulated in terms of parameter estimation from image-based measurements. This is the case of many problems addressed in the team such as pose computation or image-guided estimation of 3D deformable models... Often traditional robust techniques which take into account the covariance on the measurements are sufficient to achieve reliable parameter estimation. However, depending on their number, their spatial distribution and the uncertainty on these measurements, some problems are very sensitive to noise and there is a considerable interest in considering how parameter estimation could be improved if additional information on the noise is available. Another common problem in our field of research is the need to estimate constitutive parameters of the models, such as (bio)-mechanical parameters for instance. Direct measurement methods are destructive and elaborating image based methods is thus highly desirable. Besides designing appropriate estimation algorithms, a fundamental question is to understand what group of parameters under study can be reliably estimated from a given experimental setup.

This line of research is relatively new in the team. One of the challenges is to improve image-based parameter estimation techniques considering sensor noise and specific image formation models. In a collaboration with the Pascal Institute (Clermont Ferrand), metrological performance enhancement for experimental solid mechanics has been addressed through the development of dedicated signal processing methods [12]. In the medical field, specific methods based on an adaptive evolutionary optimization strategy have been designed for estimating respiratory parameters [7]. In the context of designing realistic simulators for neuroradiology, we are now considering how parameters involved in the simulation could be adapted to fit real images.

MAIA Project-Team

3. Research Program

3.1. Sequential Decision Making

3.1.1. Synopsis and Research Activities

Sequential decision making consists, in a nutshell, in controlling the actions of an agent facing a problem whose solution requires not one but a whole sequence of decisions. This kind of problem occurs in a multitude of forms. For example, important applications addressed in our work include: Robotics, where the agent is a physical entity moving in the real world; Medicine, where the agent can be an analytic device recommending tests and/or treatments; Computer Security, where the agent can be a virtual attacker trying to identify security holes in a given network; and Business Process Management, where the agent can provide an auto-completion facility helping to decide which steps to include into a new or revised process. Our work on such problems is characterized by three main lines of research:

- (A) *Understanding how, and to what extent, to best model the problems.*
- (B) *Developing algorithms solving the problems and understanding their behavior.*
- (C) *Applying our results to complex applications.*

Before we describe some details of our work, it is instructive to understand the basic forms of problems we are addressing. We characterize problems along the following main dimensions:

- (1) Extent of the model: full vs. partial vs. none. This dimension concerns how complete we require the model of the problem – if any – to be. If the model is incomplete, then learning techniques are needed along with the decision making process.
- (2) Form of the model: factored vs. enumerative. Enumerative models explicitly list all possible world states and the associated actions etc. Factored models can be exponentially more compact, describing states and actions in terms of their behavior with respect to a set of higher-level variables.
- (3) World dynamics: deterministic vs. stochastic. This concerns our initial knowledge of the world the agent is acting in, as well as the dynamics of actions: is the outcome known a priori or are several outcomes possible?
- (4) Observability: full vs. partial. This concerns our ability to observe what our actions actually do to the world, i.e., to observe properties of the new world state. Obviously, this is an issue only if the world dynamics are stochastic.

These dimensions are wide-spread in the AI literature and are not exhaustive, in particular the MAIA team is also interested by discrete/continuous or centralized/decentralized problems. The complexity of solving a problem – both in theory and in practice – depends heavily on where it resides in this categorization. A common practice is to address simplified problems, leading to perhaps *sub-optimal* solutions while trying to characterize how far from the *optimal* solution we stand.

In what follows, we outline the main formal frameworks on which our work is based; while doing so, we highlight in a little more detail our core research questions. We then give a brief summary of how our work fits into the global research context.

3.1.2. Formal Frameworks

3.1.2.1. Deterministic Sequential Decision Making

Sequential decision making with deterministic world dynamics is most commonly known as *planning*, or *classical planning* [49]. Obviously, in such a setting every world state needs to be considered at most once, and thus enumerative models do not make sense (the problem description would have the same size as the space of possibilities to be explored). Planning approaches support factored description languages in which complex problems can be modeled in a compact way. Approaches to automatically learn such factored models do exist, however most works – and also most of our works on this form of sequential decision making – assume that the model is provided by the user of the planning technology. Formally, a problem instance, commonly referred to as a *planning task*, is a four-tuple $\langle V, A, I, G\rangle$. Here, V is a set of variables; a value assignment to the variables is a world state. A is a set of actions described in terms of two formulas over V: their preconditions and effects. I is the initial state, and G is a goal condition (again a formula over V). A solution, commonly referred to as a *plan*, is a schedule of actions that is applicable to I and achieves G.

Planning is *PSPACE-complete* even under strong restrictions on the formulas allowed in the planning task description. Research thus revolves around the development and understanding of search methods, which explore, in a variety of different ways, the space of possible action schedules. A particularly successful approach is *heuristic search*, where search is guided by information obtained in an automatically designed *relaxation* (simplified version) of the task. We investigate the design of relaxations, the connections between such design and the search space topology, and the construction of effective *planning systems* that exhibit good practical performance across a wide range of different inputs. Other important research lines concern the application of ideas successful in planning to stochastic sequential decision making (see next), and the development of technology supporting the user in model design.

3.1.2.2. Stochastic Sequential Decision Making

Markov Decision Processes (*MDP*) [51] are a natural framework for stochastic sequential decision making. An MDP is a four-tuple $\langle S, A, T, r\rangle$, where S is a set of states, A is a set of actions, $T(s, a, s') = P(s'|s, a)$ is the probability of transitioning to s' given that action a was chosen in state s, and $r(s, a, s')$ is the (possibly stochastic) reward obtained from taking action a in state s, and transitioning to state s'. In this framework, one looks for a *strategy*: a precise way for specifying the sequence of actions that induces, on average, an optimal sum of discounted rewards $\mathrm{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right]$. Here, $(r_0, r_1, ...)$ is the infinitely-long (random) sequence of rewards induced by the strategy, and $\gamma \in (0, 1)$ is a discount factor putting more weight on rewards obtained earlier. Central to the MDP framework is the Bellman equation, which characterizes the *optimal value function*V^*:

$$\forall s \in S, \quad V^*(s) = \max_{a \in A} \sum_{s' \in S} T(s, a, s')[r(s, a, s') + \gamma V^*(s')].$$

Once the optimal value function is computed, it is straightforward to derive an optimal strategy, which is deterministic and memoryless, i.e., a simple mapping from states to actions. Such a strategy is usually called a *policy*. An *optimal policy* is any policy π^* that is *greedy* with respect to V^*, i.e., which satisfies:

$$\forall s \in S, \quad \pi(s) \in \arg\max_{a \in A} \sum_{s' \in S} T(s, a, s')[r(s, a, s') + \gamma V^*(s')].$$

An important extension of MDPs, known as Partially Observable MDPs (*POMDPs*) allows to account for the fact that the state may not be fully available to the decision maker. While the goal is the same as in an MDP (optimizing the expected sum of discounted rewards), the solution is more intricate. Any POMDP can be seen to be equivalent to an MDP defined on the space of probability distributions on states, called *belief states*. The Bellman-machinery then applies to the belief states. The specific structure of the resulting MDP makes it possible to iteratively approximate the optimal value function – which is convex in the *belief space* – by piecewise linear functions, and to deduce an optimal policy that maps belief states to actions. A further

extension, known as a DEC-POMDP, considers $n \geq 2$ agents that need to control the state dynamics in a decentralized way without direct communication.

The MDP model described above is enumerative, and the complexity of computing the optimal value function is *polynomial* in the size of that input. However, in examples of practical size, that complexity is still too high so naïve approaches do not scale. We consider the following situations: (i) when the state space is large, we study approximation techniques from both a theoretical and practical point of view; (ii) when the model is unknown, we study how to learn an optimal policy from samples (this problem is also known as Reinforcement Learning [55]); (iii) in factored models, where MDP models are a strict generalization of classical planning – and are thus at least *PSPACE*-hard to solve – we consider using search heuristics adapted from such (classical) planning.

Solving a POMDP is *PSPACE*-hard even given an enumerative model. In this framework, we are mainly looking for assumptions that could be exploited to reduce the complexity of the problem at hand, for instance when some actions have no effect on the state dynamics (*active sensing*). The decentralized version, DEC-POMDP, induces a significant increase in complexity (*NEXP*-complete). We tackle the challenging – even for (very) small state spaces – exact computation of finite-horizon optimal solutions through alternative reformulations of the problem. We also aim at proposing advanced heuristics to efficiently address problems with more agents and a longer time horizon.

3.2. Understanding and mastering complex systems

3.2.1. General context

There exist numerous examples of natural and artificial systems where self-organization and emergence occur. Such systems are composed of a set of simple entities interacting in a shared environment and exhibit complex collective behaviors resulting from the interactions of the local (or individual) behaviors of these entities. The properties that they exhibit, for instance robustness, explain why their study has been growing, both in the academic and the industrial field. They are found in a wide panel of fields such as sociology (opinion dynamics in social networks), ecology (population dynamics), economy (financial markets, consumer behaviors), ethology (swarm intelligence, collective motion), cellular biology (cells/organ), computer networks (ad-hoc or P2P networks), etc.

More precisely, the systems we are interested in are characterized by:

- *locality*: Elementary components have only a partial perception of the system's state, similarly, a component can only modify its surrounding environment.
- *individual simplicity*: components have a simple behavior, in most cases it can be modeled by stimulus/response laws or by look-up tables. One way to estimate this simplicity is to count the number of stimulus/response rules for instance.
- *emergence*: It is generally difficult to predict the global behavior of the system from the local individual behaviors. This difficulty of prediction is often observed empirically and in some cases (e.g., cellular automata) one can show that the prediction of the global properties of a system is an undecidable problem. However, observations coming from simulations of the system may help us to find the regularities that occur in the system's behavior (even in a probabilistic meaning). Our interest is to work on problems where a full mathematical analysis seems out of reach and where it is useful to observe the system with large simulations. In return, it is frequent that the properties observed empirically are then studied on an analytical basis. This approach should allow us to understand where lies the frontier between simulation and analysis.
- *levels of description and observation*: Describing a complex system involves at least two levels: the micro level that regards how a component behaves, and the macro level associated with the collective behavior. Usually, understanding a complex system requires to link the description of a component behavior with the observation of a collective phenomenon: establishing this link may require various levels, which can be obtained only with a careful analysis of the system.

We now describe the type of models that are studied in our group.

3.2.2. Multi-agent models

We represent these complex systems with reactive multi-agent systems (RMAS). Multi-agent systems are defined by a set of reactive agents, an environment, a set of interactions between agents and a resulting organization. They are characterized by a decentralized control shared among agents: each agent has an internal state, has access to local observations and influences the system through stimulus response rules. Thus, the collective behavior results from individual simplicity and successive actions and interactions of agents through the environment.

Reactive multi-agent systems present several advantages for modeling complex systems

- agents are explicitly represented in the system and have the properties of local action, interaction and observation;
- each agent can be described regardless of the description of the other agents, multi-agent systems allow explicit heterogeneity among agents which is often at the root of collective emergent phenomena;
- multi-agent systems can be executed through simulation and provide good models to investigate the complex link between global and local phenomena for which analytic studies are hard to perform.

By proposing two different levels of description, the local level of the agents and the global level of the phenomenon, and several execution models, multi-agent systems constitute an interesting tool to study the link between local and global properties.

Despite a widespread use of multi-agent systems, their framework still needs many improvements to be fully accessible to computer scientists from various backgrounds. For instance, there is no generic model to mathematically define a reactive multi-agent system and to describe its interactions. This situation is in contrast with the field of cellular automata, for instance, and underlines that a unification of multi-agent systems under a general framework is a question that still remains to be tackled. We now list the different challenges that, in part, contribute to such an objective.

3.2.3. Current challenges

Our work is structured around the following challenges that combine both theoretical and experimental approaches.

3.2.3.1. Providing formal frameworks

A widespread and consensual formal definition of a multi-agent system is lacking. Our research aims at translating the concepts from the field of complex systems into the multi-agent systems framework.

One objective of this research is to remove the potential ambiguities that can appear if one describes a system without explicitly formulating each aspect of the simulation framework. As a benefit, the reproduction of experiments is facilitated. Moreover, this approach is intended to gain a better insight of the self-organization properties of the systems.

Another important question consists in monitoring the evolution of complex systems. Our objective is to provide some quantitative characteristics of the system such as local or global stability, robustness, complexity, etc. Describing our models as dynamical systems leads us to use specific tools of this mathematical theory as well as statistical tools.

3.2.3.2. Controlling complex dynamical system

Since there is no central control of our systems, one question of interest is to know under which conditions it is possible to guarantee a given property when the system is subject to perturbations. We tackle this issue by designing exogenous control architectures where control actions are envisaged as perturbations in the system. As a consequence, we seek to develop control mechanisms that can change the global behavior of a system without modifying the agent behavior (and not violating the autonomy property).

3.2.3.3. Designing systems

The aim is to design individual behaviors and interactions in order to produce a desired collective output. This output can be a collective pattern to reproduce in case of simulation of natural systems. In that case, from individual behaviors and interactions we study if (and how) the collective pattern is produced. We also tackle "inverse problems" (decentralized gathering problem, density classification problem, etc.) which consist in finding individual behaviors in order to solve a given problem.

MULTISPEECH Team

3. Research Program

3.1. Introduction

As mentioned previously, MULTISPEECH is structured along three research directions that are associated to the three challenges previously described: explicit modeling of speech, statistical modeling of speech, and uncertainty in speech processing.

3.2. Explicit modeling of speech production and perception

Speech signals are the consequence of the deformation of the vocal tract under the effect of the movements of the jaw, lips, tongue, soft palate and larynx to modulate the excitation signal produced by the vocal cords or air turbulence. These deformations are visible on the face (lips, cheeks, jaw) through the coordination of different orofacial muscles and skin deformation induced by the latter. These deformations may also express different emotions. We should note that human speech expresses more than just phonetic content, to be able to communicate effectively. In this project, we address the different aspects related to speech production from the modeling of the vocal tract up to the production of audiovisual speech. On the one hand, we study the relationship from acoustic speech signal to vocal tract, in the context of acoustic-to-articulatory inversion, and from vocal tract to acoustic speech, in the context of articulatory synthesis. On the other hand, we work on expressive audiovisual speech synthesis, where both expressive acoustic speech and visual signals are generated from text. Phonetic contrasts used by the phonological system of any language result from constraints imposed by the nature of the human speech production apparatus. For a given language these contrasts are organized so as to guarantee that human listeners can identify sounds robustly. From the point of view of perception, these contrasts enable efficient processes of categorization in the peripheral and central human auditory system. The study of the categorization of sounds and prosody thus provides a complementary view on speech signals by focusing on the discrimination of sounds by humans, particularly in the context of language learning.

3.2.1. Articulatory modeling

Modeling speech production is a major issue in speech sciences. Acoustic simulation makes the link between articulatory and acoustic domains. Unfortunately this link cannot be fully exploited because there is almost always an acoustic mismatch between natural and synthetic speech generated with an articulatory model approximating the vocal tract. However, the respective effects of the geometric approximation, of the fact of neglecting some cavities in the simulation, of the imprecision of some physical constants and of the dimensionality of the acoustic simulation are still unknown. Hence, the first objective is to investigate the origin of the acoustic mismatch by designing more precise articulatory models, developing new methods to acquire tridimensional MRI data of the entire vocal tract together with denoised speech signals, and evaluating several approaches of acoustic simulation. This will enable the acoustic mismatch to be better controlled and the determination of the potential precision of inversion to be evaluated in particular.

Up to now, acoustic-to-articulatory inversion has been addressed as an instantaneous problem, articulatory gestures being recovered by concatenating local solutions via the determination of trajectories minimizing some articulatory cost. The second objective is thus to investigate how more elaborated strategies (a syllabus of primitive gestures, articulatory targets...) can be incorporated in the acoustic-to-articulatory inversion algorithms to take into account dynamic aspects.

This area of research relies on the equipment available in the laboratory to acquire articulatory data: articulograph Carstens AG501, head-neck antenna to acquire MRI of the vocal tract at Nancy Hospital, and multimodal acquisition system. Very few sites in France benefit from such a combination of acquisition devices.

3.2.2. Expressive acoustic-visual synthesis

Speech is considered as a bimodal communication means; the first modality is audio, provided by acoustic speech signals and the second one is visual, provided by the face of the speaker. Our research impacts both audiovisual and acoustic-only synthesis fields.

In our approach, the Acoustic-Visual Text-To-Speech synthesis (AV-TTS) is performed simultaneously with respect to its acoustic and visible components, by considering a bimodal signal comprising both acoustic and visual channels. A first AV-TTS system was developed resulting in a talking head; the system relied on 3D-visual data (3D markers on the face, data acquired by MAGRIT team) and on an extension of our non-uniform acoustic-unit concatenation text-to-speech synthesis system (SoJA). An important goal is to provide an audiovisual synthesis that is intelligible, both acoustically and visually. Thus, we continue working on adding visible components of the head through a tongue model where the tongue deformations come from EMA data analysis; and a lip-model to tackle the main recurrent problem of the lack of some lip markers in the 3D data. We will also improve the TTS engine to increase the accuracy of the unit selection simultaneously into the acoustic and visual domains (learning weights, feature selection...).

Another challenging research goal is to add expressivity in the AV-TTS. The expressivity comes through the acoustic signal (prosody aspects) and also through head and eyebrow movements. One objective is to add a prosodic component in the TTS engine in order to take into account some given prosodic entities such as emphasis, in order to highlight some important key words. Expressivity could be introduced before the unit selection step but also by developing algorithms intended to modify the parameters of prosody (in the acoustic domain, and in the visual domain as well). One intended approach will be to explore an expressivity measure at sound, syllable and/or sentence levels that describes the degree of perception or realization of an expression/emotion (audio and 3D domain). Such measures will be used as criteria in the selection process of the synthesis system. To tackle this issue we will also investigate Hidden Markov Model (HMM) based synthesis. The flexibility of the HMM-based approach enables the adjustment of the modeling parameters according to the available data and an easy adaption of the system to various conditions. This point will rely upon our experience in HMM modeling.

To acquire the facial data, we consider using marker-less motion capture system using a kinect-like system with a face tracking software. The software presents a user-friendly interface to track and visualize the motion in real time. Audio is also acquired synchronously with facial data. The advantage of this new system is to acquire rapidly the movements of the face with an acceptable quality. This system is used as an alternative relatively low-cost system to the VICON system.

3.2.3. Categorization of sounds and prosody for native and non-native speech

Discriminating speech sounds and prosodic patterns is the keystone of language learning whether in the mother tongue or in a second language. This issue is associated with the emergence of phonetic categories, i.e., classes of sounds which are related to phonemes, and prosodic patterns. The study of categorization is concerned not only with acoustic modeling but also with speech perception and phonology. Foreign language learning raises the issue of categorizing phonemes of the second language given the phonetic categories of the mother tongue. Thus, studies on the emergence of new categories, whether in the mother tongue (for people with language deficiencies) or in a second language, must rely upon studies on native and non-native acoustic realizations of speech sounds and prosody (i.e., at the segmental level and at the supra-segmental level). Moreover, as categorization is a perceptual process, studies on the emergence of categories must also rely on perceptual experiments.

Studies on native sounds have been an important research area of the team for years, leading to the notion of "selective" acoustic cues and the development of acoustic detectors. This know-how will be exploited in the study of non-native sounds. Concerning prosody, studies are focused on native and non-native realizations of modalities (e.g., question, affirmation, command ...), as well as non-native realizations of lexical accents and focus (emphasis). Results aim at providing automatic feedbacks to language learners with respect to acquisition of prosody as well as acquisition of a correct pronunciation of the sounds of the foreign language. Concerning the mother tongue we are interested in the monitoring of the process of sound categorization in the

long term (mainly at primary school) and its relation with the learning of reading and writing skills, especially for children with language deficiencies.

3.3. Statistical modeling of speech

Whereas the first research direction deals with the physical aspects of speech and its explicit modeling, this second research direction is concerned by investigating complex statistical models for speech data. Acoustic models are used to represent the pronunciation of the sounds or other acoustic events such as noises. Whether they are used for source separation, for speech recognition, for speech transcription, or for speech synthesis, the achieved performance strongly depends on the accuracy of these models, which is a critical aspect that is studied in the project. At the linguistic level, MULTISPEECH investigates models for handling the context (beyond the few preceding words currently handled by the n-gram models) and evolutive lexicons necessary when dealing with diachronic audio documents in order to overcome the limited size of the current static lexicons used, especially with respect to proper names. Statistical approaches are also useful for generating speech signals. Along this direction, MULTISPEECH mainly considers voice transformation techniques, with their application to pathological voices, and statistical speech synthesis applied to expressive multimodal speech synthesis.

3.3.1. Acoustic modeling

Acoustic modeling is a key issue for automatic speech recognition. Despite progress made for many years, acoustic modeling is still far from perfect, and current speech recognition applications rely on strong constraints (limited vocabulary, speaker adaptation, restricted syntax...) to achieve acceptable performance. As the acoustic models represent the acoustic realization of the sounds, they have to account for many variability sources, such as speaker characteristics, microphones, noises, etc. Extension of the HMM formalism based on the Dynamic Bayesian Networks (DBN) formalism are investigated further for handling such variability sources; as well as other approaches to dynamically constrain the search space according to known or estimated characteristics of the utterance being processed. Deep Neural Networks (DNN) based approaches will also be investigated as means of making speech recognition systems more accurate and robust. Speaker dependent modeling and speaker adaptation will also be investigated in relation with HMM-based speech synthesis and statistical voice conversion.

State-of-the-art speech recognition systems are still very sensitive to the quality of speech signals they have to deal with; their performance degrades rapidly when they deal with noisy signals. Accurate signal enhancement techniques are therefore essential to increase the robustness of both automatic speech recognition and speech-text alignment systems to noise and non-speech events. In MULTISPEECH, focus is set on Bayesian source separation techniques using multiple microphones and/or models of non-speech events. Some of the challenges include building a non-parametric model of the sources in the time-frequency-channel domain, linking the parameters of this model to the cepstral representation used in speech processing, modeling the temporal structure of environmental noise, and exploiting large audio data sets to automatically discover new models. Beyond the definition of such complex models, the difficulty is to design scalable estimation algorithms robust to overfitting, that will be integrated in the FASST [6] framework that was recently developed.

3.3.2. Linguistic modeling

MULTISPEECH investigates lexical and language models in speech recognition with a focus on improving the processing of proper names and the processing of spontaneous speech. Collaborations are ongoing with the SMarT team on linguistic modeling aspects.

Proper names are relevant keys in information indexing, but are a real problem in transcribing many diachronic spoken documents (such as radio or TV shows) which refer to data, especially proper names, that evolve over the time. This leads to the challenge of dynamically adjusting lexicons and language models through the use of the context of the documents or of some relevant external information possibly collected over the web. Random Indexing (RI) and Latent Dirichlet Allocation (LDA) are two possible approaches to be used for this purpose. Also, to overcome the limitations of current n-gram based language models, we investigate language

models defined on a continuous space in order to achieve a better generalization on unseen data, and to model long-term dependencies. This is achieved through neural network based approaches. We also want to introduce into these new models additional relevant information such as linguistic features, semantic relation, topic or user-dependent information.

Spontaneous speech utterances are often ill-formed and frequently contain disfluencies (hesitations, repetitions...) that degrade speech recognition performance. This is partly due to the fact that disfluencies are not properly represented in linguistic models estimated from clean text data (coming from newspapers for example); hence a particular effort will be set for improving the modeling of these events.

Attention will also be set on pronunciation lexicons in particular with respect to non-native speech and foreign names. Non-native pronunciation variants have to take into account frequent miss-pronunciations due to differences between mother tongue and target language phoneme inventories. Proper name pronunciation variants are a similar problem where difficulties are mainly observed for names of foreign origin that can be pronounced either in a French way or kept close to foreign origin native pronunciation. Automatic grapheme-to-phoneme state-of-the-art approaches, based for example on Joint Multigram Models (JMM) or Conditional Random Fields (CRF) will be further investigated and combined.

3.3.3. Speech generation by statistical methods

Voice conversion consists in building a function that transforms a given voice into another one. MULTISPEECH applies voice conversion techniques to enhance pathological voices that result from vocal folds problems, especially esophageal voice or pathological whispered voice. Voice conversion techniques are also of interest for text-to-speech synthesis systems as they aim at making possible the generation of new voice corpora (other kind of voice, or same voice with different kind of emotion).

In addition to the statistical aspects of the voice conversion approaches, signal processing is critical for good quality speech output. Information on the fundamental frequency is chaotic in the case of esophageal speech or non-existent in the case of the whispered voice. So after applying voice conversion techniques for enhancing pathological voices, the excitation spectrum must be predicted or corrected. That is the challenge that is addressed in the project. Also, in the context of acoustic feedback in foreign language learning, voice modification approaches (either statistical or not) will be investigated to modify the learner's (or teacher's) voice in order to emphasize the difference between the learner's acoustic realization and the expected realization.

Over the last few years statistical speech synthesis has emerged as an alternative to corpus-based speech synthesis. Speaker-dependent HMM modeling constitute the basis of such an approach. The announced advantages of the statistical speech synthesis are the possibility to deal with small amounts of speech resources and the flexibility for adapting models (for new emotions or new speaker), however, the quality is not as good as that of the concatenation-based speech synthesis. The reasons are twofold: first, parameters (F0, spectrum, duration...) are modeled independently and the models, even when taking into account dynamics, do not manage to generate parameters with a good precision. Second, the HMM generates sequences of feature vectors from which the actual speech signals are reconstructed, and this impacts on its quality. MULTISPEECH will focus on an hybrid approach, combining corpus-based synthesis, for its high-quality speech signal output, and HMM-based speech synthesis for its flexibility to drive selection, and the main challenge will be on its application to producing expressive audio-visual speech. One secondary objective will be to unify the HMM-based and the concatenation-based approaches.

3.4. Uncertainty estimation and exploitation in speech processing

After the explicit modeling presented and the statistical modeling that were previously described, we focus here on the uncertainty associated to some processing steps. Uncertainty stems from the high variability of speech signals and from imperfect models. For example, enhanced speech signals resulting from source separation are not exactly the clean original speech signals. Words or phonemes resulting from automatic speech recognition contain errors, and the phone boundaries resulting from automatic speech-text alignment

are not always correct, especially in acoustically degraded conditions. Hence it is important to know the reliability of the results and/or to estimate the uncertainty on the results.

3.4.1. Uncertainty and acoustic modeling

Because small distortions in the separated source signals can translate into large distortions in the cepstral features used for speech recognition, this limits the recognition performance on noisy data. One way to address this issue is to estimate the uncertainty on the separated sources in the form of their posterior distribution and to propagate this distribution, instead of a point estimate, through the subsequent feature extraction and speech decoding stages. Although major improvements have been demonstrated in proof-of-concept experiments using knowledge of the true uncertainty, accurate uncertainty estimation and propagation remains an open issue.

MULTISPEECH seeks to provide more accurate estimates of the posterior distribution of the separated source signals accounting for, e.g., posterior correlations over time and frequency which have not been considered so far. The framework of variational Bayesian (VB) inference appears to be a promising direction. Mappings learned on training data and fusion of multiple uncertainty estimators are also explored. The estimated uncertainties is then exploited for acoustic modeling in speech recognition and, in the future, also for speech-text alignment. This approach may later be extended to the estimation of the resulting uncertainty on the acoustic model parameters and the acoustic scores themselves.

3.4.2. Uncertainty and phonetic segmentation

The accuracy of the phonetic segmentation is important in several cases, as for example for the computation of prosodic features, for avoiding incorrect feedback to the learner in computer assisted foreign language learning, or for the post-synchronization of speech with face/lip images. Currently the phonetic boundaries obtained are quite correct on good quality speech, but the precision degrades significantly on noisy and non-native speech. Phonetic segmentation aspects will be investigated, both in speech recognition (i.e., spoken text unknown) and in forced alignment (i.e., when the spoken text is known). The first case (speech recognition) is connected with the computation of prosodic features for structuring speech recognition output, whereas the second case (forced alignment) is important in the context of non-native speech segmentation for automatic feedbacks in language learning.

In the same way that combining several speech recognition outputs leads to improved speech recognition performance, MULTISPEECH will investigate the combination of several speech-text alignments as a way of improving the quality of speech-text alignment and of determining which phonetic boundaries are reliable and which ones are not, and also for estimating the uncertainty on the boundaries. Knowing the reliability and/or the uncertainty on the boundaries will also be useful when segmenting speech corpora; this will help deciding which parts of the corpora need to be manually checked and corrected without an exhaustive checking of the whole corpus.

3.4.3. Uncertainty and prosody

Prosody information is also investigated as a means for structuring speech data (determining sentence boundaries, punctuation...) possibly in addition with syntactic dependencies (in collaboration with the SYNALP team). Structuring automatic transcription output is important for further exploitation of the transcription results such as easier reading after the addition of punctuation, or exploitation of full sentences in automatic translation. Prosody information is also necessary for determining the modality of the utterance (question or not), as well as determining accented words.

Prosody information comes from the fundamental frequency, the duration of the sounds and their energy. Any error in estimating these parameters may lead to a wrong decision. MULTISPEECH will investigate estimating the uncertainty on the duration of the phones (see uncertainty on phonetic boundaries above) and on the fundamental frequency, as well as how this uncertainty shall be propagated in the detection of prosodic phenomena such as accented words or utterance modality, or in the determination of the structure of the utterance. In a first approach, uncertainty estimation will rely on the comparison, and possibly the combination, of several estimators (several segmentation processes, several pitch algorithms).

ORPAILLEUR Project-Team

3. Research Program

3.1. From KDD to KDDK

Keywords: knowledge discovery in databases, knowledge discovery in databases guided by domain knowledge, data mining

Knowledge discovery in databases is a process for extracting from large databases knowledge units that can be interpreted and reused. From an operational point of view, a KDD system includes databases, data mining modules, and interfaces for interactions, e.g. editing and visualization. The KDD process is based on three main operations: selection and preparation of the data, data mining, and finally interpretation of the extracted units.

The process of "knowledge discovery in databases guided by domain knowledge" extends the KDD cycle with a fourth step, where extracted units are represented within a knowledge base to be reused. The KDDK process –as implemented in the research work of the Orpailleur team– is based on *data mining methods* that are either symbolic or numerical:

- Symbolic methods are based on frequent itemsets search, association rule extraction, Formal Concept Analysis and extensions [113].
- Numerical methods are based on higher order stochastic models, namely second-order Hidden Markov Models (HMM2) and Hidden Markov fields (HMRF), which are especially designed for an efficient modeling of space and time [12].

The principle summarizing KDDK can be understood as a process going from complex data to knowledge units being guided by domain knowledge. Two original aspects can be underlined: (i) the knowledge discovery process is guided by domain knowledge at each step of the process, and (ii) the extracted units are embedded within knowledge-based systems for problem solving purposes.

One main operation in the research work of Orpailleur on KDDK is *classification*, which is a polymorphic process involved in modeling, mining, representing, and reasoning tasks. Moreover, the KDDK process is intended to feed knowledge-based systems working in application domains, e.g. agronomy, biology, chemistry, cooking and medicine, and also in the context of semantic web, text mining, information retrieval, and ontology engineering.

3.2. Knowledge Discovery guided by Domain Knowledge

Keywords: knowledge discovery, data mining, formal concept analysis, classification, frequent itemset search, association rule extraction, second-order Hidden Markov Models

Classification problems can be formalized by means of a class of objects (or individuals), a class of attributes (or properties), and a binary correspondence between the two classes, indicating for each individual-property pair whether the property applies to the individual or not. The properties may be features that are present or absent, or the values of a property that have been transformed into binary variables. Formal Concept Analysis (FCA) relies on the analysis of such binary tables and may be considered as a symbolic data mining technique to be used for extracting a set of formal concepts then organized within a concept lattice [113] (concept lattices are also known as "Galois lattices" [103]).

In parallel, the search for frequent itemsets and the extraction of association rules are well-known symbolic data mining methods, related to FCA (actually searching for frequent itemsets can be understood as traversing a concept lattice). Both processes usually produce a large number of items and rules, leading to the associated problems of "mining the sets of extracted items and rules". Some subsets of itemsets, e.g. frequent closed itemsets (FCIs), allow to find interesting subsets of association rules, e.g. informative association rules. This is why several algorithms are needed for mining data depending on specific applications [45].

Among useful patterns extracted from a database, frequent itemsets are usually thought to unfold "regularities" in the data, i.e. they are the witnesses of recurrent phenomena and they are consistent with the expectations of the domain experts. In some situations however, it may be interesting to search for "rare" itemsets, i.e. itemsets that do not occur frequently in the data (contrasting frequent itemsets). These correspond to unexpected phenomena, possibly contradicting beliefs in the domain. In this way, rare itemsets are related to "exceptions" and thus may convey information of high interest for experts in domains such as biology or medicine.

From the numerical point of view, a Hidden Markov Model (HMM2) is a stochastic process aimed at extracting and modeling a sequence of stationary distributions of events. Such models can be used for data mining purposes, especially for spatial and temporal data as they show good capabilities to locate patterns both in time and space domains.

Moreover, stochastic models have been designed to mine temporal sequences having a spatial dimension, for example the succession of land uses in a territory. One main Markovian assumption states that the temporal event succession in a given place depends only on the temporal event successions in neighboring points. By means of stochastic models such as hierarchical hidden Markov models and Markov random fields, it is possible to perform an unsupervised clustering of a spatial territory for discovering "patches" characterized by time and space regularities in their temporal successions.

3.3. Text Mining

Keywords: knowledge discovery form large collection of texts, text mining, information extraction, document annotation, ontologies

The objective of a text mining process is to extract useful knowledge units from large collections of texts [110]. The text mining process shows specific characteristics due to the fact that texts are complex objects written in natural language. The information in a text is expressed in an informal way, following linguistic rules, making text mining a particular task. To avoid information dispersion, a text mining process has to take into account –as much as possible– paraphrases, ambiguities, specialized vocabulary and terminology. This is why the preparation of texts for text mining is usually dependent on linguistic resources and methods.

From a KDDK perspective, text mining is aimed at extracting "interesting units" (nouns and relations) from texts with the help of domain knowledge encoded within an ontology (also useful for text annotation). Text mining is especially useful in the context of semantic web for ontology engineering [105]. In the Orpailleur team, the focus is put on the mining of real-world texts in application domains such as biology and medicine, using mainly symbolic data mining methods, and especially Formal Concept Analysis. Accordingly, the text mining process may be involved in a loop used to enrich and to extend linguistic resources. In turn, linguistic and ontological resources can be exploited to guide a "knowledge-based text mining process".

3.4. Knowledge Systems and Semantic Web

Keywords: knowledge representation, ontology, description logics, classification-based reasoning, case-based reasoning, semantic web, information retrieval

Usually, people try to take advantage of the web by searching for information (navigation, exploration), and by querying documents using search engines (information retrieval). Then people try to analyze the obtained results, a task that may be difficult and tedious. Semantic web is an attempt for guiding search for information with the help of software agents, that are in charge of asking questions, searching for answers, classifying and interpreting the answers. However, a software agent may be able to read, understand, and manipulate information on the web, if and only if the knowledge necessary for achieving those tasks is available, and this is why ontologies are of main importance. Thus, there is a need for knowledge representation languages for annotating documents, describing the content of documents and giving a semantics to this content.

In particular, the knowledge representation language used for designing ontologies is the OWL language, which is based on description logics (DLs [100]). In OWL, knowledge units are represented within concepts (or classes), with attributes (properties of concepts, or relations, or roles), and individuals. The hierarchical organization of concepts (and relations) relies on a subsumption relation (i.e. a partial ordering).

The inference services are based on subsumption, concept and individual classification, two tasks related to "classification-based reasoning". Furthermore, classification-based reasoning can be extended into case-based reasoning (CBR), which relies on three main operations: retrieval, adaptation, and memorization. Given a target problem, retrieval consists in searching for a source (memorized) problem similar to the target problem. Then, the solution of the source problem is adapted to fulfill the constraints attached to the target problem, and possibly memorized for further reuse.

SEMAGRAMME Project-Team

3. Research Program

3.1. Foundations

The Sémagramme project relies on deep mathematical foundations. We intend to develop models based on well-established mathematics. We seek two main advantages from this approach. On the one hand, by relying on mature theories, we have at our disposal sets of mathematical tools that we can use to study our models. On the other hand, developing various models on a common mathematical background will make them easier to integrate, and will ease the search for unifying principles.

The main mathematical domains on which we rely are formal language theory, symbolic logic, and type theory.

3.1.1. Formal language theory

Formal language theory studies the purely syntactic and combinatorial aspects of languages, seen as sets of strings (or possibly trees or graphs). Formal language theory has been especially fruitful for the development of parsing algorithms for context-free languages. We use it, in a similar way, to develop parsing algorithms for formalisms that go beyond context-freeness. Language theory also appears to be very useful in formally studying the expressive power and the complexity of the models we develop.

3.1.2. Symbolic logic

Symbolic logic (and, more particularly, proof-theory) is concerned with the study of the expressive and deductive power of formal systems. In a rule-based approach to computational linguistics, the use of symbolic logic is ubiquitous. As we previously said, at the level of syntax, several kinds of grammars (generative, categorial...) may be seen as basic deductive systems. At the level of semantics, the meaning of an utterance is capture by computing (intermediate) semantic representations that are expressed as logical forms. Finally, using symbolic logics allows one to formalize notions of inference and entailment that are needed at the level of pragmatics.

3.1.3. Type theory and typed λ-calculus

Among the various possible logics that may be used, Church's simply typed λ-calculus and simple theory of types (a.k.a. higher-order logic) play a central part. On the one hand, Montague semantics is based on the simply typed λ-calculus, and so is our syntax-semantics interface model. On the other hand, as shown by Gallin, [56] the target logic used by Montague for expressing meanings (i.e., his intensional logic) is essentially a variant of higher-order logic featuring three atomic types (the third atomic type standing for the set of possible worlds).